Essays in Applied Causal Analysis

Development, Real Estate, and International Economics

Inauguraldissertation zur Erlangung der Würde eines

DOCTOR RERUM OECONOMICARUM

der Wirtschafts- und Sozialwissenschaftlichen Fakultät

der Universität Bern

vorgelegt von

Daniel Steffen

von Goms, Wallis

Bern, August 2020

Die Fakultät hat diese Arbeit am 10. Dezember 2020 auf Antrag der Gutachter Prof. Dr. Aymo Brunetti, Prof. Dr. Isabel Günther und Prof. Dr. Mauricio Romero als Dissertation angenommen, ohne damit zu den darin ausgesprochenen Auffassungen Stellung nehmen zu wollen.

Acknowledgments

For the last four years, I have had the opportunity to write my dissertation in a very inspiring environment. It has been four intense, enlightening, and exciting years. I am very grateful to a multitude of people without whom my doctoral thesis would not have been possible.

I would like to express my deepest gratitude to my main thesis supervisor, Aymo Brunetti. He has always given me the freedom to pursue research questions that interest me and was always there for me when I needed support. I have been able to grow enormously thanks to him.

I am also indebted to my additional supervisors, Isabel Günther and Mauricio Romero. As part of the "Impact Award", which we won together with Consciente, we had the opportunity to present our work to Isabel Günther and her team. Their feedback was invaluable for us. Mauricio Romero gave us precious feedback on our work countless times and always took the time to thoroughly discuss research ideas and projects.

Further, I want to thank Blaise Melly, Max von Ehrlich, and Michael Gerfin. They patiently discussed the empirical strategy of each of my papers with me. I am also thankful to all my colleagues at the Department of Economics and the Center for Regional Economic Development. Working with them was a great pleasure.

My deepest appreciation goes to all my co-authors; I have learned so much from all of them. I will especially remember the field visits that I was able to experience with Konstantin and Martina: All those challenges which seemed to be insurmountable at first, but which we finally managed to overcome; all the impressive experiences that we shared; and most importantly, the countless times we laughed heartily together (best remembered are the "crema de chocolate" for breakfast, daring fights against spiders, the slow but hard-earned progress in playing "Trompo", the good old firecrackers in the middle of the night and many more). These visits were certainly the most intense, thought-provoking, and memorable moments of my dissertation.

A special thanks goes to the staff of Consciente, the NGO that implemented our projects in El Salvador. Their hard work made our projects possible and thanks to them I felt always at home in El Salvador. Their dedication to fighting for a better future is truly inspiring.

This thesis would not have been possible without Simone, my family, and my friends. I owe my deepest gratitude to Simone for her patience (when I lay out the long version of my research to her), our engaged discussions, and her tremendous support. My parents made it possible that I could always do what inspires me and my siblings were and always will be my idols. Their unconditional support and love made this journey possible. Last but not least, I want to thank my friends for having me accompanied during all those years.

Daniel Steffen, Bern, August 2020

Contents

1 Introduction

Every day, individuals, companies, NGOs, and political actors make countless decisions. These decisions shape the world of tomorrow and have an impact on billions of people. Often, however, decision-makers lack the information to make targeted decisions. We invest enormous resources or adopt far-reaching political measures to achieve our goals, yet, we often fail to measure thoroughly enough whether these efforts are effective in reaching our set targets or whether the opposite is true. For example, scientists and policy-makers believed that providing textbooks or other non-teacher inputs to schools in low-income countries could substantially increase test scores (Glewwe, Kremer and Moulin, 2009). Consequently, vast resources were invested in such non-teacher inputs for years. However, it is known today that providing inputs as textbooks or flip charts without changing pedagogy or governance has no or only limited impact on education quality (Kremer, Brannen and Glennerster, 2013; Glewwe, Kremer and Moulin, 2009; Mbiti et al., 2019). So, even though the intentions behind our decisions and actions may be meaningful, it often happens that our talent, time, and dedication to certain issues or our monetary resources are wasted due to uninformed decisions.

Data is key to avoid such uninformed decisions and rather taking more effective action. Data allows us to examine the (unintended) effects of actions or policies and draw the right conclusions. It gives us the possibility to measure how we achieve our goals effectively and more efficiently. It also enables us to identify the causes of problems that we would otherwise not fully understand. To make such evidence-based decisions, data must be collected, analyzed with credible methods, and presented understandably.

Such credible empirical studies were scarce – or even inexistent – 30 years ago or as Leamer (1983, cited in Angrist and Pischke, 2010) describes it in the 1980es: "Hardly anyone takes data analysis seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analysis seriously." The trust in empirical analyses was low and data analysis in political or scientific debates played a neglectable role. However, in the last 30 years, empirical methods have experienced a "credibility revolution" (Angrist and Pischke, 2010). Policy and scientific debates can no longer be imagined without empirical analysis. Imbens and Wooldridge (2009) conclude that after two decades of research, the econometric and statistical analysis of causal effects has reached a level of maturity that makes it an important tool in many areas. For example, data-based impact analyses become more and more important in public administrations around the world and shape critical policies. This development goes so far that some speak of an "empirical revolution in public administration" (Pomeranz, 2019, p.4).

How did empirical work become such an important and irreplaceable tool for decision-makers in public administration, companies, and also in the scientific debate? One reason is certainly the fact that the availability of data has increased enormously, which has also made empirical analyses much more affordable (Angrist and Pischke, 2010; Pomeranz, 2019). However, Angrist and Pischke (2010, p. 6) argue that the "clear eyed focus on research design is at the heart of the credibility revolution in empirical economics." First, the focus on research design has led to a more transparent discussion of empirical strategies. Second, and more importantly, the use of better strategies per se improved empirical work largely and increased the trust in its results. These design-based methods either use (field) experiments or carefully implemented quasi-experimental methods, which can credibly identify causal relationships. The idea behind these methods is to explicitly or implicitly build hypothetical counterfactual worlds to compare with the "treated" world. This thesis is to be understood in the spirit of this credibility revolution and the goal to estimate causal relationships with experimental or quasi-experimental methods is the overarching topic of the thesis. It consists of four essays on three different topics using three different (quasi-)experimental methods – always with the aim of combining data with design-based methods to find efficient solutions to fundamental problems and to quantify the causal effects of certain policies or actions of market players.

Chapters 2 and 3 are *field experiments* addressing the question of how the problem of low education quality in many low- and middle-income countries can be tackled. Basic education is recognized as a key factor for economic development and the fight against poverty. Yet, the quality of education is alarmingly low in many low- and middle-income countries despite an impressive increase in enrollment rates in the last two decades. El Salvador – where both experiments take place – is an example of this situation: Sixth graders, after having attended more than 1000 math lessons, are on average only able to answer 57 percent of first and second grade math questions correctly and a mere 14 percent of questions at their current level. We argue that teachers are the key to quality education and therefore, we focus on the often insufficient content knowledge of teachers. Computer-Assisted Learning (CAL) has proven to be a very promising approach to improve education quality, because it mitigates many of the challenges public education systems in low- and middle-income countries face. We designed two experiments that allow us to gain further insights into how CAL might enhance quality learning in an environment of poorly qualified teachers and how policy-makers and NGO leaders can create more cost-effective education interventions.

Chapter 2 provides novel evidence on the relative effectiveness of CAL software and traditional teaching. Based on a randomized controlled trial in Salvadoran primary schools, we evaluate three interventions that aim to improve learning outcomes in mathematics: *(i)* teacher-led classes, *(ii)* CAL classes monitored by a technical supervisor, and *(iii)* CAL classes instructed by a teacher. As all three interventions involve the same amount of additional mathematics lessons, we can directly compare the productivity of the three teaching methods. The main contributions of this study are the assessment of the value-added of CAL software compared to traditional

lessons as well as the complementarities between CAL software and teachers. CAL lessons lead to larger improvements in students' mathematics skills than traditional teacher-centered classes. Also, teachers add little to the effectiveness of learning software. Overall, our results highlight the value of CAL approaches in an environment of poorly qualified teachers.

Chapter 3 addresses the poor mathematics content knowledge of primary school teachers directly. At baseline, the primary school teachers participating in this study master on average 43 percent of the official Salvadoran curriculum that they are supposed to teach. Based on a randomized controlled trial, we study the impact of an in-service teacher training program in El Salvador that targets the content knowledge of primary school teachers. The five-month training combines *(i)* computer-assisted self-studying and *(ii)* monthly workshops. It is the first study that examines an intervention that explicitly focuses on teachers' content knowledge. After the training, program teachers score significantly higher in mathematics than their peers from the control group. The teacher training program proved more successful in raising the competence for concepts taught in higher grades (i.e. grades five and six) compared to lower grades (i.e. grades two to four), and it was particularly effective among young teachers.

Chapter 4 follows a quasi-experimental approach and examines the effects of a drastic real estate regulation in Switzerland, the Second Home Initiative (SHI). The SHI prevents the construction of second homes in all municipalities with a second home share of 20 percent or more. This initiative is drastic because one in five municipalities in Switzerland is affected and a total of 17 percent of all housing units are second homes. This setup allows applying Difference-in-Differences (DD) type methods to estimate the causal effects: Municipalities affected by SHI form the treatment group, while unaffected municipalities build the control group. In a first step, I show that a classical DD approach is not valid in this context. Therefore, I use an extension of the *synthetic control method* (SCM) in which I deal with multiple treatment units instead of one treatment unit as in the classic approach. By analyzing the SHI, this study contributes to a small but growing literature investigating the effects of interventions that aim at restricting non-local real estate demand. It is also one of the first studies to apply the SCM with multiple treatment units and develops an innovative approach to compute precise statistical significance. Results show that the SHI caused a delayed decrease in real estate prices by –10 percent to –19 percent three to five years after the acceptance of the SHI. The decrease in prices is likely to be caused by adverse effects on local economies and due to a "lock-in" effect caused by the SHI.

Chapter 5 investigates the effect of outward foreign direct investments (FDI) of Swiss multinational enterprises on home employment. In this chapter, we do not explicitly build a control group as a counterfactual, but we try to come as close as data and the setup allow to a causal interpretation of the relationship by applying an *instrumental variable approach* – another prominent approach of the arsenal of quasi-experimental methods (Angrist and Pischke, 2010). In a context where the debate on globalization is reviving, we aim at providing empirical findings that

allow for an evidence-based discussion about the effects of internationalization of firms on the domestic labor market. Using Swiss firm-level data we construct a novel instrument to identify a direct negative displacement effect and an indirect positive output effect. Our study contributes to the literature by investigating the effect of outward FDI in the context of a small but relatively heavily exposed economy – Switzerland. We find that FDI to high-income countries have a positive effect on domestic jobs, while FDI to lower middle-income countries are associated with a loss of domestic jobs. Overall, the effect of outward FDI on home employment is small and tends to create more domestic jobs than it relocates.

2 The Relative Effectiveness of Teachers and Learning Software: Evidence from a Field Experiment in El Salvador*

2.1 Introduction

While net primary school enrollment rates in low-income countries climbed from 56 percent in 2000 to 81 percent in 2019, learning outcomes have failed to keep pace. Less than 15 percent of primary school children in low-income countries pass minimum proficiency thresholds in reading and math, compared to about 95 percent of pupils in high-income countries (World Bank, 2018, p. 8). Public schooling systems in developing countries face multiple challenges that curb their productivity. These include a mismatch between national curricula and student abilities (Pritchett and Beatty, 2014), large and heterogeneous classes (Mbiti, 2016; Glewwe and Muralidharan, 2016), and low levels of effort among poorly trained teachers (Chaudhury et al., 2006; Bold et al., 2017a). A much-noticed approach to overcome these barriers is the use computer-assisted learning software (e.g. The Economist, 2017). Computer-assisted learning (CAL) has several advantages over traditional teaching methods, as it allows for self-paced learning that is tailored to the abilities of the student, provides instant feedback and is less sensitive to the motivation and skills of teachers. Previous studies on the impact of technology-based teaching methods on learning outcomes are encouraging. CAL interventions are usually found to improve students' test scores and seem to be particularly beneficial if the software is used to personalize instructions (for a review see Escueta et al., 2020).[1]

Yet, most studies evaluate CAL lessons that were offered as a supplement to regular classes, meaning that beneficiaries experienced a considerable expansion of school time compared to the untreated students in the control group. Thus, it remains unclear whether learning gains are actually attributable to the use of the software or if additional lessons conducted by a teacher

[1] Experimental studies on CAL interventions in low- and middle-income countries include Banerjee et al. (2007, math in Indian primary schools), Carrillo, Onofa and Ponce (2011, language and math in Ecuadorian primary schools), Yang et al. (2013, language and math in Chinese primary schools), Mo et al. (2015, math in Chinese primary schools), Lai et al. (2015, language and math in Chinese primary schools), and Muralidharan, Singh and Ganimian (2019, language and math with Indian secondary school pupils). They consistently report positive intent-to-treat estimates on learning outcomes that range between 0.1 standard deviations (σ) and 0.4σ.

*This chapter is joint work with Konstantin Büchel, Martina Jakob, Christoph Kühnhanss and Aymo Brunetti.

might have produced similar or even better results.[2] In addition, there is little evidence on whether CAL is a substitute for certified teachers or if it is a complement to them. Finally, previous research has mostly evaluated specifically customized software which is available in a limited number of languages. As a result, many policy-makers with an interest in implementing CAL cannot draw on software that is readily available and has been successfully evaluated.

Based on a randomized controlled trial, this paper examines the relative effectiveness of primary school math teachers and a freely available CAL software that features content in more than 30 languages. To disentangle the effects of additional teaching and the use of a learning software, the experimental design features three different treatments: The first treatment comprises additional math lessons instructed by a teacher (henceforth labeled as TEACHER). The second and third treatments are additional math lessons based on CAL software; one group of classes is monitored by technical supervisors (CAL + SUPERVISOR), while the other group is taught by teachers (CAL + TEACHER). Teachers had to be officially qualified to teach math in primary schools, whereas supervisors were laymen instructed to provide no content-related help to students. CAL lessons were taught using an offline application of the "Khan Academy" platform, and the three treatment arms were implemented by the Swiss-Salvadoran NGO *Consciente*.

We conducted the experiment between February and October 2018 with a sample of 198 primary school classes spanning grades three to six in the rural district of Morazán, El Salvador. 29 out of 57 eligible schools were randomly selected for program participation. The 158 classes from these 29 schools were then randomly assigned to either Treatment 1 (i.e. TEACHER, 40 classes), Treatment 2 (i.e. CAL + SUPERVISOR, 39 classes), Treatment 3 (i.e. CAL + TEACHER, 39 classes) or a within-school control group (40 classes). In non-program schools, a random sample of 40 classes was drawn resulting in a "pure" control group that is not subject to potential treatment externalities.

Our analysis establishes four key findings. First, the additional CAL classes had a considerable impact on students' math skills. Being assigned to additional CAL lessons increased their math scores by 0.21σ (p-value=0.00) when overseen by a supervisor and by 0.24σ (p-value=0.00) when instructed by teachers. These are intent-to-treat estimates reflecting an average attendance rate of 59 percent. Using the treatment assignment as instrumental variable for attendance, we estimate that participating in all 46 additional CAL lessons (each lasting 90 minutes) translates to average learning gains of 0.38σ (p-value=0.00) and 0.40σ (p-value=0.00), respectively. This is equivalent to the average increase in math abilities over 1.2 school years.

Second, additional CAL lessons seem to have been more productive than the additional math lessons instructed by a teacher. The intent-to-treat effect of being assigned to additional teacher-led classes without CAL was 0.15σ (p-value=0.01). Hence, students assigned to CAL + TEACHER

[2] To our knowledge, the only study that evaluates the effectiveness of CAL lessons as a substitute to regular teaching in the development context was conducted by Linden (2008) in India. While attending *additional* CAL lessons raised math scores of second and third graders, CAL had a negative impact when it *substituted* regular classes. As the author points out, the study sample only covers NGO-run schools with well trained staff and innovative teaching methods. While it is unclear whether these findings translate to the challenging contexts of public education systems in developing countries, they still raise doubts about the inherent benefits of technology-based instruction.

outperformed students assigned to TEACHER by 0.09σ (p-value=0.10); when analyzing percentage scores instead of standardized IRT-scores the according p-value decreases to 0.06. The CAL treatment overseen by technical supervisors (CAL + SUPERVISOR) was also slightly more successful in raising student learning than traditional teaching, even though this difference clearly falls short of statistical significance (p-value=0.24). The advantage of CAL lessons relative to teacher-centered lessons was most pronounced in the domain of number sense and elementary arithmetic, and less so with respect to geometry, measurement and data. Focusing on number sense and elementary arithmetic, the difference between the CAL and non-CAL treatments increases to 0.11σ (p-value=0.06) for CAL classes instructed by teachers and to 0.09σ (p-value=0.12) for the CAL lessons monitored by supervisors.

Third, we present multifaceted evidence that points to a rather low productivity of teachers. The difference in learning gains between within-program school control classes and those classes receiving additional teacher-centered math lessons was close to zero and statistically insignificant (p-value=0.78). Similarly, teachers did not provide much "value added" to the learning software: the estimated impact for CAL lessons instructed by teachers is slightly higher than for CAL lessons conducted by supervisors but the difference is negligible and statistically insignificant (p-value=0.65). Moreover, the productivity of teachers dropped as the complexity of concepts increased: The impact of additional math lessons instructed by teachers decreased sharply with both the grade level and the baseline achievement of their students, while the effect of the CAL-based lessons was largely insensitive to students' grades and initial ability levels. To gain a better understanding of the mechanisms behind these findings, we conducted a comprehensive teacher math assessment covering the primary school curriculum of El Salvador. This assessment documents very poor content knowledge among the teachers hired by the NGO. Furthermore, regular math teachers in local primary schools are even less proficient in their subject. Potential productivity gains resulting from an introduction of CAL to regular classes may thus be larger than suggested by our estimates, since the NGO's contract teachers had better content knowledge and employed more modern pedagogical techniques than regular math teachers.

Finally, we document substantial treatment externalities. At endline, students in within-program school control classes outperformed pure control classes by 0.14σ (p-value=0.02), although they were only indirectly exposed to the three treatments. In particular, we find evidence for spillovers from the two CAL treatments. While we cannot comprehensively pin down the mechanisms at work, suggestive evidence points toward social learning. At the same time, the data rejects hypotheses operating via direct exposure of control students to the treatments (i.e. non-compliance) or behavioral adjustments in response to the experimental design.

This study makes several contributions to the literature on educational interventions in developing countries. First, it improves our understanding of how CAL performs relative to alternative teaching models. To our knowledge, this is the first well-identified study assessing the value-added of CAL in a public school setting of a developing country. As opposed to Linden (2008), who documents a negative value-added of CAL in NGO-administered schools in India, our findings suggest that CAL has the potential to outperform traditional teacher-led instruc-

tion, especially if teachers are poorly qualified. While CAL has been regularly praised in terms of its individualized and interactive pedagogy (e.g. Banerjee et al., 2007; Muralidharan, Singh and Ganimian, 2019), our findings highlight that it may also be a promising approach to mitigate the adverse effects of teachers' inadequate content knowledge and pedagogical knowledge, that has been recently documented for several developing countries (e.g. Bold et al., 2017a).

Second, we present the first experimental test of the complementarities between teachers and learning software. In our setting, teachers seem to play a marginal role in the success of technology-based instruction, with CAL lessons being almost equally effective when provided by a supervisor rather than a certified teacher. Thus, teachers and learning software are likely substitutes and not complements. Only few experimental studies aspire to distinguish between complementary and substitutable inputs entering the educational production function; notable exceptions are recent papers by Mbiti et al. (2019) on the complementarity between school grants and teacher incentives in Tanzanian primary schools, and by Attanasio et al. (2014) on the complementarity between psychosocial stimulation programs and nutritional supplements in early childhood development.

Third, we contribute to the broader literature on treatment externalities (e.g. Miguel and Kremer, 2004; Baird et al., 2015). By including control classes from treatment schools as well as spatially separated pure control classes from non-treatment schools into our experimental design, this study provides a credible identification of potential externalities. Our findings underscore the importance of appreciating the possibility of externalities in the design of experimental evaluation studies, even when such effects appear unlikely at first sight. Moreover, the presence of positive treatment externalities provide a strong rationale in favor of scaling the evaluated program.

Finally, this study adds to the accumulated evidence on the effectiveness of CAL by evaluating a widely available off-the-shelf software. In contrast to software tested in previous evaluations, Khan Academy is freely available and features extensive math contents in more than 30 languages.[3] Since the employed software is arguably one of the most important features of a CAL intervention, our findings bear direct policy relevance for decision-makers around the globe that are looking for a readily available learning software suitable in non-English speaking countries.[4]

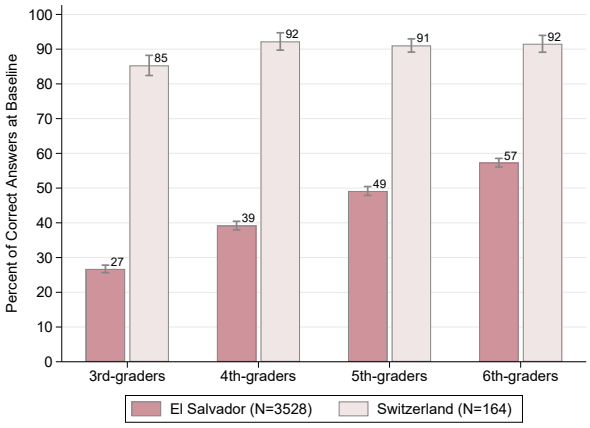
2.2 Context and Intervention

2.2.1 Context

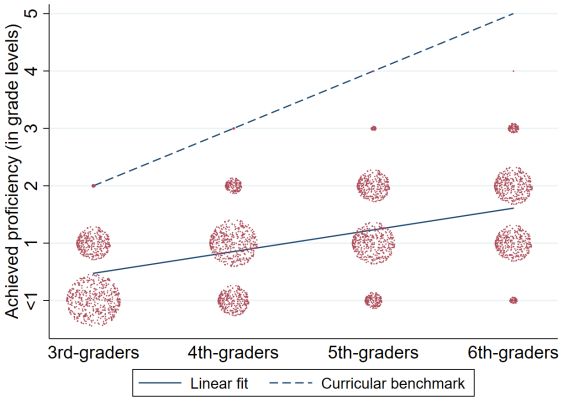
El Salvador is a lower middle-income country in Central America. The country's net primary enrollment rates are estimated at 80 percent, which is 7 percentage points below the average of lower middle-income countries. While most children get to attend primary school, access becomes

[3] The full version is available in 16 languages including Spanish, English, Chinese, French and Portuguese. A subset of content is available in another about 20 languages including Russian, Hindi and Swahili. For further information see the Khan Academy website `https://www.khanacademy.org/` (last visit: 01.12.2019).

[4] Another off-the-shelf learning software that has been successfully evaluated is Mindspark (see Muralidharan, Singh and Ganimian, 2019), which operates in English and Hindi for math and language training.



(a) Share of correct answers on 1st/2nd grade math questions among Salvadoran and Swiss pupils.



(b) Assessed grade level in math among third to sixth graders in Morazán early in their school year.

Figure 2.1: Math learning outcomes in Morazán (Panels a & b) and Switzerland (Panel a).
Note: Panel (b) illustrates the achieved proficiency in math (measured in grade levels) among third to sixth graders in Morazán at the beginning of their school year. A student, each represented by a dot, needs to score at least 50 percent correct answers on grade specific items in order to reach the next proficiency level. Since the test was administered in the first weeks of their school year, a third grader answered first and second grade items and therefore may be assigned to grade level 2, 1 or <1 depending on her performance. The size of the bubbles are proportional to the number of students they represent. Further explanations are provided in Appendix A.1.1.
Source: Baseline data, February 2018.

more selective at later stages of an educational career with secondary and tertiary enrollment standing at 67 percent and 28 percent, respectively.[5]

The department of Morazán is a poor and rural region in the northeast of the country with roughly 200,000 inhabitants. An average person in Morazán lives on 3.80 USD per day and, according to national definitions, almost 50 percent of the households face multifaceted poverty. With an illiteracy rate of more than 20 percent, Morazán ranks second-last among all Salvadorian departments in terms of educational attainment (DIGESTYC, 2018).

Our math assessments with 3,528 third to sixth graders conducted in February 2018 further reveal that primary school children barely grasp the most elementary concepts in math. Figure 2.1a shows that the share of correct answers to first and second grade questions increases from 27 percent among third graders to 57 percent among sixth graders, who by then should have attended more than 1000 math lessons. To put these numbers into context, we conducted the same test with 164 pupils in Switzerland, who answered on average between 85 percent and 92 percent of the items correctly. Even the worst performing Swiss third grader outperformed the median sixth grader in Morazán.

Several challenges that plague Morazán's schooling system can help to explain its low productivity. For instance, our monitoring data from school visits reveal high rates of teacher absenteeism so that, on average, 25 percent of regular lessons are canceled. Low teacher motivation mixes with outdated pedagogical techniques that basically follow the logic of "copy, learn by

[5]Enrollment statistics according to the *World Development Indicators* provided online by the World Bank, see `https://data.worldbank.org/indicator` (last access: 26.10.2019)

heart, and reproduce". And despite relatively small class sizes – the pupil-teacher ratio averages 28-to-1 at the national level and 19-to-1 in our sample – heterogeneous student performance and an overambitious curriculum make it difficult to teach at an appropriate level. As Figure 2.1b shows, third to sixth graders lag considerably behind the official curriculum and this gap widens as children move up to higher grade levels. Moreover, performance heterogeneity within classes is considerable. In the majority of classes, students' math ability spans three grades or more (for further explanations see Appendix A.1.1). In general, the public schooling system in El Salvador faces similar issues to those reported for other low- and middle income countries.[6]

The Salvadoran Ministry of Education has recently put considerable effort into addressing learning deficiencies in public schools. While primary schooling has been typically confined to either morning or afternoon lessons throughout El Salvador, the new SI-EITP policy[7] aims to extend school time over a full day and to complement traditional teaching with innovative learning approaches (MINED, 2013). The government hopes that longer schooldays will not only boost learning outcomes, but also shield children from the influence of criminal gangs. Within the scope of this countrywide program, the Ministry of Education seeks to cooperate with NGOs in order to collectively promote an open and flexible curriculum. While all schools received official instructions to expand their school days, most of them have not put the policy into practice due to a lack of resources to pay for further teaching staff.

2.2.2 Intervention

In this context, we evaluate the impact of an educational initiative on math abilities of primary school children of grades three to six. The program features three intervention arms that offer two additional lessons of 90 minutes per week and almost double the beneficiaries' number of math lessons during the program phase. The first intervention arm comprises additional math lessons instructed by a teacher without using software. The second and third intervention arms are additional math lessons based on computer-assisted learning software; one group of classes is taught by teachers, while the other group is instructed by supervisors.

The *CAL-lessons* in the second and third intervention arm were based on an offline application of the learning platform *Khan Academy*, which is known as *KA Lite*. This freely available software provides a wide range of instructional videos and exercises for every difficulty level. While the learning tool is not directly adaptive, it allows teachers to track the progress of each student

[6] The pupil-teacher ratio in middle-income countries averages 24-to-1, while it climbs to 40-to-1 in low income countries (Unesco, 2019); in some contexts, such as rural India, it can even reach 90-to-1 (Mbiti, 2016). Besides the large class size, students' abilities and preparation levels are often very heterogeneous, which is also the case in our data. For example, Muralidharan, Singh and Ganimian (2019) report for their sample of 116 Indian middle schools that students' ability in the median classroom spans four grades in both math and language, while we obtain 3 grade levels for primary schools. Moreover, Pritchett and Beatty (2014) show that the pace of learning is very slow in developing countries and that there is a mismatch between curriculum and student abilities. This is consistent with what we observe in Figure 2.1b. Finally, low teacher motivation is a well-known issue: Chaudhury et al. (2006) find that 19 percent of teachers in developing countries are absent during unannounced visits, while our monitoring data suggests that 25 percent of classes in Morazán's primary school are canceled.

[7] SI-EITP stands for *Sistema Integrado de Escuelas Inclusivas de Tiempo Pleno*, which translates to *Integrated System of Inclusive Full Time Schools*.

and assign appropriate contents based on prior performance. To tailor instruction to students' learning levels, a set of working plans covering different content units was prepared. Based on a placement test, children received a plan that was viewed as accurate for their respective level and they could then proceed to subsequent plans at their own pace. Since one computer was available per student, each child could follow its individual learning path. Typically, students started with materials from lower grades and then slowly progressed towards contents corresponding to their actual grades.

A similar methodology was used for the first intervention arm that features more traditional *math lessons instructed by a teacher*. According to their initial math skills, children were arranged in two different table groups where they worked on plans tailored to their ability. Teachers were instructed to explain important concepts, correct students' work at home and promote children (or entire table groups) to subsequent plans when necessary. While this strategy only allows for a crude approximation of teaching to each child's ability level, it represents a degree of individualization that can realistically be achieved without the help of technology.

To pay credit to the *social component of learning*, all treatments combined individualized learning with educational games. For this purpose, a manual containing animation, concentration and math games was developed. The manual compiles simple techniques to promote students' collective learning as well as their motivation to participate in class. Games were usually played at the beginning or at the end of each session. While supervisors were instructed to use animation and concentration games, teachers were additionally introduced to a series of math games.

The contracted *teachers* were required to be officially qualified to instruct grades three to six in math. That is, they all possessed a university degree and had either completed a teacher education, or another study program combined with a one-year pedagogical course. Teachers were selected based on a brief math assessment and a job interview. They were employed on short-term contracts and earned 300 USD per month for assuming four classes.[8] For lessons that were canceled, teachers received no remuneration. To optimize the comparability of treatments, all teachers were assigned an equal number of CAL and non-CAL lessons. Before and during the intervention, teachers were trained to operate the learning software and they reviewed mathematical concepts as well as central pedagogical strategies including the use of educational games. Teaching was tightly monitored by our partner NGO through monthly meetings and unannounced classroom visits during the implementation phase.

The *supervisors* received only technical training and were paid substantially less than teachers, that is 180 USD for taking care of four classes. They were required to have minimal IT skills and some experience in dealing with children, while teaching credentials and a specific educational degree were not among the selection criteria. During the intervention, supervisors were instructed to restrain from providing any content-specific help. Like teachers, supervisors were employed on short-term contracts and paid conditional on the number of classes they conducted.

[8] This corresponds to 8 $\times$ 90 minutes of teaching per week, or – including preparatory work – to a 60 percent job. A smaller group of teachers only assumed two classes (i.e 4 $\times$ 90 minutes of teaching per week, or approx. a 30 percent job).

2.3 Research Design

This study is built around an RCT to identify the causal impact of the three interventions arms. It started in February 2018 with a baseline assessment and a survey covering all control and program classes. The additional math classes began in April 2018 and were implemented until the end of the school year in fall 2018.[9] The endline tests took place in October 2018, six months after the start of the intervention. Again, all program and control classes took part in the endline tests.

2.3.1 Sampling and Randomization

Our sampling and randomization scheme has three layers, as exemplified in Figure 2.2. Starting point are all 302 primary schools in Morazán. In coordination with the NGO and the regional Ministry of Education, we defined the following eligibility criteria for a pre-selection of schools:

- ***School size**, eliminates 221 schools:* A school was considered too small, if it had integrated classes (across grades) or gaps in its grade structure (i.e. not at least one class per grade). This guarantees that every eligible school has at least four different classes in grades three to six, and therefore can participate with at least (*i*) one CAL+TEACHER, (*ii*) one CAL+SUPERVISOR, (*iii*) one TEACHER, and (*iv*) one control class;
- ***Security**, eliminates 14 of the remaining 81 schools:* Based on an assessment by the local staff and the regional Ministry of Education, schools located in areas dominated by criminal gangs were excluded due to security concerns;
- ***Accessibility**, eliminates 7 of the remaining 67:* Schools that are hardly accessible by car were discarded. To inform this decision we relied on Google-Maps driving times and a validation by local staff and the regional Ministry of Education;
- ***Electricity**, eliminates 3 of the remaining 60 schools:* Schools without a (close-by) power supply did not qualify for the program.

After this pre-selection, 57 schools with a total of 320 eligible classes and about 6400 students remained in the sample. In *randomization stage 1*, 29 of the 57 schools were randomly chosen to participate in the program. To improve balance, the assignment was stratified by school size, local population density and students' access to a computer room.

In *randomization stage 2a*, we randomly assigned the 158 classes in the 29 selected program schools to the control group or one of the three intervention arms. Following Morgan and Rubin (2012) we re-run the randomization routine until the interventions were balanced across schools and grades. This mechanism assigned 39 classes to CAL+TEACHER, 39 classes to CAL+SUPERVISOR, 40 classes to TEACHER, and 40 classes to the control group. We account for

[9] The school year in El Salvador starts in mid-January and ends in November.

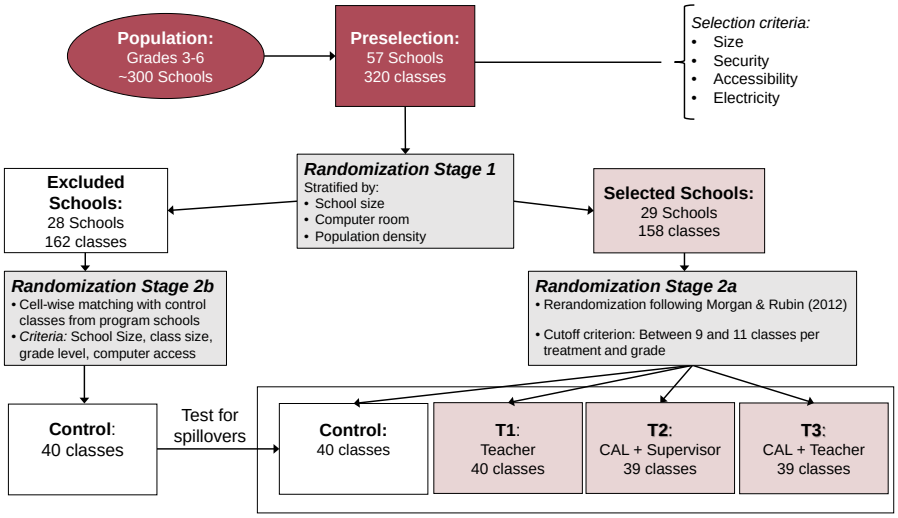


Figure 2.2: Sampling and randomization scheme.

the re-randomization procedure when comparing estimates within program schools by computing randomization inference test statistics based on 2000 random draws subject to the identical cut-off criterion. Our choice to run 2000 draws is guided by Young (2019, p. 572), who finds no appreciable change in rejection rates beyond this threshold. To implement the randomization tests we rely on Stata's RITEST-package developed by Hess (2017).

As prominently discussed in Miguel and Kremer (2004), interventions can have spill-over effects on non-participating students from the same school or area. A unique feature of our design allows us to estimate the size of such treatment externalities. For this purpose, in *randomization stage 2b*, 40 additional control classes from non-treatment schools were included in the study. These additional "pure" control classes are spatially separated from the intervention, and thus not affected by the NGO's work. The *pure control classes* were randomly selected from the 28 control schools by matching them cell-wise to the distribution of control classes from program schools, accounting for school size, grade level, class size and students' access to computers.

This procedure yields five different groups of primary school classes, namely the 39 or 40 classes assigned to each of the three treatment groups, 40 control classes from the 29 program schools, and 40 pure control classes from the 28 control schools.

2.3.2 Data

In the course of the evaluation, four types of data were gathered: (*i*) Math learning outcomes of students were assessed before and after the intervention, (*ii*) socio-demographic statistics stem from a survey that children answered prior to the baseline math assessment, (*iii*) administrative data on schools was collected between October 2017 and February 2018, and (*iv*) monitoring data was recorded during unannounced school visits throughout the program phase. Table 2.1 shows summary statistics for the main variables collected before the start of the program as well as absence rates at the endline and baseline assessment. In particular, it displays means and

standard errors for the different variables by treatment status, and tests whether the mean is equal across the five groups.

While the treatment and control groups do not differ significantly on any observable dimension at baseline, Table 2.1 shows a sizeable increase in the absence rates between baseline and endline assessment. Before both rounds of data collection we updated comprehensive class lists of registered pupils. This revealed that large numbers of children either migrated out of Morazán or discontinued their education. We achieved an attendance of about 95 percent registered pupils in both rounds, but since classes shrank during the school year, the overall attrition at endline almost hits the 10 percent mark. Importantly, Table 2.1 does not point toward systematic differences in attrition.[10] Moreover, compliance with the protocol was very good in the sense that only 38 out of 3197 students (i.e. 1.2 percent) within our estimation sample switched between different classes, grades or schools.

Math Learning Outcomes. The math assessments include 60 items covering the primary school curriculum of El Salvador. The weighting of questions across the three main topics (a) number sense & elementary arithmetic (∼65 percent), (b) geometry & measurement (∼30 percent), and (c) data & statistics (∼5 percent) was closely aligned with the national curriculum. Moreover, we verified the appropriateness of each question through a careful mapping to the national curriculum and a feedback loop involving the regional Ministry of Education and local education experts. The math problems presented to the children mostly required a written answer (as opposed to a multiple choice format) and were inspired by El Salvador's official textbooks as well as various international sources of student assessments; the Appendix Section A.2 explains the design of our assessments step by step.

In the appendix, we further present detailed statistics on the distribution of student test scores and the difficulty of the items. Top or bottom coding is neither an issue with respect to students nor the selected items: Table A.5 shows that virtually all of the items (except one for fifth graders in the endline assessment) were at least once answered correctly or incorrectly. Likewise, Table A.4 documents that only about 0.5 percent of test-takers scored zero points, while nobody achieved the maximum score. In general, the assessments seem to nicely capture the different performance levels, with the scores being roughly normally distributed around a median of 30 percent (3rd graders) to 40 percent (6th graders) correct answers (see Figure A.3).

A particularly nice feature of our math assessments is that they allow us to project all outcomes on a common ability scale by drawing on techniques from psychology called *Item Response Theory (IRT)*(e.g. de Ayala, 2009). This implies that we can directly compare children across grades and express their learning gains between base- and endline assessment in terms of how many additional school years would be required to reproduce the same effect. The conversion of our estimates into program effects measured in terms of additional school years is explained in the Appendix A.2.

[10] We examine this more closely in Table A.1 in the appendix, confirming that the treatment status is not significantly correlated with presence at the endline test.

Table 2.1: Balance at baseline and absence rates during assessments

Panel A: Math scores (N=3528)	(1) **T1:** Math w. teacher	(2) **T2:** CAL w. supervisor	(3) **T3:** CAL w. teacher	(4) Within school controls	(5) Pure control classes	(6) p-value
%-share correct answers	30.33	33.47	31.97	32.60	30.80	0.45
	(1.80)	(1.90)	(2.07)	(1.32)	(2.00)	
Std. IRT math score	0.01	0.18	0.08	0.08	0.00	0.72
	(0.14)	(0.14)	(0.16)	(0.10)	(0.15)	
Panel B: Sociodemographics (N=3528)						
Female student	0.50	0.52	0.55	0.51	0.49	0.43
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	
Student age	−0.09	−0.01	0.02	−0.03	−0.03	0.70
	(0.08)	(0.09)	(0.09)	(0.06)	(0.09)	
Household size	5.56	5.61	5.57	5.55	5.50	0.92
	(0.13)	(0.12)	(0.12)	(0.08)	(0.12)	
Household assets index	0.55	0.55	0.54	0.56	0.56	0.88
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	
Panel C: Class room variables and absence rates during assessments (N=198)						
Class size	18.40	19.33	18.69	18.13	18.32	0.92
	(1.37)	(1.35)	(1.37)	(0.96)	(1.54)	
Female teacher	0.80	0.77	0.77	0.73	0.55	0.14
	(0.10)	(0.10)	(0.10)	(0.07)	(0.11)	
Absence rate at baseline (%)	3.88	3.15	5.39	4.39	3.38	0.59
	(1.33)	(1.16)	(1.74)	(0.95)	(1.15)	
Absence rate at endline (%)	9.09	9.72	10.50	9.99	8.10	0.72
	(2.09)	(2.04)	(2.18)	(1.63)	(2.00)	
Panel D: School variables (N=49)				Treatment schools	Pure control schools	p-value
# classes grade 3–6				5.48	6.25	0.32
				(0.43)	(0.76)	
Computer lab				0.79	0.75	0.73
				(0.08)	(0.13)	
Local population density				0.18	0.19	0.63
				(0.01)	(0.02)	

Notes: This table presents the mean and standard error of the mean (in parenthesis) for several characteristics of students (Panels A & B), class rooms (Panel C), and schools (Panel D), across treatment groups. The student sample consists of all students tested by the research team during the baseline survey in February 2018. Column 6 shows the p-value from testing whether the mean is equal across all treatment groups. IRT-scores are standardized such that $\mu = 0$ and $\sigma = 1$ for the pure control group. The household asset index measures what share of the following assets a household owns: Books, electricity, television, washmachine, computer, internet and car. Local population density is the municipality's population density measured in 1000 inhabitans per km^2. Standard errors are clustered at the class level in Panels A & B, and at the school level in Panel C. * p<0.10, ** p<0.05, *** p<0.01.

Socio-Demographic Survey. The socio-demographic survey was distributed 15 minutes before the baseline math assessment began. It asked students about their age, gender, household composition, household assets and parental education. Since literacy can be an issue, questions were illustrated with pictures and the enumerators helped children to understand and answer them correctly.

Administrative Data on Schools. In the run-up to the study we collected various administrative data on Morazán's school. While the government gathers thematically broad data on the school environment through a paper-and-pencil survey administered to school principals, the data turned out to be of rather poor quality. To obtain utilizable information on the class structure, enumerators had to call each school during the first weeks of January, because the planning data from official sources was too unreliable. Moreover, the paper-and-pencil surveys left many missing values, so that we had to discard most items due to an insufficient coverage. We therefore decided to use a minimal set of school level variables, which were either comprehensively available, relatively cheap to supplement, or essential for the study. These include the number of grade three to grade six classes (school size), information on the presence of gangs (security at school), accessibility measures based on Google-Map estimates and validated by local staff, power supply according to the administrative survey and validated via phone calls, student access to computer labs according to the administrative survey and validated via phone calls, and local population density from the National Bureau of Statistics.

Monitoring Data. From May to September 2018, NGO staff made on average five unannounced school visits (about 1000 visits in total) to collect monitoring data. They covered both regular lessons as well as program lessons and collected data on teacher attendance, student attendance, computer usage, and the implementation of the additional math lessons in the afternoon.

2.4 Results

2.4.1 The Overall Program Effects

We begin by estimating *intent to treat* (ITT) effects of being assigned to one of the three programs (i.e. $\beta_{T1}, \beta_{T2}, \beta_{T3}$) or the within-program school control classes (i.e. β_{CX}) using

$$Y_{ics}^{EL} = \alpha + \beta_{T1} T1_{cs} + \beta_{T2} T2_{cs} + \beta_{T3} T3_{cs} + \beta_{CX} CX_{cs} + \delta Y_{ics}^{BL} + X_{ics}^{'}\gamma + V_{cs}^{'}\lambda + \phi_k + \epsilon_{1ics}, \quad (2.1)$$

where Y_{ics}^{EL} is the endline math score of student i in class c and school s; math scores are either measured as percentage of correct answers or as the IRT-score normalized to μ=0 and σ=1 based on the baseline score of the pure control group. The binary treatment indicators are defined as follows: $T1$ equals one for those assigned to extra math lessons conducted by a teacher, $T2$ equals one for those assigned to extra CAL lessons overseen by a supervisor, $T3$ equals one

for those assigned to extra CAL lessons instructed by a teacher, and CX equals one for those assigned to within-program school control classes that are potentially subject to externalities. Our control variables include Y_{ics}^{BL} that stands for the baseline math score, X_{ics} representing a set of student-level control variables (i.e. age standardized by average grade age, gender, household size and household assets), and V_{cs} comprising a set of classroom-level variables (i.e. indicator for grade level, class size and teacher gender). Finally, ϕ_k stands for k strata fixed effects and ϵ_{1ics} represents the error term.

The upper panel of Table 2.2 displays the program effect relative to pure control classes (i.e. $\hat{\beta}_{T1}$, $\hat{\beta}_{T2}$, $\hat{\beta}_{T3}$ and $\hat{\beta}_{CX}$) and the lower panel of Table 2.2 presents estimates for the pairwise differences between the three treatment groups in program schools. The lower panel reports p-values obtained from a randomization inference test statistic based on 2000 random draws subject to the identical cut-off criterion as used in our re-randomization scheme (see Section 2.3). In the upper panel, however, p-values are based on traditional clustered standard errors, since the assignment to program schools and pure control schools did not involve re-randomization.[11]

Students who were assigned to one of the treatments perform significantly better in the endline assessment than students assigned to the pure control classes. Compared to the pure control students, participants assigned to extra classes with math teachers (i.e. $T1$) score 2.6 percentage points or 0.15σ better, students assigned to CAL classes with supervisors (i.e. $T2$) score about 3.9 percentage points or 0.21σ better, and students assigned to CAL classes with a teacher (i.e. $T3$) score 4.3 percentage points or 0.24σ better. Remarkably, students in control classes within program schools (i.e. CX) also perform 2.4 percentage points or 0.14σ better than students in pure control classes. As we discuss in Section 2.5.1, our analysis points towards spillovers from CAL-lessons to the within program school control classes, while we find no evidence for direct exposure of control units (i.e. non-compliance) or behavioral changes at the level of the school administration or regular teachers.

Finally, we test whether the gaps in the endline performance of students assigned to one of the three treatments (defined as β_{T4}, β_{T5}, and β_{T6}) are statistically different from zero. While we find that the two CAL treatments outperform additional math classes, only the difference between additional math classes and CAL classes conducted by a teacher is statistically significant at the 10 percent level: students assigned to CAL+TEACHER outperform students assigned to TEACHER by 1.7 percentage points or 0.085σ with p-values ranging from 0.059 to 0.117.

On the one hand, this is novel evidence that CAL delivers measurable learning gains in a Latin American context using off-the-shelf learning software. And the impact of additional CAL

[11] Moreover, we cannot properly apply randomization inference to the upper panel due to missing information on ability levels of non-selected classes from pure control schools. As we show in Appendix A.1.3, randomization inference in the upper panel is based on draws that include on average 37 percent missing data points. Consequently, p-values obtained from these randomization tests increase by a factor of about 5 to 10 compared to p-values from traditional inference with clustered standard errors. While this is clearly too conservative, our main conclusion are not altered when we apply randomization inference to the upper panel (see Table A.3). The only notable difference is that program externalities, captured by β_{CX}, turn insignificant with p-values around 0.13. When we apply traditional inference to the lower panel, as shown in Table A.2, changes in p-values are very small and do not show a clear pattern.

Table 2.2: ITT-Estimates on the effects of the different interventions on children's math scores

	Percent Correct		IRT-Scores	
	(1)	(2)	(5)	(6)
T1: Lessons with Teacher	2.904***	2.643**	0.165***	0.152**
	(0.005)	(0.012)	(0.006)	(0.013)
T2: CAL-Lessons with Supervisor	4.095***	3.869***	0.226***	0.214***
	(0.000)	(0.000)	(0.000)	(0.000)
T3: CAL-Lessons with Teacher	4.554***	4.328***	0.250***	0.238***
	(0.000)	(0.000)	(0.000)	(0.000)
CX: Control Classes for Externalities	2.595**	2.407**	0.147**	0.137**
	(0.011)	(0.017)	(0.013)	(0.020)
$\beta_{T4} \coloneqq \beta_{T2} - \beta_{T1} = 0$	1.191	1.226	0.061	0.063
p-value (β_{T4}=0)	(0.214)	(0.194)	(0.268)	(0.244)
$\beta_{T5} \coloneqq \beta_{T3} - \beta_{T1} = 0$	1.650*	1.686*	0.084	0.086
p-value (β_{T5}=0)	(0.069)	(0.059)	(0.117)	(0.102)
$\beta_{T6} \coloneqq \beta_{T3} - \beta_{T2} = 0$	0.459	0.460	0.024	0.023
p-value (β_{T6}=0)	(0.618)	(0.615)	(0.650)	(0.653)
Adjusted R^2	0.66	0.67	0.69	0.70
Observations	3197	3197	3197	3197
Individual & Classroom Controls	No	Yes	No	Yes
Baseline Score	Yes	Yes	Yes	Yes
Stratum & Grade FE	Yes	Yes	Yes	Yes

Notes: In the upper panel (coef. $\beta_{T1} - \beta_{CX}$), p-values are based on traditional clustered standard errors. In the lower panel (coef. $\beta_{T4} - \beta_{T6}$), p-values are based on a two-sided randomization inference test statistic that the placebo coefficients are larger than the actual; randomization inference is based on 2000 random draws.
* p<0.10, ** p<0.05, *** p<0.01.

lessons is considerable: Expressing the impact estimates in terms of school years suggests that the effect of the CAL interventions is equivalent to the average student's progress in 0.6 to 0.7 school years (see Appendix A.2). On the other hand, traditional math classes conducted by teachers are relatively ineffective compared to additional math lessons with CAL-software: In comparison to the within program school control classes, boosting the supply of conventional math lessons by roughly 80 percent delivered no measurable impact. Importantly, the performance difference between CAL classes taught by teachers and additional teacher-centered math classes is statistically (marginally) significant. We interpret this as suggestive evidence that the learning gains reported in a series of CAL-evaluations can – at least partially – be attributed to the learning software and not necessarily to the increase in number of math lessons.

2.4.2 Heterogeneity Analysis

We now examine effect heterogeneity along several dimensions. We first decompose effects by subtopics, before we explore effect heterogeneity along baseline ability, grade level and class size.

Program Effects by Subtopic

In this subsection, we explore the impact of the three interventions on learning outcomes by topics. In accordance with the official curriculum, 65 percent of the items cover number sense and arithmetic (NSEA), 30 percent of the items cover geometry and measurement (GEOM), and 5 percent of the items cover data, probability and statistics (DSP). In particular, we re-estimate Equation (2.1) but calculate separate math scores based on *(i)* NSEA-questions and *(ii)* GEOM- as well as DSP-questions.

The ITT-effects on students' NSEA skills are shown in Table 2.3. We find that both CAL treatments had a more pronounced effect on the NSEA score than on the overall math ability. Students who were assigned to CAL classes with supervisors score 4.6 percentage points or 0.24σ higher in NSEA questions than students assigned to pure control classes; this is an increase of 10 percent to 20 percent compared to the overall impact reported in Table 2.2. The NSEA math score of students assigned to CAL classes with teachers is 4.9 percentage points or 0.26σ higher than the score of students assigned to pure control classes; again this effect is 10 percent to 15 percent larger compared to estimates based on all questions. Since the impact on the NSEA math score remains about the same for students receiving additional math classes instructed by teachers, the gap between CAL and conventional teaching widens.

When we compare the learning gains attributed to CAL with the gains attributed to the additional math classes without software the differences range between 1.7 and 2.1 percentage points or 0.092σ and 0.115σ. The corresponding p-values lie in between 0.046 and 0.055 for the CAL classes with teachers and between 0.093 and 0.129 for CAL classes with supervisors. Hence, when focusing on NSEA questions, the overall pattern remains qualitatively similar to the estimations including all subject domains, but the gap between the two CAL treatments and additional math classes in the traditional sense (i.e. without the use of software) becomes more pronounced.

Table 2.4 shows the results that are based on GEOM- and DSP-items. Focusing on these topics mitigates the impact of both CAL treatments. The effects compared to pure control classes remain significant but they decrease considerably in magnitude. The results show, for instance, that additional CAL lessons conducted by a teacher increase the NSEA-score by about 5 percentage points, while the increase in the combined GEOM- and DSP-score is only 3.5 percentage points. Since this drop is less pronounced for those classes receiving additional math lessons instructed by a teacher, the within treatment school comparisons yield insignificant effects.

These results show that computer-assisted learning software can be a valuable substitute to traditional teaching, but its impact seems to be sensitive to the concepts that are taught. While the lower-bound effects net of any spillovers are consistently significant for CAL + TEACHER and just at the edge of the 0.1-threshold for CAL + SUPERVISOR, the measured differences seem primarily driven by the pronounced improvements in the domains of number sense and elementary arithmetic. The intervention was less successful in shifting abilities to solve questions on geometry, measurement, data and statistics: the point estimates decrease by about 30 percent,

Table 2.3: ITT-Estimates on the effects of the interventions on children's *NSEA*-scores

	Percent Correct		IRT-Scores	
	(1)	(2)	(5)	(6)
T1: Lessons with Teacher	3.174***	2.849***	0.166***	0.146**
	(0.002)	(0.006)	(0.006)	(0.013)
T2: CAL-Lessons with Supervisor	4.907***	4.581***	0.258***	0.238***
	(0.000)	(0.000)	(0.000)	(0.000)
T3: CAL-Lessons with Teacher	5.225***	4.895***	0.279***	0.259***
	(0.000)	(0.000)	(0.000)	(0.000)
CX: Control Classes for Externalities	2.711***	2.463**	0.145**	0.130**
	(0.008)	(0.012)	(0.013)	(0.020)
$\beta_{T4} \coloneqq \beta_{T2} - \beta_{T1} = 0$	1.733	1.732*	0.092	0.091
p-value (β_{T4}=0)	(0.103)	(0.093)	(0.129)	(0.115)
$\beta_{T5} \coloneqq \beta_{T3} - \beta_{T1} = 0$	2.051**	2.047**	0.113*	0.112*
p-value (β_{T5}=0)	(0.046)	(0.047)	(0.051)	(0.055)
$\beta_{T6} \coloneqq \beta_{T3} - \beta_{T2} = 0$	0.318	0.315	0.021	0.021
p-value (β_{T6}=0)	(0.750)	(0.752)	(0.706)	(0.714)
Adjusted R^2	0.62	0.63	0.65	0.65
Observations	3197	3197	3197	3197
Individual & Classroom Controls	No	Yes	No	Yes
Baseline Score	Yes	Yes	Yes	Yes
Stratum & Grade FE	Yes	Yes	Yes	Yes

Notes: In the upper panel (coef. $\beta_{T1} - \beta_{CX}$), p-values are based on traditional clustered standard errors. In the lower panel (coef. $\beta_{T4} - \beta_{T6}$), p-values are based on a two-sided randomization inference test statistic that the placebo coefficients are larger than the actual; randomization inference is based on 2000 random draws.
* p<0.10, ** p<0.05, *** p<0.01.

and the p-values clearly exceed the 0.1-threshold for statistical significance. This sub-analysis also confirms the strikingly low productivity of certified teachers: Whether we compare the performance across items on basic arithmetic or items on geometry and data analysis, classes receiving additional math lessons conducted by teachers do not perform better than control classes subject to externalities.

Effect Heterogeneity by Baseline Ability, Grade Level and Class Size

We continue the heterogeneity analysis by discussing Figure 2.3, which plots kernel-weighted locally-smoothed means of the endline test score at each percentile of the baseline test score by treatment status. Figure 2.3a shows that endline tests scores in the control group for spillovers are slightly higher than those in the pure control group at all percentiles of the baseline test score, but the 95 percent confidence bands mostly overlap. Comparing pure control classes to the TEACHER classes in Figure 2.3b shows that the latter outperform the former at low percentiles of the baseline score, while there is no difference at higher percentiles. Both CAL intervention

Table 2.4: ITT-Estimates on the effects of the interventions on children's *GEOM & DSP*-scores

	Percent Correct		IRT-Scores	
	(1)	(2)	(5)	(6)
T1: Lessons with Teacher	2.433*	2.132*	0.155**	0.140*
	(0.055)	(0.093)	(0.035)	(0.057)
T2: CAL-Lessons with Supervisor	3.207***	3.014**	0.196***	0.187***
	(0.009)	(0.014)	(0.006)	(0.009)
T3: CAL-Lessons with Teacher	3.646***	3.472***	0.201***	0.193**
	(0.006)	(0.008)	(0.008)	(0.010)
CX: Control Classes for Externalities	2.773**	2.561**	0.159**	0.149**
	(0.032)	(0.048)	(0.036)	(0.050)
$\beta_{T4} \coloneqq \beta_{T2} - \beta_{T1} = 0$	0.775	0.882	0.041	0.047
p-value (β_{T4}=0)	(0.498)	(0.432)	(0.543)	(0.464)
$\beta_{T5} \coloneqq \beta_{T3} - \beta_{T1} = 0$	1.213	1.340	0.046	0.053
p-value (β_{T5}=0)	(0.279)	(0.221)	(0.481)	(0.412)
$\beta_{T6} \coloneqq \beta_{T3} - \beta_{T2} = 0$	0.438	0.458	0.005	0.006
p-value (β_{T6}=0)	(0.692)	(0.669)	(0.934)	(0.926)
Adjusted R^2	0.46	0.47	0.49	0.50
Observations	3197	3197	3197	3197
Individual & Classroom Controls	No	Yes	No	Yes
Baseline Score	Yes	Yes	Yes	Yes
Stratum & Grade FE	Yes	Yes	Yes	Yes

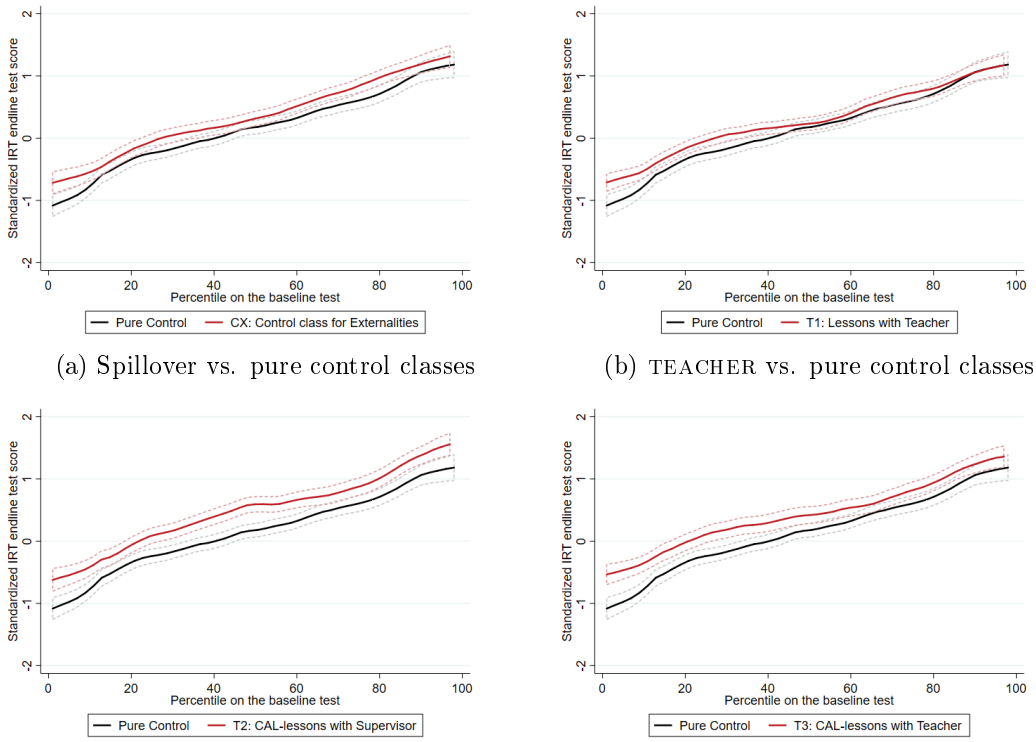
Notes: In the upper panel (coef. $\beta_{T1} - \beta_{CX}$), p-values are based on traditional clustered standard errors. In the lower panel (coef. $\beta_{T4} - \beta_{T6}$), p-values are based on a two-sided randomization inference test statistic that the placebo coefficients are larger than the actual; randomization inference is based on 2000 random draws.
* p<0.10, ** p<0.05, *** p<0.01.

groups, as illustrated in Figures 2.3c and 2.3d, achieve considerably higher endline scores than pure control classes across all percentiles in the baseline achievement, although the gap seems to narrow at higher percentiles in the CAL + TEACHER group.

In a next step, we examine the functional relation between treatment effect and baseline achievement more closely. Similarly, we further investigate whether the reported effects vary by grade level or class size. To do so, we estimate

$$\begin{aligned} Y_{ics}^{EL} = \alpha &+ \beta_{T1} T1_{cs} + \beta_{T2} T2_{cs} + \beta_{T3} T3_{cs} + \beta_{CX} CX_{cs} \\ &+ \theta_1 (T1_{cs} \times Var_{ics}) + \theta_2 (T2_{cs} \times Var_{ics}) \\ &+ \theta_3 (T3_{cs} \times Var_{ics}) + \theta_{CX} (CX_{cs} \times Var_{ics}) \\ &+ \delta Y_{ics}^{BL} + X_{ics}^{'} \gamma + V_{cs}^{'} \lambda + \phi_k + \epsilon_{2ics} \end{aligned} \tag{2.2}$$

where $(T_{cs} \times Var_{ics})$ is the interaction of the treatment dummy with the variable of interest (i.e. baseline math score, grade level and class size). Except for the four interaction terms, Equation (2.2) is identical to our benchmark estimation equation, i.e. Equation (2.1).



(a) Spillover vs. pure control classes

(b) TEACHER vs. pure control classes

(c) CAL + SUPERVISOR vs. pure control classes

(d) CAL + TEACHER vs. pure control classes

Figure 2.3: Endline test scores by treatment status and baseline percentiles.
Note: The figures present kernel-weighted local mean smoothed plots relating endline test scores to percentiles in the baseline achievement by treatment status alongside 95 percent confidence bands.

In terms of baseline math ability, the regression analysis confirms our visual analysis of Figure 2.3. Regarding the effect of additional math classes instructed by teachers, the effect size and baseline achievement are indeed negatively correlated (see column 1 in Table 2.5). This suggests that teachers were more effective in improving the performance of children with low math ability than those children who performed well in the baseline assessment. The regression also yields negative signs for the interaction between the baseline math score and T2 (i.e. CAL + SUPERVISOR) and T3 (i.e. CAL + TEACHER), but the p-values do not reach the 10 percent threshold. Hence, the benefit of attending CAL-based lessons was independent of initial ability levels, while the effectiveness of teachers without software was particularly low among well-performing students.

A similar pattern emerges when we study effect heterogeneity by grade level of the participating students (see column 2 in Table 2.5). The effects of the CAL treatments do not significantly vary with the grade level of students, but we find that additional math lessons taught by a teacher are least effective in higher grades. This corroborates the finding that without the help of learning software, teachers in Morazán seem to be least effective when explaining more complex concepts.

Finally, we find that large class sizes reduce the effectiveness of teachers (see column 3 in Table 2.5), no matter whether they use CAL software or not. This pattern does not emerge

Table 2.5: Effect heterogeneity along baseline ability, grade level and class size

Treatment indicators interacted with:	Baseline Math Score	Grade Level	Class Size (log)
Dependent variable: Std. IRT-Score	(1)	(2)	(3)
T1: Lessons with Teacher × Var.	−0.105***	−0.140***	−0.437***
	(0.004)	(0.000)	(0.004)
T2: CAL-Lessons with Supervisor × Var.	−0.014	−0.052	−0.109
	(0.741)	(0.250)	(0.434)
T3: CAL-Lessons with Teacher × Var.	−0.038	−0.058	−0.270*
	(0.284)	(0.181)	(0.052)
CX: Classes exposed to Externalities × Var.	−0.004	−0.023	−0.118
	(0.913)	(0.675)	(0.482)
Adjusted R^2	0.70	0.70	0.70
Observations	3197	3197	3197
Baseline Score	Yes	Yes	Yes
Individual & Classroom Controls	Yes	Yes	Yes
Stratum & Grade Level FE	Yes	Yes	Yes

Notes: p-values are based on class-level clustered standard errors and are shown in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

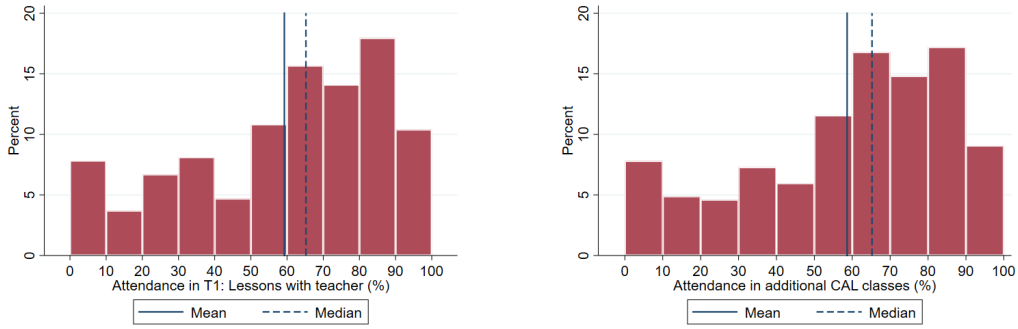
for CAL classes overseen by supervisors, which seems plausible since supervisors were directed to refrain from explaining math contents but solely provided technical assistance. Comparing the point estimates of the interaction terms of the two treatments conducted by teachers, we find that the effect of traditional classes ($\hat{\theta}_1$=-0.437, p-value=0.005) is more sensitive to class size than the effect of CAL-lessons instructed by teachers ($\hat{\theta}_3$=-0.270, p-value=0.052). Overall, this confirms the notion that computer-based learning can mitigate the problems related to large class sizes (e.g. Banerjee and Duflo, 2011; Muralidharan, Singh and Ganimian, 2019).

2.4.3 Program Attendance and IV-Estimates

Our benchmark analysis focuses on ITT-estimates that do not account for the actual attendance rate of students in the additional math lessons. In this section, we therefore present data on the overall compliance, examine the correlation between individual attendance rates and individual endline scores, and finally discuss instrumental variable estimates for the impact of the three interventions assuming full attendance.

Figure 2.4 plots the distribution in attendance rates across all eligible students. With an average attendance rate of 59 percent, participation of students was a weak spot of the program. Attendance rates slightly varied across the three treatments, although the differences are statistically insignificant: Additional CAL classes instructed by teachers achieved the highest participation (60 percent), followed by additional classes instructed by teachers (59 percent) and CAL classes conducted by a supervisor (57 percent).

The individual attendance rate of students is strongly correlated with their performance in the endline math assessment, as one would expect considering that the programs successfully



(a) T1: Additional math lessons with teachers (b) T2 & T3: Additional CAL-lessons

Figure 2.4: Attendance of students in additional math lessons.

increased math learning outcomes. Figure 2.5 plots the residual endline IRT-score (net of all control variables including baseline scores) on the y-axis, and the attendance rates of the students on the x-axis. We aggregated the individual data points into 15 bins in order to improve readability, and plot the correlation by treatment type. Figure 2.5a covers those students that were assigned to additional math classes taught by teachers, while Figure 2.5b illustrates the correlation between attendance and residual endline scores for the two CAL interventions.[12]

We next appraise the question, how much children would have learned had they fully participated in the additional math lessons they were offered. To do so, we estimate an IV-model, with the first-stage estimation being specified as

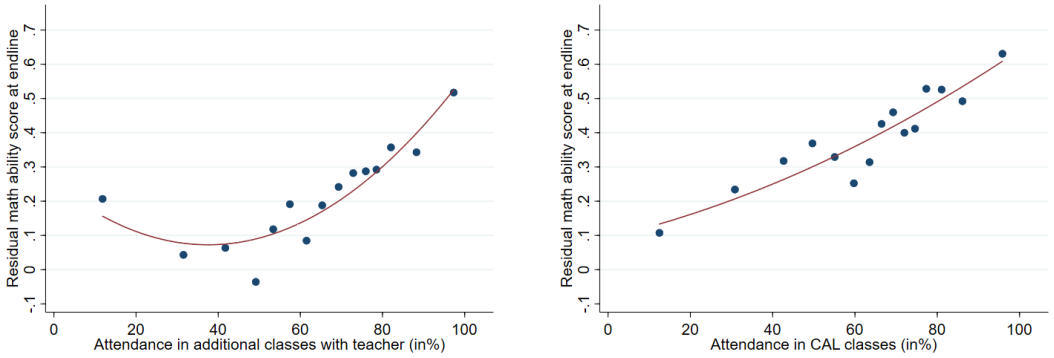
$$Att_{ics}^{T=t} = \alpha + \pi_1 T1_{cs} + \pi_2 T2_{cs} + \pi_3 T3_{cs} + \delta Y_{ics}^{BL} + X_{ics}^{'}\gamma + V_{cs}^{'}\lambda + \phi_k + \epsilon_{3ics} \quad for \quad t \in [1,2,3] \tag{2.3}$$

where $Att_i^{T=t}$ is student's i attendance rate in treatment t and takes values between 0 and 1. All other variables are defined as in the benchmark estimation equation, i.e. Equation (2.1). In the second stage, we replace the binary treatment indicators with the predicted attendance rates from stage 1, i.e. $\widehat{Att}_{ics}^{T=t}$, and estimate

$$Y_{ics}^{EL} = \alpha + \beta_1 \widehat{Att}_{ics}^{T=1} + \beta_2 \widehat{Att}_{ics}^{T=2} + \beta_3 \widehat{Att}_{ics}^{T=3} + \delta Y_{ics}^{BL} + X_{ics}^{'}\gamma + V_{cs}^{'}\lambda + \phi_k + \epsilon_{4ics}. \tag{2.4}$$

In order to interpret $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ as the treatment effects of attending all 46 additional math lessons, we have to impose two (restrictive) properties that go beyond the standard monotonicity and independence assumptions (see Angrist and Pischke, 2008; Muralidharan, Singh and Ganimian, 2019). *First*, the treatment effect needs to be homogenous across students. *Second*, the functional form between attendance and math score gains has to be linear.

[12] Regressing endline IRT scores on attendance rates (continuous between 0 and 1), baseline scores, individual and classroom controls yields the following correlations between attendance and performance: $\hat{\gamma}_{T1}$=0.40 (t-value=5.0); $\hat{\gamma}_{T2}$=0.56 (t-value=4.2); $\hat{\gamma}_{T3}$=0.55 (t-value=3.6). Including a quadratic term we get: $\hat{\gamma}^1_{T1}$=−0.53 (t-value=−1.9), $\hat{\gamma}^2_{T1}$=0.89 (t-value=3.1); $\hat{\gamma}^1_{T2}$=0.56 (t-value=1.1), $\hat{\gamma}^2_{T2}$=0.01 (t-value=0.0); $\hat{\gamma}^1_{T3}$=−0.41 (t-value=−1.0), $\hat{\gamma}^2_{T3}$=0.94 (t-value=2.1).



(a) T1: Additional math lessons with teachers (b) T2 & T3: Additional CAL-lessons

Figure 2.5: Residual endline test scores and attendance in additional math lessons.
Note: The figures present the partial correlation between individual attendance rates and residual endline test scores after controlling for baseline scores, individual and classroom characteristics. To ease readability, we aggregated individual data points into 15 bins.

Our data suggest, that these two additional assumptions may be violated and that the IV-estimates are potentially *downward* biased. *Effect homogeneity* seems questionable, since the impacts of the interventions are homogenous (both CAL treatments) or decreasing (TEACHER) in initial ability (see Section 2.4.2), even though attendance rates are positively correlated with baseline scores. Attending an additional math lesson thus had a stronger effect on low ability than high ability students. Hence, the IV-estimates might undervalue the true effect under full participation. Moreover, the functional form between attendance and ability gains appears to be (slightly) convex rather than linear, suggesting that children experienced increasing returns to attending the additional math lessons. Again this would lead to a downward bias in the reported IV-estimates.

Table 2.6 presents the IV-estimates, that can be interpreted as the (potentially downward biased) treatment effects of attending all 46 additional math lessons. Attending the full CAL program during the intervention period leads to an increase in the endline score of about 7 percentage points or 0.38σ to 0.41σ, which is about equivalent to the average student's progress in 1.2 school years.[13] This is comparable in magnitude to effects of technology-aided instructions found in India, where Muralidharan, Singh and Ganimian (2019) report average learning gains in math of 0.6σ for a 90 days attendance at CAL learning centers.

2.5 Discussion

2.5.1 Treatment Externalities

Our research design allows us to quantify spillovers on non-treated classes in program schools. As discussed in Section 2.4.1, we find positive and significant externalities: Students assigned to

[13] We refrain from presenting F-tests that formally test whether the difference between the three interventions are statistically significant because our re-randomization scheme for the within school assignment of treatments would require randomization inference, which we cannot implement in the IV-setting.

Table 2.6: IV-Estimates: Program effects with full participation

	Percent Correct		Std. IRT-Scores	
	(1)	(2)	(3)	(4)
T1: Lessons with Teacher	5.066***	4.739***	0.286***	0.269***
	(0.002)	(0.004)	(0.002)	(0.005)
T2: CAL-lessons with Supervisor	7.104***	6.859***	0.390***	0.378***
	(0.000)	(0.000)	(0.000)	(0.000)
T3: CAL-lessons with Teacher	7.517***	7.236***	0.411***	0.396***
	(0.000)	(0.000)	(0.000)	(0.000)
Kleibergen-Paap F-statistic	214.45	193.47	213.78	192.89
Adjusted R^2	0.65	0.66	0.69	0.69
Observations	2570	2570	2570	2570
Baseline Score	Yes	Yes	Yes	Yes
Individual & Classroom Controls	No	Yes	No	Yes
Stratum & Grade Level FE	Yes	Yes	Yes	Yes

Notes: p-values are based on class-level clustered standard errors and are shown in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

control classes in program schools scored 0.14σ higher in the endline assessment than students assigned to pure control classes. This effect is comparable in magnitude to the treatment effect for additional math lessons instructed by teachers. While we do not have rigorous experimental evidence to pin down the mechanisms with certainty, the data we collected from different sources allows for a discussion of what may (or may not) explain these externalities. In the following we distinguish between three broad explanations: *(i)* direct exposure of control students, *(ii)* behavioral adjustments to the experimental design, and *(iii)* social learning among peers.

Direct Exposure. We begin with examining the hypothesis that control students in program schools may have been directly exposed to one of the treatments, either by (illicitly) participating in the additional math lessons, by targeted migration and class changes, or by using CAL-software in regular lessons or at home.

To prevent direct exposure of control students to the treatments, the implementing NGO instructed contract teachers and supervisors to confine access to children that were registered as official participants. Our monitoring data shows that compliance with this directive was high, as unauthorized participation was only recorded during 6 out of about 750 unannounced visits in NGO-run math classes.

Likewise, we aimed to eliminate any incentives to change classes or schools and therefore barred students that changed into treatment classes during the school year from attending the additional math lessons. Only 38 (about 1 percent) students in our estimation sample changed classes or schools during the program and excluding these students from the estimation models leaves the results virtually unchanged.

Control students in program classes may also have been exposed to the learning software in

regular classes or at home. Again, our data suggests otherwise: The enumerators recorded computer usage in only 5 out of about 1000 regular class visits. Similarly, computer usage at home is an unlikely candidate to account for treatment externalities: According to our socio-demographic survey, only 576 students (about 18 percent) live in a household that owns a computer with internet access and this asset class is not correlated with learning outcomes in the endline assessment.

Behavioral Adjustments to the Experimental Design. We now discuss the likelihood of behavioral adjustments of teachers and students to the experimental design, namely unintended incentives to improve performance at the school level or reactive behavior of the control group.

The presence of the NGO might have incentivized school staff to make a good impression, for instance to be allowed to keep the IT equipment after the intervention or to be considered for future collaborations. We first examine this reasoning by using class cancellation and attendance rates as proxies for the effort by school staff/teachers, and then continue by testing whether a more generous supply of computer hardware raised performance in control classes. Contrary to expectations, cancellation rates appear to be slightly higher in program schools than in control schools although the difference is not statistically significant (see columns 4 & 5 in Table 2.7).[14] Similarly, student attendance rates do not point towards intensified efforts in program schools, as the estimated differences in columns (1) and (2) of Table 2.7 yield p-values larger than 0.8. Finally, we test whether a more generous furnishing of computer-labs by the NGO has pushed schools to better performances that is not necessarily reflected in attendance and cancellation rates. Consistent with the previous results, columns (3) and (4) in Table 2.8 show no relevant correlation between the number of NGO computers installed in a school and the endline performance of students in control classes.

The difference between control classes within an outside treatment schools could also be driven by a so-called *John Henry Effect*: a bias resulting from reactive behavior of the control group (e.g. Glennerster and Takavarasha, 2013). In our setting, such a bias could result either from student or from teacher behavior. As to the former, students in control classes might have worked harder to make up for their disadvantage. Similarly, teachers could have redirected resources and effort towards control classes to compensate them for their relative deprivation. For example, teachers may have given more weight to math relative to other subjects when attending control classes. If such behavior arises within treatment schools, but not in geographically (and thereby socially) separated schools, it could account for the observed treatment externalities. This mechanism has similar implications, but is distinguishable from those discussed in the previous paragraph. While the last paragraph explores the possibility of a general boost in student or teacher motivation

[14] The project could also have affected class cancellation rates directly, e.g. due to space limitations inducing the conduction of the additional lessons at the expense of regular classes. Furthermore, differences in recorded cancellation rates may (partly) be an artifact of the data collection process. To minimize transport expenses, we randomly selected entire *schools* rather than *classes* to be visited on a given day. Thus, enumerators were faced with slightly different settings in treatment and control schools: They had to record data from all classes on grades 3–6 in treatment schools, and only about one to two classes during visits to control schools. One could hypothesize that, in control schools, data collectors might have been inclined to wait patiently for the teacher to turn up (to be able to conduct the classroom observations), while, in treatment schools, they may have moved on to the next class.

Table 2.7: Externality channel (I): Motivation proxied with class attendance and cancellations

Dependent variable:	Student Attendance (%)			Class Cancellations (%)		
	(1)	(2)	(3)	(4)	(5)	(6)
Program Schools	−0.304	−0.287	−0.978	6.879	6.537	8.215
	(0.891)	(0.896)	(0.648)	(0.238)	(0.264)	(0.140)
Adjusted R^2	0.07	0.06	0.00	0.08	0.08	0.07
Observations	198	198	80	198	198	80
Control Classes Only	No	No	Yes	No	No	Yes
Classroom Controls	No	Yes	Yes	No	Yes	Yes
Grade Level FE	Yes	Yes	Yes	Yes	Yes	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: p-values are based on school-level clustered standard errors and are shown in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

across all groups in treatment schools, the *John Henry Effect* would only operate for the control group. As shown in columns (3) and (6) of Table 2.7, limiting the analysis to the control classes does not alter our conclusions: The difference in class cancellation rates between program school control classes and pure control classes is small and remains aloof from any conventional level of statistical significance. The same applies for students' attendance rates.

Social Learning among Peers. Treatment externalities may also stem from *social learning and peer effects*, as participating students could have shared their knowledge and motivation with their peers from other classes. Results in columns (1) and (2) of Table 2.8 suggest that this may have been the case: What explains part of the performance differential between within-program school control classes and pure control classes is the share of children that participated in the CAL treatments. One explanation is that the learning gains produced by CAL were (partly) passed on by the participants to their peers from non-treated classes. Another explanation for this pattern would be that hosting many CAL classes came about with a more generous furnishing of computer-labs by the NGO, which might have incentivized school staff to make a good impression with the NGO so that they could keep the equipment even after the NGO-run program expired. However, as discussed above, columns (3) and (4) in Table 2.8 show no relevant correlation between the number of NGO computers installed in a school and the endline performance of students. Hence, the interpretation that CAL beneficiaries passed on their learning gains to their peers seems more plausible than behavioral adjustments in prospect of being donated new equipment. This finding is consistent with a broad literature of peer-effects that documents how the performance of each student affects achievements of their class-mates (see Sacerdote, 2011).

Although we cannot comprehensively pin down the channels through which the observed externalities operate, *social learning among peers* is the mechanism that can be reconciled best with the data at hand. In contrast, we are confident to rule out *direct exposure* of control units

Table 2.8: Externality channel (II): Proxies for social learning and in-kind incentives

Dependent variable: Std. IRT Score	Treatment Intensity		Installed NGO computers	
CX-indicator interacted with:	All Treatments	CAL	Per Student	Total
	(1)	(2)	(3)	(4)
CX: Control Classes for Externalities	0.146**	0.135**	0.142**	0.146**
	(0.019)	(0.023)	(0.020)	(0.037)
CX: Control Classes for Externalities × Var.	0.010	0.015***	0.031	0.001
	(0.290)	(0.001)	(0.950)	(0.865)
Adjusted R^2	0.73	0.74	0.73	0.73
Observations	1279	1279	1279	1279
Individual & Classroom Controls	No	No	No	No
Baseline Score	Yes	Yes	Yes	Yes
Stratum & Grade FE	Yes	Yes	Yes	Yes

Notes: Treatment intensity defined as share of treated students in a school. p-values are based on school-level clustered standard errors and are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

to the evaluated treatments, as the monitoring data documents excellent compliance with the experimental protocol. *Behavioral adjustments to the experimental design* may unfold in many ways, which makes it difficult to track them exhaustively. We tested several potential channels operating via school teachers' and students' attendance (a proxy for motivation), but the data consistently rejects this set of claims. Considering that social learning remains as the most plausible explanation for the observed treatment externalities further strengthens the case in favor of the CAL interventions.

2.5.2 Cost-Effectiveness

Since all three interventions were assessed within the same context and framework, we can directly compare their cost-effectiveness. The bulk of expenditures comes from salaries to teachers and supervisors (65 percent for TEACHER, 41 percent for CAL + SUPERVISOR, and 51 percent for CAL + TEACHER). The two computer treatments additionally entail costs for acquiring the IT equipment, shipping it to El Salvador and maintaining it. Since our partner NGO acquired most computers as in-kind donations, the factual IT-related costs incurred by the NGO (about 18 USD per computer) provide a poor guidance for educational policy-makers aiming to implement CAL interventions at scale. To make the cost-effectiveness calculations more insightful for a generic setting, we assume costs of 200 USD per work station and an average of five years of usage time.

Based on these assumptions for the costs of the computer hardware, the cost accounting of our partner NGO, and the guidelines developed by Dhaliwal et al. (2014), we estimate the cost per child to be 44 USD for TEACHER, 43 USD for CAL + SUPERVISOR and 56 USD for CAL + TEACHER. Assuming a linear dose-response-relationship, TEACHER can thus be expected to yield a 0.35σ increase in test scores per 100 USD, while investing the same amount of money in CAL lessons would produce 0.49σ and 0.43σ, respectively. This implies that even when the computers

have to be acquired at a considerable price, the two CAL interventions outperform additional teacher-led classes in terms of cost-effectiveness. Moreover, hiring lower-paid supervisors rather than certified teachers to conduct the CAL classes might be slightly more cost-effective, as supervisors were paid only about 60 percent of a teacher's wage. Note, however, that these conclusions have to be cautiously interpreted: Not only is precision impaired by the statistical uncertainty of our impact estimates, but relative cost-effectiveness is also dependent on different contextual factors such as the local wage levels, the wage premium for certified teachers or the availability of affordable hardware.

2.5.3 The Role of Teacher Ability

Multifaceted evidence derived in our analysis points to a relatively low productivity of teachers. *First*, the difference in learning gains between within-program school control classes and classes receiving additional teacher-centered math lessons are close to zero and statistically insignificant (p-values around 0.7). Similarly, teachers do not seem to add much to the effect of computer-assisted learning lessons: The estimated impact for CAL lessons instructed by teachers is slightly higher than for CAL lessons conducted by supervisors but statistically speaking they are not distinguishable (again the p-values are in the 0.7 range). *Second*, the heterogeneity analysis shows that the productivity of teachers declines as the complexity of concepts increases: The impact of the additional math lessons instructed by a teacher is decreasing in both the grade level as well as the baseline achievement of their students. *Third*, both CAL interventions (at least marginally) outperform the additional math lessons instructed by teachers: The point estimates of the CAL interventions are consistently larger, and the impact of neither CAL + TEACHER nor CAL + SUPERVISOR decreases with student baseline performance or grade level. Hence it appears that in our setting, learning software is more productive in teaching basic math than certified teachers, especially as the complexity of the contents increases.

In order to analyze the root cause of the low productivity of teachers, we asked the instructors hired by the NGO to participate in an 90 minutes math assessment covering the primary school curriculum of grade two to grade six. Moreover, we administered the same assessment to a representative sample of regular math teachers of grade three to grade six classes which allows us to learn how the contract teachers compare to the regular teaching staff (see Brunetti et al., 2020, for details on the assessment). Figure 2.6 illustrates the main insights from this assessment: primary school math teachers in the department of Morazán insufficiently master the contents they are supposed to teach. The contract teachers hired by the NGO answered on average only 75 percent of the second and third grade questions correctly and this share declines to 54 percent for the sixth grade questions. Hence, even for the simplest questions, the average contract teacher does not meet the minimum proficiency of 80 percent correct answers as advocated in recent World Bank contributions (see Bold et al., 2017b; World Bank, 2018). This direct evidence on the lack of content knowledge conforms with our finding on the teachers' low productivity in conveying math concepts, especially those concepts pertaining to higher grades.[15]

[15]Since the teachers performed considerably worse than expected, we also validated the questions by adminis-

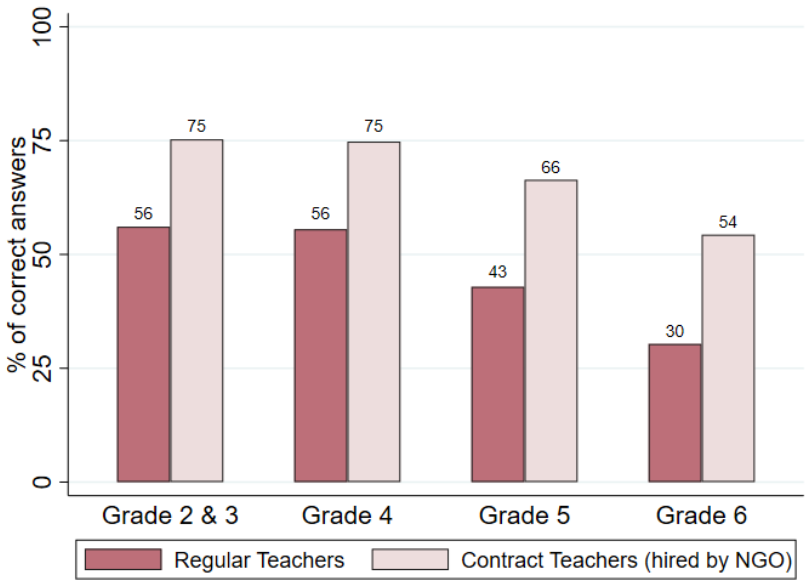


Figure 2.6: Math proficiency among regular teachers and teachers hired for additional math lessons. *Note:* The graph shows the share of correct answers on questions covering the official math curriculum of grades 2 & 3, grade 4, grade 5, and grade 6. This data was collected after the endline assessment for students in late 2018 and early 2019. The sample includes all program teachers as well as a representative sample of regular primary school teachers teaching math in grades three to six in the department of Morazán. *Source:* Brunetti et al. (2020).

These insights raise the question, whether the teachers hired for the intervention have a particularly low proficiency in math – which could explain why they are not part of the publicly employed teaching staff. Figure 2.6 suggests otherwise: Regular teachers performed considerably worse than the contract teachers, as they achieved on average only 56 percent correct answers on second and third grade questions and alarmingly low 30 percent on items pertaining to the sixth grade curriculum.[16]

In view of drawing a general conclusion for the effectiveness of additional math lessons instructed by regular teachers, the results reported in Figure 2.6 are particularly grim. The relatively low impacts found for the additional math lessons instructed by contract teachers may be even too optimistic when aiming for a scale up with regular teachers, who have on average an even lower math proficiency than the (much younger) contract teachers hired by the implementing NGO. At the same time, these results highlight the value of learning software that can compensate for the poor content knowledge of teaching staff: Earlier contributions on the value of computer-assisted learning emphasized its advantages in terms of mitigating issues of large class sizes and the challenges of "teaching at the right level" (e.g. Banerjee and Duflo, 2011; Mu-

tering the identical test (translated to German) to 16 Swiss primary school teachers, who achieved a median score of 90 percent (i.e. 45 correct answers out of 50 questions).

[16]Note that the implementing NGO administered a very short math assessment in the hiring process in order to eliminate the least qualified candidates. Moreover, the hired teachers participated in several workshops to prepare them for the teaching assignment. Since the assessment reported in Figure 2.6 was conducted *after* the intervention finished, it is likely that the NGO's selection process and the additional training for the contract teachers partly explains the pronounced differences in content knowledge between the regular teachers and the contract teachers.

ralidharan, Singh and Ganimian, 2019). While our heterogeneity analysis corroborates this line of reasoning, we further show that computer-assisted learning can help to remedy shortcomings related to low teacher ability. Since teachers are considered to be the most pivotal input to the learning production function (e.g. Hanushek, 2011; Baumert and Kunter, 2013), these findings also raise the question, how teacher quality can be improved in an effective manner.

2.6 Conclusion

Computer-assisted learning (CAL) is widely perceived as a promising approach to address the low quality of teaching in developing countries. While encouraging, previous research is inconclusive regarding the value of technology-based instruction relative to traditional teaching and and has little to say on the complementarities between teachers and learning software. The evidence presented in this paper suggests that CAL can not only produce substantial learning gains, but may also outperform traditional instruction. In our setting, this relative advantage seems to be driven by a mismatch between teacher preparation and the complexity of the concepts they have to teach: Under traditional teaching models, it seems questionable that children are able to master what their teachers fail to understand, while CAL allows them to make progress beyond their teachers' content knowledge. Overall, our findings point to an alarmingly low productivity of teachers. Not only is the effect of additional teacher-led instruction comparatively low (and might be partly if not completely attributable to treatment externalities), but poorly trained teachers also do little to improve the productivity of CAL lessons. In light of the fact that they do not master a large share of the contents they are required to teach, these results are hardly surprising.

Promoting the targeted use of computers may therefore be an attractive option for governments and NGOs operating in settings with low teacher quality. When teachers are struggling with the concepts they have to teach, learning software can be an important remedy allowing them to improve the quality of their teaching. Another approach would be to invest in the skills of teachers, for instance by offering professional development programs: Teachers may not make much of a difference when they do not master what their students are supposed to learn, but vast empirical evidence from developed countries suggests that they can matter a great deal when they are well prepared and adequately qualified (Rockoff, 2004; Chetty, Friedman and Rockoff, 2014). Hence, gaining a better understanding of how teachers' preparedness, and particularly their content knowledge, can be improved seems to be crucial for researchers as well as policy makers. Since hardly any rigorous evidence on this aspect is available (see Muralidharan, 2017; Bold et al., 2017a), we teamed up with the same implementing partner to examine, whether computer-assisted learning software can help to effectively improve the content knowledge of teachers and therewith their productivity in the classroom (see Chapter 3 of this thesis).

A. Appendices

A.1 Appendix: Additional Analysis

A.1.1 Learning Gap and Grade Level Heterogeneity in our Sample

In order to examine the learning gap and grade level heterogeneity in our sample of primary school pupils, we follow the approach by Muralidharan, Singh and Ganimian (2019) and convert the pupils' performance in the baseline assessment into a proficiency measure expressed in grade levels. As point of origin, we calculate for each participant her share of correct answers by item grade level. The score that a child obtains in our discrete proficiency measure is determined by those grade specific set of items, where the child scores at least 50 percent correct answers. To be assigned to a certain grade level, a participant needs to reach the 50 percent-threshold that corresponds with said grade level and all preceding grades. For example, a fourth grader that scored 80 percent on first grade items, 55 percent on second grade items and 40 percent on third grade items would be assigned to a second grade proficiency level. Participants answering less than 50 percent of first grade items correctly, are assigned to grade level<1.

Based on the previously specified measure, which is plotted in Figure 2.1b, we obtain a performance gap of two grades between the best and worst student in the *median* class of our sample. By construction the *mean* in the within-class performance range is lowest in third grade classes (about 1.3, i.e. the math abilities of students' within the same class cover on average 2.3 grades) and highest in sixth grade classes (about 2.4). A simple regression analysis also confirms that within-class variation is substantial, as classroom fixed effects only account for about 25 percent of the total variation at a certain grade level.

A.1.2 Attrition

In Table A.1 we examine whether the attrition at endline is correlated with the treatment status. To do so, we present results based on Linear Probability Models in columns (1) to (3), and on Logit Models in columns (4) to (6). The results unequivocal suggest, that the probability to miss the endline test did not depend on the treatment status.

Table A.1: Differences in attrition across treatments

Dependent var.: Attrition at endline	OLS			Logit		
	(1)	(2)	(3)	(4)	(5)	(6)
T1: Lessons with Teacher	0.018		0.017	0.224		0.224
	(0.302)		(0.318)	(0.298)		(0.293)
T2: CAL-Lessons with Supervisor	0.021		0.026	0.263		0.330
	(0.203)		(0.115)	(0.202)		(0.116)
T3: CAL-Lessons with Teacher	0.023		0.025	0.280		0.315
	(0.226)		(0.190)	(0.215)		(0.173)
CX: Control Classes for Externalities	0.019		0.022	0.236		0.282
	(0.307)		(0.237)	(0.298)		(0.228)
Baseline math score		−0.002***	−0.002***		−0.024***	−0.024***
		(0.000)	(0.000)		(0.000)	(0.000)
Adjusted R^2	0.00	0.01	0.01	-	-	-
Pseudo R^2	-	-	-	0.00	0.02	0.02
Observations	3528	3528	3528	3528	3528	3528

Notes: p-values (in parentheses) are based on class-level clustered standard errors. * p<0.10, ** p<0.05, *** p<0.01.

A.1.3 Method of Inference and Robustness of our Results

As explained in Section 2.4.1, we apply two methods of inference. When we assess the impact of the different treatments relative to the children in pure control classes, the reported p-values are based on class-level clustered standard errors. Inference on within program school comparisons between the different treatments (including control classes subject to externalities), however, are based on a randomization inference test statistic with 2000 random draws subject to the identical cut-off criterion as used in our re-randomization scheme.

This mixed estimation approach directly follows from our two-step randomization design (see Figure 2.2). Randomization inference is indispensable when comparing experimental groups within program schools since the underlying assignment process involved re-randomization. Conversely, selection of program schools and pure control schools was not based on re-randomization, making the use of randomization inference less critical.

While randomization inference is also preferable for assignment processes based on plain (or stratified) randomization (e.g. Young, 2019), its application is problematic in our case due to a particular feature of our study design: Out of the 162 eligible classes in pure control schools, we

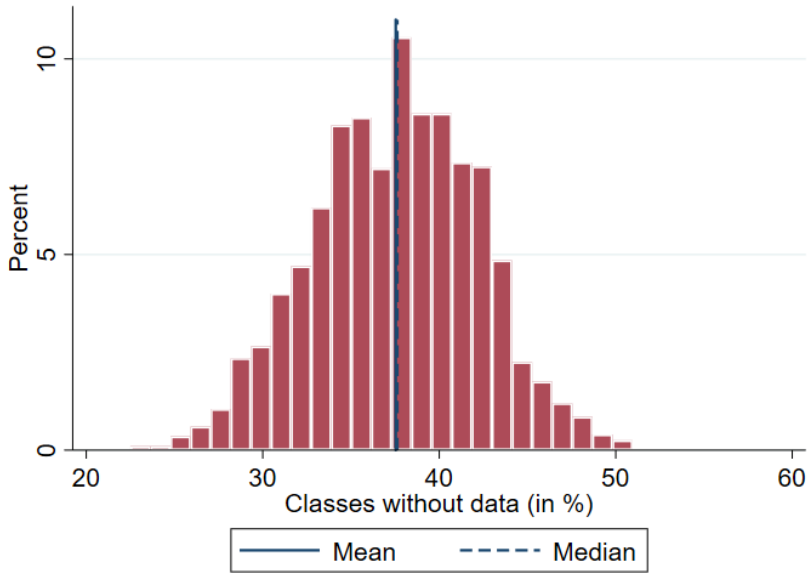


Figure A.1: Full re-randomization (incl. steps 1, 2a, and 2b) and the share of classes without data points (N=2000 draws).
Notes: This graph plots the distribution of the share of missing data points, when we conduct randomization inference by reiterating both stages of our randomization procedure. The large number of missing data points weakens the precision of our estimates, which explains why the p-values in the upper panel of Table A.2 increase by a factor of 5 to 10 compared to the p-values in Table 2.2.

only collected data for a random sample of 40 classes. Implementing randomization inference for both stages of the randomization process thus comes with the downside that each draw will contain a considerable number of classes that did not participate in the assessments. As illustrated in Figure A.1, re-iterating the full randomization procedure yields an average of 37 percent of classes without data per draw. Even though missing data points in the replication procedure create an artificial loss of statistical power, we present the respective estimates as a conservative reference point.

To assess the robustness of our results with respect to the method of inference, we report three versions of our benchmark analysis: In Table 2.2, the upper panel p-values are based on class-level clustered standard errors, while we run randomization tests in the lower panel. Table A.2 replicates these results, but inference is consistently based on class-level clustered standard errors. Finally, Table A.3 presents all results with p-values based on a full randomization tests.

Reassuringly, our main conclusion do not depend on the method of inference. When we apply traditional inference to the lower panel, as in Table A.3, changes in p-values are very small and do not show a clear pattern. And despite losing a lot of power when applying randomization inference to the upper panel, as in Table A.2, the only notable difference is, that the program externalities captured by β_{CX} turn insignificant with p-values around 0.13.

Table A.2: ITT-Estimates on the effects of the different interventions on children's math scores with p-values based on clustered standard errors

	Percent Correct		IRT-Scores	
	(1)	(2)	(5)	(6)
T1: Lessons with Teachers	2.904***	2.643**	0.165***	0.152**
	(0.005)	(0.012)	(0.006)	(0.013)
T2: CAL-Lessons with Supervisor	4.095***	3.869***	0.226***	0.214***
	(0.000)	(0.000)	(0.000)	(0.000)
T3: CAL-Lessons with Teacher	4.554***	4.328***	0.250***	0.238***
	(0.000)	(0.000)	(0.000)	(0.000)
CX: Control Classes for Externalities	2.595**	2.407**	0.147**	0.137**
	(0.011)	(0.017)	(0.013)	(0.020)
$\beta_{T4} \coloneqq \beta_{T2} - \beta_{T1} = 0$	1.191	1.226	0.061	0.063
p-value (β_{T4}=0)	(0.203)	(0.180)	(0.267)	(0.241)
$\beta_{T5} \coloneqq \beta_{T3} - \beta_{T1} = 0$	1.650*	1.686*	0.084	0.086*
p-value (β_{T5}=0)	(0.080)	(0.063)	(0.115)	(0.093)
$\beta_{T6} \coloneqq \beta_{T3} - \beta_{T2} = 0$	0.459	0.460	0.024	0.023
p-value (β_{T6}=0)	(0.606)	(0.599)	(0.637)	(0.636)
Adjusted R^2	0.66	0.67	0.69	0.70
Observations	3197	3197	3197	3197
Individual & Classroom Controls	No	Yes	No	Yes
Baseline Score	Yes	Yes	Yes	Yes
Stratum & Grade FE	Yes	Yes	Yes	Yes

Notes: p-values based on traditional clustered standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

Table A.3: ITT-Estimates on the effects of the different interventions on children's math scores with p-values based on randomization inference

	Percent Correct		IRT-Scores	
	(1)	(2)	(5)	(6)
T1: Lessons with Teacher	2.904*	2.643*	0.165*	0.152*
	(0.073)	(0.089)	(0.083)	(0.097)
T2: CAL-Lessons with Supervisor	4.095***	3.869**	0.226**	0.214**
	(0.009)	(0.013)	(0.015)	(0.018)
T3: CAL-Lessons with Teacher	4.554***	4.328***	0.250***	0.238**
	(0.006)	(0.006)	(0.007)	(0.011)
CX: Control Classes for Externalities	2.595	2.407	0.147	0.137
	(0.117)	(0.136)	(0.120)	(0.140)
$\beta_{T4} \coloneqq \beta_{T2} - \beta_{T1} = 0$	1.191	1.226	0.061	0.063
p-value (β_{T4}=0)	(0.214)	(0.194)	(0.268)	(0.244)
$\beta_{T5} \coloneqq \beta_{T3} - \beta_{T1} = 0$	1.650*	1.686*	0.084	0.086
p-value (β_{T5}=0)	(0.069)	(0.059)	(0.117)	(0.102)
$\beta_{T6} \coloneqq \beta_{T3} - \beta_{T2} = 0$	0.459	0.460	0.024	0.023
p-value (β_{T6}=0)	(0.618)	(0.615)	(0.650)	(0.653)
Adjusted R^2	0.66	0.67	0.69	0.70
Observations	3197	3197	3197	3197
Individual & Classroom Controls	No	Yes	No	Yes
Baseline Score	Yes	Yes	Yes	Yes
Stratum & Grade FE	Yes	Yes	Yes	Yes

Notes: p-values based on a two-sided randomization inference test statistic that the placebo coefficients are larger than the actual are shown in parentheses. The p-values were computed based on 2000 random draws.
* p<0.10, ** p<0.05, *** p<0.01.

A.2 Appendix: Measuring and Converting Learning Outcomes

To measure math skills of third to sixth graders, we conducted two standardized math assessments during the school year 2018. Both assessments include 60 items and were designed as follows:

1. We summarized the Salvadoran math curriculum for grades 1–6 along the three topics (a.) number sense & arithmetic, (b.) geometry & measurement, and (c.) data & probability.
2. We then mapped test items from various sources on the Salvadoran curriculum. These sources are (a.) official text books of El Salvador, (b.) publicly available items from the STAR[1] evaluations in California, (c.) publicly available items from the VERA[2] evaluations in Germany, and (d.) exercises from the Swiss textbook MATHWELT.
3. We then gathered pilot data on 180 test items answered by 600 Salvadoran pupils in October 2017 and estimated the difficulty and discrimination parameters of test questions based on *Item Response Theory* (e.g. de Ayala, 2009).

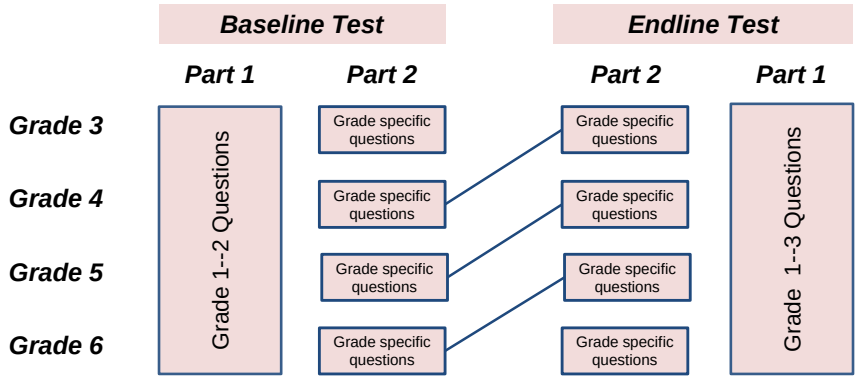
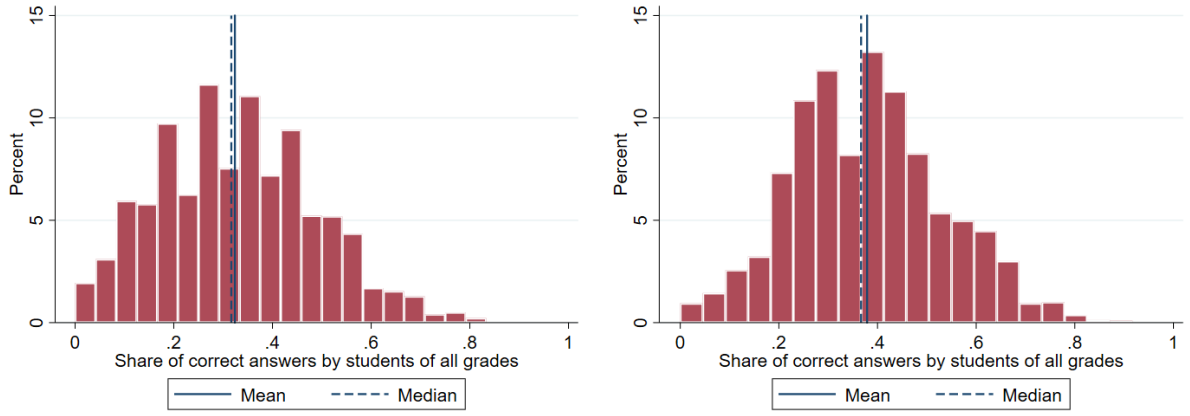


Figure A.2: Stylized illustration of the assessment design.
Note: Each part covers 30 items, adding up to 60 items per wave.

4. Finally, we designed paper and pencil maths tests using insights from step 3. The 60 items are selected such that they reflect the weighting in the official curriculum: 60–65 percent number sense & arithmetic, 30 percent geometry & measurement, 5–10 percent data & probability. Most items required a written answer, while the share of multiple choice questions varied between 10 percent and 15 percent depending on grade level. Figure A.2 illustrates how the math assessments at baseline and endline were structured and linked. Both assessments had two parts, with the first part being answered by all children independent of their grade. Moreover, the grade specific second part of 3rd/4th/5th graders in the endline assessment included many baseline questions of the 4th/5th/6th graders. This linking across grades and waves was essential to infer a commonly scaled ability score, i.e. the IRT scores.

[1] Further information on the Standardized Testing and Reporting (STAR) programme in California is available online: `www.cde.ca.gov/re/pr/star.asp` (last accessed: 14.01.2018).
[2] VERA is coordinated by the Institut für Qualitätsentwicklung im Bildungswesen (IQB), see `www.iqb.hu-berlin.de/vera` (last accessed: 14.01.2018).

Diagnostics. Table A.4 shows summary statistics on test items for each grade and wave of the assessment. In Table A.5 and Figure A.3, similar statistics are displayed for students' percentage scores. As can be seen, our test is not subject to relevant floor or ceiling effects: Hardly any students could not answer a single question on a given assessment and not a single student scored all items correctly. Similarly, only one item was not solved by anyone and no question could be answered by all students. On average, students gave correct answers to about 25-43 percent of the questions in a test booklet (column 2 in Tables A.4 and A.5). Figure A.4a shows the corresponding IRT-based test information function for the entire assessment, i.e. for all grades and waves combined (see below for details on IRT). As can be seen, our test is very informative for students across all ability levels. However, the assessment is skewed towards high difficulty levels, meaning that it allows to differentiate very precisely among high-achieving, but less precisely among low-achieving students. Ideally, the precision (or "information") of an assessment is highest around Theta = 0 where most students are located (see Figure A.4b). This implies that, on average, students should be able to answer about 50 percent of the test items. This reflects our decision to construct the assessment based on the official Salvadoran curriculum in spite of the mismatch between the curriculum and students' actual ability levels. Consequently, most of the included items could be answered by less than half of the students. While this curriculum-based approach allows for a more meaningful interpretation of results, it represents a slight loss in terms of test information. Nevertheless, sufficient questions of differing difficulty levels are covered to warrant the conclusion that, overall, our item battery provides a fairly reliable measurement instrument.



(a) Baseline test (b) Endline test

Figure A.3: Distribution of percentage scores across students

Calculating IRT-Scores. Our math assessments allows us to project all outcomes on a common ability scale by using Item Response Theory. Instead of summing up the correct answers to a total score taken to represent a person's ability, Item Response Theory proposes a probabilistic estimation procedure. Ability is then viewed as a latent variable influencing the responses of each individual to each item through a probabilistic process: The higher a person's ability and

Table A.4: Item diagnostic: The distribution of correct answers across items

	Share of correct answers across items (in %)					
a. Baseline	Minimum	Mean	Median	Maximum	Share 0%[a]	Share 100%[b]
3rd Graders	0.4	24.9	18.3	87.3	0.0	0.0
4th Graders	2.4	30.9	25.5	94.2	0.0	0.0
5th Graders	0.4	34.9	26.6	96.6	0.0	0.0
6th Graders	0.4	38.7	27.4	96.4	0.0	0.0
b. Endline	Minimum	Mean	Median	Maximum	Share 0%[a]	Share 100%[b]
3rd Graders	0.9	34.1	23.5	95.8	0.0	0.0
4th Graders	0.5	36.0	31.0	98.0	0.0	0.0
5th Graders	0.0	38.9	32.3	98.8	1.7	0.0
6th Graders	1.3	42.6	37.2	98.9	0.0	0.0

Notes: The share of correct answers bases on those students that participated in both assessments, and hence constitute the main estimation sample. *a. Share 0%:* This column displays the share of items with zero correct answers. *b. Share 100%:* This column displays the share of items that were answered correctly by all test-takers.

Table A.5: Item diagnostic: The distribution of percentage scores across students

	Percentage score across students (in %)					
a. Baseline	Minimum	Mean	Median	Maximum	Share 0%[a]	Share 100%[b]
3rd Graders	0.0	24.9	21.7	78.3	0.9	0.0
4th Graders	0.0	30.9	28.3	83.3	0.6	0.0
5th Graders	0.0	34.9	35.0	80.0	0.2	0.0
6th Graders	1.7	38.7	38.3	80.0	0.0	0.0
b. Endline	Minimum	Mean	Median	Maximum	Share 0%[a]	Share 100%[b]
3rd Graders	0.0	34.1	33.3	83.3	0.8	0.0
4th Graders	0.0	36.0	35.0	91.7	0.2	0.0
5th Graders	0.0	38.9	38.3	81.7	0.1	0.0
6th Graders	0.0	42.6	40.0	90.0	0.1	0.0

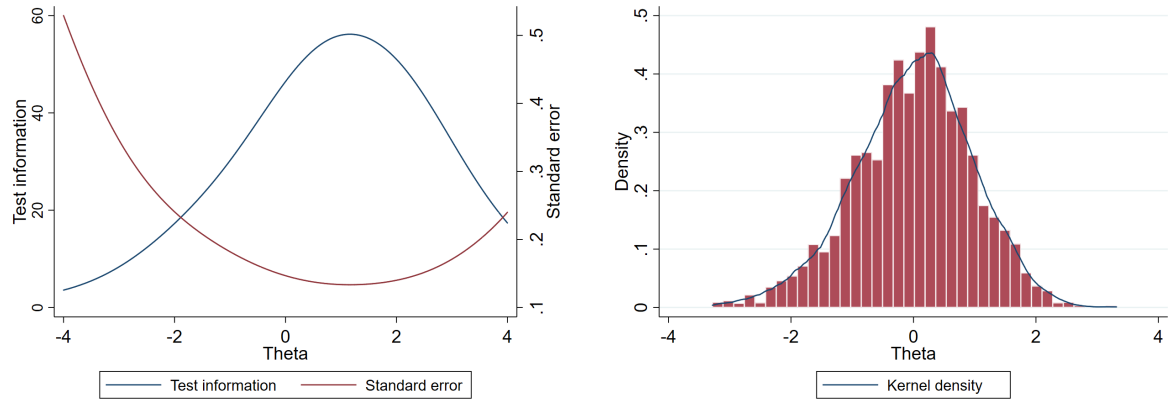
Notes: The distribution of percentage scores bases on those students that participated in both assessments, and hence constitute the main estimation sample. *a. Share 0%:* This column displays the share of students that answered zero questions correctly. *b. Share 100%:* This column displays the share of students that answered all questions correctly.

the lower the difficulty of a particular test item, the higher the probability of a correct answer. In the simplest form of the model, the probability that individual i succeeds on item j can be expressed by the following function:

$$Pr(success_{ij}|b_j, \theta_i) = \frac{exp\{a(\theta_i - b_j)\}}{1 + exp\{a(\theta_i - b_j)\}}$$

with θ_i denoting the ability of student i, and b_j representing the difficulty of item j.

In this so-called *one-parameter model*, the probability that an individual correctly solves



(a) IRT-based test information function (b) Distribution of student abilities (Theta)

Figure A.4: Test information figure and distribution of students' abilities.

a particular item is thus a logistic function of the distance between the ability level of that individual and the difficulty of the item. Ability levels for each person and difficulties for all items can be computed through joint maximum likelihood estimation. IRT has many advantages over classical test theory. It tends to produce more reliable ability estimates, allows to link the scores of different individuals in different tests through overlapping items, and can help to better understand and improve the quality of a test (e.g. de Ayala, 2009).

As illustrated in Figure A.2 a selection of items overlap *(i)* between the baseline and endline assessments and *(ii)* across test booklets of different grades within an assessment wave. This allowed us to project the performance in the baseline and endline assessment onto a common scale through the estimation of an IRT one-parameter model. This procedure yields for every student i two ability estimates, namely one for the baseline assessment, i.e. θ_i^{BL}, and one for the endline assessment, i.e. θ_i^{EL}. The latter serves as outcome variable in the regression models that are labeled with "IRT-Scores".

Converting IRT-Scores to School Year Equivalents. To allow for an intuitive interpretation, IRT scores can be represented as school year equivalents. For this purpose, we re-scale ability estimates based on between-grade ability differences among pure control students at the time of the endline assessment; that is they are divided by the average difference between adjacent grades, which we calculated to be 0.31. That means, that the average ability difference between third and fourth graders, fourth and fifth graders, and fifth and sixth graders in October 2018 equaled 0.31.

The estimated program effects can then be interpreted as a proportion of the children's average progress during one school year. Note, however, that ability differences between grades do not only represent what children learn in their regular math classes at school but also reflect age-based cognitive development, learning at home or spillovers from other subjects (e.g. literacy or science).

3 How Effective Are Computer-Based Teacher Training Programs? Experimental Evidence from El Salvador*

3.1 Introduction

Basic education is universally recognized as a key requirement for economic development, strong institutions, human flourishing, and therefore a central pillar of development cooperation. However, the quality of schooling often remains alarmingly low in developing countries, despite impressive improvements in the accessibility of primary education (see e.g. World Bank, 2018). Teachers are arguably the most essential input factor to the educational production function (Baumert and Kunter, 2013; Hanushek, 2011). Indeed, several studies from the US and developing countries document a sustainable impact of teaching quality on student learning (Hanushek, 1992; Rockoff, 2004; Araujo et al., 2016; Bau and Das, 2017). Barber and Mourshed (2007, p.43) conclude from their study of high-performing school systems that "the quality of an education system cannot exceed the quality of its teachers" and that "the only way to improve outcomes is to improve instruction". Teachers are not only one of the key inputs for quality education, payments to the teaching staff constitute by far the largest share of educational expenditures. Teacher salaries in Sub-Saharan Africa amount to 70 percent of educational expenditures and 12 percent of total government expenditures (Bold et al., 2017b). In Latin America, they consume almost 4 percent of the region's GDP (Bruns and Luque, 2014).

Yet, teachers are often strikingly ineffective in many low- and middle-income countries. While previous research on the underlying mechanisms has mainly focused on factors such as incentives or pedagogical practices (for an overview, see Kremer, Brannen and Glennerster, 2013; Glewwe and Muralidharan, 2016), relatively little attention has been payed to the role of teachers' content knowledge. Recent evidence from African countries and India suggests that many primary school teachers do not possess sufficient mastery of the concepts they are supposed to teach (Bold et al., 2017a; Sinha, Banerji and Wadhwa, 2016). Our findings from a regionally representative math assessment in El Salvador confirm this pattern (Brunetti et al., 2020): On average, teachers could answer less than 50 percent of the assessment questions covering materials from grades two to six – i.e. materials they are supposed to teach. Only 14 percent of these Salvadoran primary school teachers pass the minimum proficiency threshold for effective teaching defined

*This chapter is joint work with Aymo Brunetti, Konstantin Büchel, Martina Jakob, Ben Jann and Christoph Kühnhanss.

by Bold et al. (2017a). Non-experimental evidence from Peru and several African countries further indicates that subject-related teacher skills have a sizable impact on students' learning (Bietenbeck, Piopiunik and Wiederhold, 2018; Metzler and Woessmann, 2012; Bold et al., 2019). For example, Bold et al. (2019) find that deficiencies in teachers' content knowledge account for 30 percent of the shortfalls in student learning relative to the curriculum, and about 20 percent of the cross-country difference in student performance in their sample. Based on these findings, one may conclude that the low quality of education is, to a considerable degree, the result of a *teaching crisis* (for a similar conclusion see Molina et al., 2018). Without joint efforts, this situation is likely to reproduce itself: Many of today's poorly qualified teachers will continue teaching for years to come and consequently shape tomorrow's teachers. Despite these facts, we know strikingly little as to how teacher content knowledge can be effectively improved.

Based on a field experiment in El Salvador, this study provides novel evidence on the effectiveness of a content-based teacher training program in a developing country. Throughout the rural department of Morazán, 175 primary school teachers were randomly assigned either to participate in a five-month in-service program or to a control group. The program combines self-study elements based on a well-proven computer-assisted learning (CAL) software with group workshops, where the basic math concepts covered in the self-study modules are subsequently discussed by handpicked expert teachers. All teachers were administered a math assessment covering the curriculum of grades two to six as well as a survey prior to randomization and participated in an endline assessment after the end of the intervention allowing us to assess the short-term impact of the program.

A particularly promising feature of the training program we evaluate in this study is the use of CAL software. CAL is a popular approach to improve students' learning outcomes in the developing world (e.g. Banerjee et al., 2007; Muralidharan, Singh and Ganimian, 2019; Snilstveit et al., 2015). In Chapter 2 of this thesis, we document substantial learning gains for students in El Salvador as a result of additional software-based math lessons. However, providing and maintaining computer labs for schools can be costly. Building on the success of CAL in improving students' learning levels, we want to explore its potential to raise teacher subject knowledge. This innovative approach has a key advantage over student-centered initiatives: Instead of one computer for each student, only one computer per teacher is needed. Considering that teachers instruct hundreds or even thousands of children in the course of their careers, sustainably improving their skills might have a vast multiplier effect and thus be a highly cost-effective strategy to boost student learning. Hence, a technology-based approach targeting teachers seems to be a promising, though currently unexplored way to address learning shortfalls in the developing world. Since we have implemented a CAL for students program in the very same context (see Chapter 2 of this thesis), we are able to compare the cost-effectiveness of the "CAL for teachers" version directly to the "CAL for students" version. Finally, the learning software used in this study, "Khan Academy", is freely available in more than 30 languages. Hence, CAL for teachers, has the potential not only to be a very cost-effective intervention but as well has the potential to easily reach teachers in almost any region.

While there is recent experimental evidence on the impact of teacher-oriented interventions (see Popova et al., 2018, for an overview), previous studies focus on either purely pedagogical or multifaceted programs, and thus, do not allow to isolate the effect of the subject-matter training component. In our extensive literature review we have identified 31 studies that estimate the impact of teacher-oriented programs with rigorous methods. Of these 31 studies, 22 examine interventions without a content knowledge component, while 9 studies examine mixed interventions containing some content knowledge elements. We are not aware of research that examines an intervention focusing exclusively on teacher content knowledge (see Appendix A.3 for our literature review). For example, Cilliers et al. (2019) experimentally evaluate teacher-oriented interventions with a coaching approach and find substantial effects on student learning in literacy of about 0.2 standard deviations (σ). However, this study focuses on pedagogy, which – at best – affects participants' content knowledge only indirectly. Similarly, Zhang et al. (2013) analyze the impact of a randomly assigned three-week training in English language for Chinese teachers, which combined training in language skills with sessions on pedagogical techniques. They report insignificant impact estimates for both teachers and students concluding that the intensity of the program might have been insufficient. The most closely related study evaluates an intervention targeting 14 secondary school math teachers in Johannesburg. Using matching techniques, Pournara et al. (2015) estimate the effect of a teacher training program, which mainly focused on refreshing content (75 percent weight) but also featured pedagogical training (25 percent weight). While they report promising effects of the intervention on students' math skills, their non-experimental identification strategy and small sample size preclude strong conclusions. Hence, little is known about the effectiveness of content-based training programs – or as Bold et al. (2017a, p. 202) have put it: "Unfortunately, there are few, if any, well-identified studies on how to effectively improve teacher knowledge and skills and the impact thereof." Similarly, Muralidharan (2017, p. 377) concludes that "we still do not have good experimental evidence on many important questions including [...] teacher-training programs [...] This in turn means that the evidence base for the design of teacher training programs is also very limited." This study addresses this gap in the literature.

Our analysis establishes three sets of findings. *First*, we find that the five-month program significantly improves teachers' math content knowledge. Program teachers scored 0.28σ higher than teachers in the control group, i.e. the program increases teachers' math content knowledge by 12 percent.[1] *Second*, the heterogeneity analysis shows that teachers make relatively greater progress in areas where they struggled most in the baseline assessment and that the effect varies considerably by teacher characteristics. Program teachers outperformed their peers from the control group by 0.34σ (i.e. an increase of 25 percent compared to the control group) on grade five questions and by 0.36σ (25 percent) on grade six questions. Further, the program is more effective for younger teachers and teachers performing better at the baseline assessment. Teachers under the age of 40 years score 0.52σ (19 percent) better than their peers from the control group. The program appears to be ineffective for teachers above 50 years of age (0.03σ, p-value=0.879). *Third*, we find that the program is cost-effective, if the effect on teachers' content knowledge

[1] These results are based on intent-to-treat estimates with a teacher compliance rate of 75 percent.

is persistent. During the first year, the program's cost-effectiveness (0.22σ per 100 USD) is clearly inferior to the cost-effectiveness of a comparable CAL-intervention targeting students (0.49σ per 100 USD). Depending on the assumed annual depreciation in the gained teacher content knowledge, the cost-effectiveness of the teacher training gradually improves over time and eventually overtakes CAL for students after four years, if we assume an annual depreciation of the newly gained content knowledge of 40 percent.

This study makes several contributions to the literature on educational interventions in developing countries. Most importantly, it is the first experimental study that evaluates an intervention focusing exclusively on content knowledge of primary school teachers. Results show that computer-based teacher training significantly improves the math content knowledge of teachers. Teachers are at the center of a high-quality education and, yet, recent evidence on the alarmingly poor content knowledge of teachers from different developing countries shows how inadequately teachers are prepared for their job. Therefore, it is decisive to address the largely understudied question of how to advance teaching quality in a schooling system staffed with poorly qualified teachers.

Second, our study also connects to the literature on aging and work performance. Ng and Feldman (2008) find a negative relationship between age and performance in training programs in general and conclude that programs with a focus on technical skills (e.g. computer training) may not be effective for older participants. Furthermore, older persons are often more reluctant to engage in new skill training (Kanfer and Ackerman, 2004). Finally, cognitive ability to learn new skills may decline with age. While a decline in cognitive abilities starts usually only in the seventies (Schaie and Willis, 2013), fluid intellectual skills,[2] which are particularly important for acquiring new competences (e.g. computer skills), tend to decline earlier than other intellectual abilities (Kanfer and Ackerman, 2004). Our heterogeneity analysis is in line with these findings and shows that the program is particularly effective for young teachers, while it is ineffective for older teachers. In addition, young teachers are at an earlier stage of their careers and will therefore instruct more students in the future than older teachers. Therefore, it makes sense to focus on younger teachers from a cost-effectiveness perspective – in particular for technology-based training programs.

Third, this study shows the importance of the persistence of a treatment effect for the cost-effectiveness. This is in line with Cilliers et al. (2020), who show that the cost-effectiveness of a pedagogical training and coaching intervention increases by 50 percent when also learning gains of the second cohort are included. Cilliers et al. (2020) argue that many successful, low-cost education interventions, such as parent information campaigns or accountability measures, are unlikely to have a persistent impact on future cohorts of students after the program ends, while relatively expensive teacher trainings are likely to also affect future cohorts. Therefore, the cost-effectiveness of investments in human capital, such as teacher training, is often underestimated

[2]Fluid intellectuals skills are associated with working memory, abstract reasoning and processing novel information (Kanfer and Ackerman, 2004). Our program focuses on math and is technology-based which is new to many participants. Therefore, the program requires specifically these fluid intellectual abilities in order to make progress.

when costs and benefits of future students are not considered. We are able to compare the cost-effectiveness of CAL for teachers with CAL for students in the very same context and show the importance of considering future costs and benefits. While CAL for students is more cost-effective in the first year, CAL for teachers becomes more cost-effective in later years, depending on the persistence of the effect on teachers.

3.2 Context and Intervention

3.2.1 Context

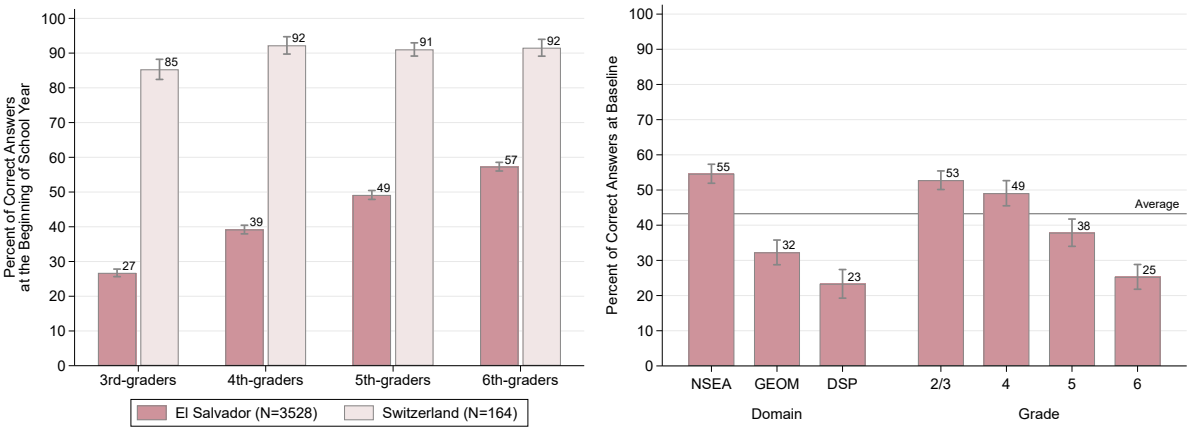
The intervention is situated in the department of Morazán in El Salvador. El Salvador's net primary school enrollment rate is estimated at roughly 80 percent and is thus below the average of lower middle-income countries.[3] Morazán is located in the northeastern part of the country and belongs to the poorest regions in El Salvador. The illiteracy rate of Morazán is considerably high with 20 percent (DIGESTYC, 2018). In 2019, Morazán's average score in the "PAES" examination, a standardized test administered to all secondary school students throughout the country, was slightly above the national average and the department ranked seventh among the 14 Salvadoran departments (MINED, 2019).

Findings in Chapter 2 (see Figure 3.1a) with more than 3500 primary school students throughout Morazán point to strikingly low learning levels: Sixth graders, who by then should have attended more than 1000 math lessons, only answered 57 percent of first and second grade math questions correctly and this share declined to 14 percent for math questions at their current level. The math proficiency among third, fourth and fifth graders can be characterized along the same lines. The identical assessment translated to German was administered to a convenience sample of 164 primary school students in Switzerland. With 85 to 92 percent of correct answers Swiss pupils clearly outperform their peers from El Salvador.

The public school system suffers from many shortcomings that partially account for the low quality of education and which are typical for many low and middle-income countries. While class size is relatively low with an average of 28 pupils per teacher in El Salvador and 19 pupils per teacher in Morazán, heterogeneity within grades and an overambitious curriculum often prevent adequate teaching. For instance, the average sixth grader in Morazán lags more than three grades behind the official curriculum and in most classes students' math ability spans over three or more grades. In addition to this challenging environment with heterogeneous classes and an overambitious curriculum, teachers are often not motivated or simply not well prepared for the job. Teacher absenteeism is common, as average class cancellation rates of 25 percent indicate, and teachers follow outdated pedagogical techniques that focus on memorization and recitation (see Chapter 2).

As baseline results in Figure 3.1b show, teachers do not master the math contents they are

[3] Enrollment statistics according to the *World Development Indicators* provided online by the World Bank: `https://data.worldbank.org/indicator` (last accessed: 21.06.2019)



(a) Share of correct answers on 1st/2nd grade math questions among Salvadoran and Swiss pupils

(b) Share of correct answers among Salvadoran teachers at baseline

Figure 3.1: Students' and teachers' math content knowledge

Note: The results in Figures 3.1a and 3.1b cannot be compared because we used mutually exclusive items to assess teachers' and students' math skills. Teacher results are divided into the domains Number Sense & Elementary Arithmetic (NSEA), Geometry & Measurement (GEOM) and Data, Statistics & Probability (DSP). 95% confidence spikes in both figures. *Source*: Figure 3.1a, Chapter 2 and Figure 3.1b, teacher baseline assessment in September 2018 resp. March 2019.

supposed to teach. On average teachers scored only alarming 43 percent on an exam covering math contents of grades two to six of the official Salvadoran curriculum.[4] Knowledge gaps are most apparent in *Data, Statistics and Probability* (DSP) with 23 percent correct answers and *Geometry and Measurement* (GEOM) with 32 percent correct answers. Teachers performed better with questions on *Number Sense and Elementary Arithmetic* (NSEA), yet even there they only answered 55 percent correctly. When we group questions by grade level, we find that the share of correct answers declines from 53 percent for second and third grade questions to 25 percent for sixth grade questions. Bold et al. (2017a) argue that a teacher masters the student curriculum if she or he is able to mark at least 80 percent of the questions of a assessment covering the curriculum correctly. The average teacher of Morazán does not reach this minimum proficiency threshold in any domain or grade level, suggesting that the low learning outcomes of students may be largely explained by the insufficient teacher content knowledge.

3.2.2 Intervention

In this context we evaluate an intervention that aims at improving teachers' content knowledge. To this end, we cooperate with the NGO Consciente, which implemented an in-service teacher training program between April and August 2019. The intervention targets 87 primary school math teachers (teaching grades three to six) and consists of two elements: *(i)* computer-assisted self-studying at home, and *(ii)* monthly revision workshops. Both elements focus, as far as possible, exclusively on teachers' math content knowledge and not on pedagogical aspects.

[4]Note, that the teacher sample of this study is not representative for Morazán. However, Brunetti et al. (2020) find that the average teacher of a representative sample of Morazán only scores marginally higher (47 percent correct answers) than the average teacher of the sample considered in this study.

Self-Studying. Drawing on the extensive contents of the learning software *Khan Academy*, 16 study modules covering selected contents of the Salvadoran primary school math curriculum were designed. In accordance with the official curriculum, the main focus of the training program is on number sense & arithmetic, but concepts pertaining to geometry & measurement and data, statistics & probability are covered as well. In an initial meeting, participants received a laptop equipped with the learning software, which allows offline access to the selected learning videos and exercises from *Khan Academy*.[5] Teachers had to complete one module per week, corresponding to a workload of four to eight hours, and then took a short assessment administered by the software. Since module completion had to be accomplished outside working hours, teachers received monetary compensation for it. Payments were conditional on the completion of the assigned exercises and videos (85 percent) and on test performance at the end of each module (15 percent). For the first module, teachers could earn up to 18.00 USD. In terms of Salvadoran wage levels, this roughly corresponds to a regular teacher salary for half a workday. With each subsequent module, maximum compensation increased by 0.50 USD yielding 25.50 USD for the final assignment. Thus, payments are determined by the following formula

$$C = \sum_{i=1}^{16} (0.85 * E_i + 0.15 * T_i) * (18.00\ USD + 0.50\ USD * (i-1)), \tag{3.1}$$

where C is the total compensation, E_i represents the share of completed exercises (and videos) in module i and T_i denotes the share of correct answers on the assessment in module i. Hence, maximal compensation is 348 USD per teacher which amounts to roughly 50 to 60 percent of a teacher's monthly salary. Throughout the intervention, the software monitored teachers' progress and participants received regular reminders and individual support in case of technical problems (e.g. if they were not able to log in their account).

Monthly Workshops. At the monthly workshops, participants submitted the work they accomplished on the previous four self-studying modules. While teachers took part in a tutoring session, their learning progress was evaluated to determine the compensation they were to receive. During the workshops, expert teachers recapitulated key concepts and addressed teachers' questions. Meetings were scheduled for half a day and organized in four separate groups. Since the courses took place during work hours, teachers were only compensated for travel expenses and did not receive further remuneration for attending them. Note, however, that receiving compensation for completing the self-studying modules was conditional on workshop participation. Hence, teachers were indirectly incentivized to attend. High completion rates for modules (75 percent on average) and attendance rates at workshops (85 percent on average) show that teachers complied well with the protocol (see Section 3.3.4).

[5] Technically, the offline access to the *Khan Academy* content is carried out via *Kolibri*, see www.learningequality.org/kolibri/ (last accessed: 21.06.2019). For *Khan Academy* see www.khanacademy.org/ (last accessed: 21.06.2019).

3.3 Research Design

This study is set up as a randomized controlled trial (RCT), with applicants to the teacher training program being randomly assigned to either the treatment or the control group. Before the start of the intervention all candidates took an unannounced baseline assessment.[6] After the completion of the five-month training program for the treatment group, study participants were administered a follow-up assessment in September 2019. The endline assessment in September 2019 was remunerated generously with 40 USD to incentivize participation. A comparison between the two groups allows for a causal identification of the effect of the program on teacher content knowledge.

3.3.1 Sampling and Randomization

All public primary schools with students in grades three to six in Morazán serve as the starting point for the sampling process. For implementation purposes, the 49 smallest schools with fewer than a total of 15 students in grades one to six were excluded, resulting in a target population of 253 schools. The NGO sent out invitations to all grade three to six math teachers in eligible schools. Prospective participants were asked to subscribe through a registration sheet, an online form or a phone call. In total, 313 teachers from 175 different schools applied for the program.

In a next step, all applicants were invited to an information meeting, where the training program and the design of the study were presented. In particular, teachers were informed about the random element in the admission procedure and the data that would be collected as part of the evaluation. Applicants willing to participate in the study were then administered a baseline assessment including a math assessment and short survey.[7] Based on the results of the baseline assessments, the worst-performing applicant of every school was selected for participation. Note, however, that this part of the sampling procedure was not communicated to applicants to avoid misaligned incentives during the assessments. At the end of this procedure 175 teachers from 175 different schools across Morazán remained in the sample.

Finally, the 175 pre-selected teachers were randomly assigned to either the control group (88 teachers) or the treatment group (87 teachers). To enhance the efficiency of the estimates, randomization was stratified by the teachers' baseline score and gender. For this purpose, teachers were grouped by performance quartiles using the baseline assessment and by gender so that we obtained eight strata. Even though we randomized at the teacher level, the pre-selection left only one teacher per school in the sample. This prevents the risk of biased estimates due to spillover effects within schools.

[6] A sub-group of the applicants took the math assessment in the context of a diagnostic pilot study in September 2018 (see Brunetti et al., 2020). In March 2019, the same assessment was administered to all other applicants. Table 3.1 further below shows that the proportion of teachers who took the exam in September 2018 (instead of March 2019) does not differ significantly between the control and the treatment group (see Table 3.1). In both cases, the assessment was unannounced.

[7] As mentioned, some applicants had already taken the same assessment during a representative teacher assessment six months earlier. This subgroup of teachers only received information about the program.

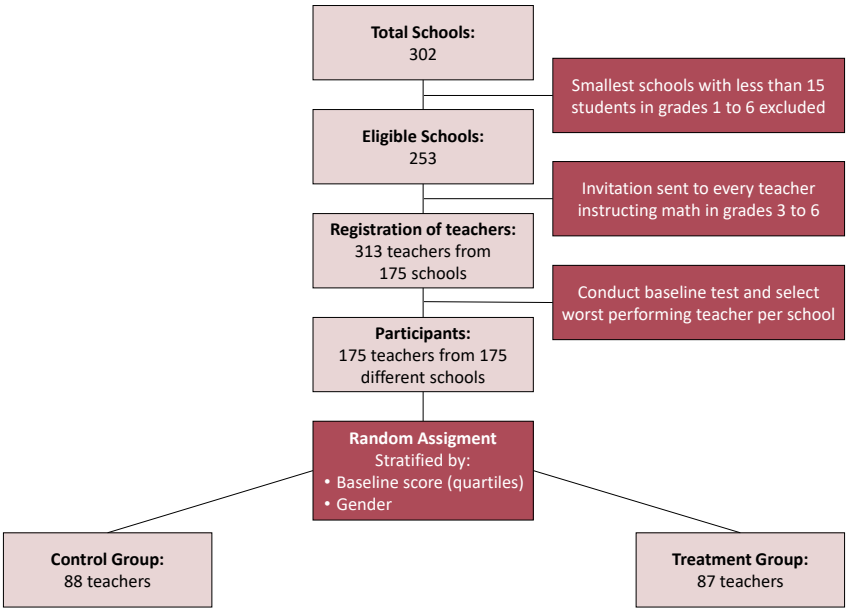


Figure 3.2: Sampling and randomization scheme

3.3.2 Data and Measurement

The data for this study is based on four sources: *(i)* base- and endline teacher assessments, *(ii)* survey on teacher characteristics, *(iii)* administrative data on school characteristics provided by the Ministry of Education and *(iv)* monitoring data collected during the intervention.

Teacher Assessments. To assess teachers' math skills before and after the intervention, we proceed as follows: First, we summarized the Salvadoran math curriculum for grades two to six along the three topics *Number Sense & Elementary Arithmetic* (NSEA), *Geometry & Measurement* (GEOM), and *Data, Statistics & Probability* (DSP). For the assessments, we then map test items from various sources on the Salvadoran curriculum. These sources include official text books of El Salvador, publicly available items from the STAR[8] evaluations in California, publicly available items from the VERA[9] evaluations in Germany, and publicly available items from the SAT[10] assessments in Great Britain. Finally, we design paper and pencil math assessments including a total of 50 questions on materials from grade two (∼6 items) and grades three to six (between 10 and 13 items) reflecting the official national curriculum. The assessments cover questions from NSEA (∼30 items), GEOM (∼15 items), and DSP (∼5 items) and have to be completed in 90 minutes. The relative weighting of the three main domains emulates the weighting in the national primary school math curriculum. To make sure that questions are

[8]Further information on the Standardized Testing and Reporting (STAR) program in California is available online: `www.cde.ca.gov/re/pr/star.asp` (last accessed: 17.06.2019).

[9]VERA is coordinated by the Institut für Qualitätsentwicklung im Bildungswesen (IQB), see `www.iqb.hu-berlin.de/vera` (last accessed: 17.06.2019).

[10]SAT is an acronym for *standardized assessment tests* coordinated by the UK's Standards and Testing Agency, see `https://www.gov.uk/government/organisations/standards-and-testing-agency` (last accessed: 26.06.2019).

suitable for the Salvadoran context, assessments are reviewed by local teaching experts and the local education ministry.

Based on these assessments, we use two different main outcome measures at the teacher level. The share of correctly answered questions and standardized test scores with $\mu = 0$ and $\sigma = 1$. We further compute domain-specific test scores for NSEA and GEOM & DSP as well as grade-level specific test scores. Assessment diagnostics can be found in Appendix A.2.

Survey on Teacher Characteristics. The baseline math assessments were preceded by a short questionnaire in order to collect basic information on teachers' socio-demographic characteristics as well as information on the classes they instruct. Variables in Table 3.1 panel B stem from this survey.

Administrative Data. In regular administrative surveys, the Salvadoran Ministry of Education collects information on a wide range of school characteristics. Data from our own assessments can thus be complemented with official school level indicators provided by local education authorities.

Monitoring Data. Monitoring data on teachers' progress in the self-study modules is regularly collected by the software. Further, we collected data on the attendance in the monthly workshops. This monitoring data allows us to compute an overall compliance rate of teachers.

3.3.3 Descriptive Statistics and Balance at Baseline

Table 3.1 presents summary statistics for teacher and school characteristics collected at baseline as well as the attrition at the endline. Panels A and B present *teacher characteristics*. Across both groups, the average score attained by teachers was 43 percent correct answers, while the median teacher answered only 38 percent of the questions correctly. Moreover, the average teacher is 44 years old, graduated in 2000 and has to commute 66 minutes to reach his school. 64 percent of the teachers are female, 7 percent are specialized teachers giving only math lessons. 75 percent have completed 2 to 3 years of teacher's formation (Profesorado) and 23 percent have obtained a Bachelor's degree (Licenciatura). The attrition at endline was 6.3 percent (i.e. 11 teachers).

As to *school characteristics* (panel C), students have access to computers in 42 percent of schools in our sample, 86 percent of the schools are located in rural regions and the average school is 49 driving minutes away from the departments' capital. Activities of criminal gangs on school premises are reported for 10 percent of the schools in our sample. The average school is equipped with about 26 percent of potential items from a list containing 12 standard technical devices such as printers, copy machines or overhead projectors (i.e. equipment index) and about 27 percent of all items on a similar list of facilities like bathrooms, a cafeteria or a sports pitch (i.e. infrastructure index).

Table 3.1: Balance at baseline and absence rate during endline assessment

	Treatment Group	Control Group	p-Value
Panel A: Baseline Math Scores (N=175)	(1)	(2)	(3)
%-Share Correct Answers	43.26	43.27	1.00
	(2.94)	(2.07)	
Standardized Math Score	0.00	0.00	1.00
	(0.15)	(0.11)	
Baseline Test Group: March 2019[a]	0.32	0.36	0.56
	(0.07)	(0.05)	
Panel B: Sociodemographics (N=175)			
Age	44.36	43.78	0.64
	(1.21)	(0.85)	
Female	0.64	0.64	0.92
	(0.07)	(0.05)	
Highest Degree[b]	2.22	2.25	0.65
	(0.07)	(0.05)	
Years since Highest Degree	19.77	18.82	0.44
	(1.22)	(0.86)	
Math Specialization[c]	0.08	0.06	0.54
	(0.04)	(0.03)	
Travel Time to School (min.)	58.80	72.28	0.17
	(9.86)	(6.95)	
Absence at Endline (%)	6.90	5.68	0.74
	(3.69)	(2.60)	
Panel C: School Information (N=175)			
Computer Access Students	0.46	0.38	0.26
	(0.07)	(0.05)	
Equipment Index[d]	0.27	0.26	0.63
	(0.03)	(0.02)	
Infrastructure Index[e]	0.27	0.27	0.89
	(0.02)	(0.02)	
Gang Activities on School Grounds	0.11	0.09	0.60
	(0.05)	(0.03)	
Rural	0.86	0.85	0.85
	(0.05)	(0.04)	
Travel Time to Department Capital (min.)	47.70	50.22	0.56
	(4.26)	(3.00)	

Notes: This table presents the mean and standard error of the mean (in parenthesis) for several characteristics of teachers and schools by treatment status. The sample consists of all teachers of the control and treatment group (N=175). Column 3 shows the p-value (based on two-sided t-tests) from testing whether the mean is equal across control and treatment group. *a*: Dummy variable indicating whether the teacher took the baseline in September 2018 (0) or March 2019 (1). *b*: Four categories: 1=bachillerato (high school), 2=profesorado (2–3 years of tertiary education), 3=licenciatura (5–6 years of tertiary education, equiv. to a bachelor's degree) and 4=maestria (equiv. to a master's degree). *c*: Teaching math only. *d*: For each school a list covering twelve technical equipments is available. The index refers to the share of items on this list that a school possesses. *e*: For each school a list covering eleven facilities is available. The index refers the share of facilities on this list that a school possesses.

Random assignment was successful in the sense that teacher characteristics appear to be well-balanced, as the *p-values* presented in Table 3.1 remain aloof from conventional levels of statistical significance. Most importantly, the average pre-treatment content knowledge of teachers is almost identical across the treatment and control group. Further, the probability to miss the endline assessment is not correlated with the treatment status as the the last row of panel B in Table 3.1 as well as the additional analysis in Appendix A.1.1 shows.

3.3.4 Program Compliance

In this section we discuss teachers' compliance with the protocol. The program included 16 self-study modules with an average workload of roughly 6 hours (total workload of 96 hours) and 4 half-day meetings (total workload of 16 hours). We define the overall compliance rate as the weighted average of module completion and attended meetings, where the weights correspond to the respective total workload in hours.

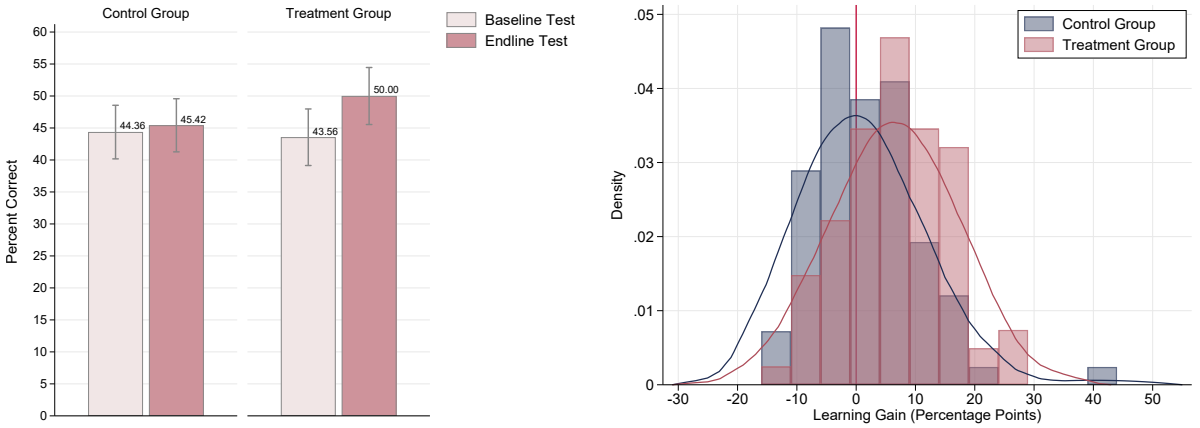
The average compliance is 75 percent and the median compliance 91 percent. Nine program teachers (10 percent) completed no module and attended no meetings, while nine of them have a compliance rate of 100 percent. Three out of four teachers completed at least 70 percent of the program. Hence, the vast majority of teachers worked on all modules and almost all teachers attended the meetings. Figure A.1 in Appendix A.1.2 presents the distribution of the module completion and meeting attendance rate. Teachers solved on average 74 percent of the modules and 85 percent attended the meetings on average. Almost 80 percent of the teachers attended all four meetings. The results discussed in the following section are ITT-estimates that do not account for the actual engagement of the teachers with the modules and monthly meetings in the training program.[11]

3.4 Results

Before we turn to the multivariate analysis, we have a look at simple group means and the distribution of learning gains. Figure 3.3a presents the mean math scores of program and control teachers before and after the intervention. In the baseline assessment teachers of both groups score about 44 percent correct answers. Teachers of the control group reach with 45 percent correct answers almost the same score in the endline assessment, while program teachers achieve 50 percent correct answers. Thus, after the program, treatment teachers outperform their peers from the control group by about 5 percentage points. Hence, the endline score of program teachers increases by 11 percent compared to the endline score of control teachers.

Figure 3.3b shows the distribution of learning gains among control and program teachers.

[11] We refrain from conducting and discussing IV estimates, because underlying assumptions of the IV estimates are likely to be violated. Further, since compliance rate is based on completion of modules and some teachers would struggle to complete more complex modules even if they try, it makes no sense to assume a hypothetical compliance rate of 100 percent. In this context, the average compliance rate of 75 percent already seems fairly high and IV estimates would not deliver meaningful additional insights.



(a) Mean test scores at base- and endline

(b) Distribution of learning gains

Figure 3.3: Mean scores between groups and distribution of learning gains
Note: Only teachers that attended the endline assessment are included in these figures (N=164). 95% confidence spikes in Figure 3.3a. Learning gain in Figure 3.3b is defined as the difference of the endline and baseline percentage score.

While the control group distribution is centered around zero, the treatment group distribution is shifted to the right and centered around a learning gain of 8 percentage points. This indicates a positive effect of the program. There is a salient outlier in the control group that advanced by 40 percentage points compared to the baseline. The effect of this outlier on our estimates will be discussed in Section 3.4.1.

3.4.1 The Overall Program Effects on Teachers

We estimate the intent-to-treat (ITT) treatment effect of being randomly assigned to the treatment group using the following equation:

$$Y_{jk}^{EL} = \alpha + \beta Treat_{jk} + \delta Y_{jk}^{BL} + X_{jk}^{'}\gamma + S_{jk}^{'}\rho + \phi_k + \epsilon_{jk}, \tag{3.2}$$

where Y_{jk}^{EL} represents the endline math score of teacher j in stratum k and is either measured in percent correct answers or standardized such that $\mu = 0$ and $\sigma = 1$. $Treat_{jk}$ is a dummy variable that indicates whether teacher j belongs to the treatment group, Y_{jk}^{BL} denotes the baseline test score. X_{jk} are pre-determined control variables on teacher characteristics and $S_{jk}^{'}$ stands for covariates on school characteristics. ϕ_k denotes stratum fixed effects and ϵ_{jk} is the error term. The vector of teacher controls includes age, gender, highest degree[12], years since graduation, math specialization and travel time to school (see Table 3.1, panel B for more details). The vector of school level covariates includes an equipment and an infrastructure index, travel time to the department capital as well as indicator variables for computer access of students, gang activities on school grounds and whether the school is located in a rural area (see Table 3.1, panel C for more details).

[12]Categories of the covariate *Highest Degree* are included as indicator variables.

Table 3.2: ITT-Estimates on the effects of the program on teachers' math scores

Dependent variable:	Percent Correct			Std. Math Score		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	5.38***	5.31***	5.45***	0.28***	0.27***	0.28***
	(1.46)	(1.51)	(1.52)	(0.08)	(0.08)	(0.08)
Baseline Score	0.90***	0.84***	0.85***	0.90***	0.84***	0.85***
	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)	(0.10)
Adjusted R^2	0.80	0.81	0.80	0.80	0.81	0.80
Observations	164	164	164	164	164	164
Teacher Controls[a]	No	Yes	Yes	No	Yes	Yes
School Controls[b]	No	No	Yes	No	No	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: *a* Teacher level controls correspond to variables in panel B in Table 2.1 (except *Absence at Endline*).
b School level controls correspond to variables in panel C in Table 2.1. Robust standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

Table 3.2 presents the main effect of the program on teacher math scores. Columns (1) to (3) show that teachers that were (randomly) assigned to the program score between 5.31 and 5.45 percentage points higher than teachers of the control group. This corresponds to an increase of 12 percent relative to the endline score of the control group.[13] Considering the generally more comparable standardized test scores as outcome measure in columns (4) to (6), we find that teachers of the treatment group outperform control teachers by 0.27σ to 0.28σ. These results are highly significant (p-value=0.00).

As mentioned above, one teacher of the control group experienced a learning gain of 40 percentage points and is a notable outlier (see Figure 3.3b). Excluding this data point, the treatment effect increases from 5.45 to 6.04 percentage points or from 0.28 to 0.31σ in columns (3) and (6). Hence, excluding the outlier increases the effect size by 11 percent (see Table A.2 in Appendix A.1.3).

In summary, benchmark results show that the five-month in-service teacher training with an average compliance rate of 75 percent has a highly significant positive effect on teacher content knowledge in math. Program teachers outperform their peers by 0.28σ or 5.45 percentage points. If we exclude the outlier in the control group, effect sizes increase to 0.31σ or 6 percentage points.

3.4.2 Effect Heterogeneity

In this section, we investigate whether the ITT effect varies by subtopic, content complexity (i.e. grade level of the questions) and teacher characteristics. We start with the analysis of the effect heterogeneity by subtopic.

[13] The increase of 12 percent is calculated as follows: The effect in percentage points is divided by the average endline percentage test score of the control group: $5.45pp./45\% = 0.12$

Program Effects by Subtopic

The base- and endline math assessment combine test items pertaining to the subtopics NSEA, GEOM and DSP.[14] In columns (1) and (3) of Table 3.3, we re-estimate Equation (3.2) considering only NSEA questions and in columns (2) and (4) only GEOM & DSP questions.

Table 3.3: ITT-Estimates on the effects on teacher's math scores by subtopic

Dependent variable:	Percent Correct		Std. Math Score	
Subject Domain:	NSEA	GEOM & DSP	NSEA	GEOM & DSP
	(1)	(2)	(3)	(4)
Treatment	4.85***	6.19***	0.27***	0.27***
	(1.81)	(2.01)	(0.10)	(0.09)
Baseline Score	0.66***	0.43***	0.66***	0.43***
	(0.11)	(0.11)	(0.11)	(0.11)
Adjusted R^2	0.72	0.72	0.72	0.72
Observations	164	164	164	164
Additional Controls	Yes	Yes	Yes	Yes
Stratum FE	Yes	Yes	Yes	Yes

Notes: Additional controls include teacher and school level controls. Robust standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

Results in Table 3.3 show that program teachers significantly outperform their peers in the control group in all domains, i.e. NSEA as well as GEOM & DSP. Using percent of correct answers as outcome variable, we find that the effect of the treatment on GEOM & DSP questions is with 6.2 percentage points (p-value=0.003) slightly larger than for NSEA questions with 4.9 percentage points (p-value=0.008). However, the effect difference between the two domains is not significant. If we are using standardized math scores, the effect difference between the two subject domains disappears and the effect in both domains is 0.27σ. Note, that effect size using standardized math scores as outcome decreases when standard deviation is large. While the average endline score in NSEA is 54.6 percent and the standard deviation is 18.1, the average score in GEOM & DSP is 29.9 percent and the standard deviation is 23.2. This explains, why effect size is the same for both domains using standardized test scores, while it varies for the percentage score. Summing up, the program is effective for NSEA as well as GEOM & DSP contents and the effect magnitude does not vary significantly across subdomains.

Program Effects by Grade Level of Questions

In a next step, we estimate the treatment effect by grade level of test questions. To this end, each question is assigned to a grade level based on the official curriculum and then we compute scores along these grade levels. Based on the grade specific scores, we re-estimate Equation (3.2)

[14] Following the official Salvadoran math curriculum, 60 percent of the questions belong to the domain NSEA, 30 percent to GEOM and 10 percent to DSP.

to assess the impact of the program at different levels of complexity. Grades two and three are combined, because the base- and endline assessments contain only six questions pertaining to grade two.

Table 3.4: ITT-Estimates on the effects on teacher's math scores by grade level

Dependent variable:	Percent Correct				Std. Math Score			
Grade level:	Gr. 2/3	Gr. 4	Gr. 5	Gr. 6	Gr. 2/3	Gr. 4	Gr. 5	Gr. 6
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	2.66	3.88*	8.88***	8.52***	0.15	0.16*	0.34***	0.36***
	(1.84)	(2.06)	(2.88)	(2.72)	(0.10)	(0.09)	(0.11)	(0.12)
Baseline Score	0.28***	0.07	0.64***	0.41***	0.28***	0.07	0.64***	0.41***
	(0.08)	(0.09)	(0.13)	(0.10)	(0.08)	(0.09)	(0.13)	(0.10)
Adjusted R^2	0.58	0.62	0.68	0.61	0.58	0.62	0.68	0.61
Observations	164	164	164	164	164	164	164	164
Additional Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Additional controls include teacher and school level controls. Robust standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

Table 3.4 shows that the impact of the teacher training was largest for concepts pertaining to grades five and six, while the effect was considerably smaller when we analyze questions of grades two to four. In particular, we find a positive but insignificant effect on grade 2 & 3 questions (p-value=0.15), while the effect is positive and significant on all other grade levels. We obtain a positive effect of 3.9 percentage points or 0.16σ (p-value=0.062) on grade level four, whereas we find a more substantial and highly significant effect of 8.9 percentage points or 0.34σ on grade level five and 8.5 percentage points or 0.36σ on grade level six. This means that the program increased teacher assessment scores for grade five and six by 25 percent determined by the respective endline score of the control group. It appears that the effect is larger for more complex contents of grade five and six and contents with which teachers struggle most according to the baseline assessment results (see Figure 3.1b).

Effect Heterogeneity by Baseline Ability, Age and Gender

We conclude the heterogeneity analysis by looking at how the effect varies with three teacher characteristics: baseline math ability, gender and age. To do so, we estimate the following regression equation:

$$Y_{jk}^{EL} = \alpha + \beta Treat_{jk} + \lambda(Treat_{jk} * Covariate_{jk}) + \delta Y_{jk}^{BL} + X_{jk}^{'}\gamma + S_{jk}^{'}\rho + \phi_k + \epsilon_{jk}, \quad (3.3)$$

where $Treat_{jk} * Covariate_{jk}$ denotes the interaction of the treatment dummy and the specific variable of interest (i.e. teacher baseline score, gender, age). The coefficient λ then captures the

extent to which the effect of the treatment differs along these interacted characteristics. All other terms are defined as in Equation (3.2).[15]

Table 3.5 presents the results of the heterogeneity analysis with standardized math scores as the dependent variable.[16] Columns (1) and (2) show that the treatment does not vary with the baseline ability of teachers. Hence, it appears that the program's effectiveness was independent of teachers' baseline ability. However, we find that the effect size significantly varies with gender and age. Columns (3) and (4) show that the effect is significantly lower for female teachers (p-values of 0.015 and 0.025). One explanation for this gender effect might be that self-study modules had to be done in free time and that female teachers have more responsibilities at home with domestic work compared to their male peers. Average module completion rate of female participants is with 69.7 percent lower compared to 80.4 percent for male participants. However, the difference is not significant (p-value=0.147 with 87 observations). Furthermore, results in columns (5) and (6) show that the program effect decreases with teachers' age (p-values of 0.050 and 0.063). The literature on aging and performance delivers three explanations, why teachers' age affects the impact: First, many of the older teachers are less familiar with computers than their younger peers and might have struggled with technical issues using the computer. Ng and Feldman (2008) find in their literature review on age and job performance that older persons perform lower in training programs and in particular that trainings focusing on technical skills (e.g. computer training) are less effective for older participants. Second, it is commonly noted that older workers are often more reluctant to engage in skill training and prefer collaborative instead of competitive tasks (Kanfer and Ackerman, 2004). Our program focusing on self-study, acquiring new skills and including a competitive component with an incentivized test at the end of each module might, therefore, be less motivating for older teachers. Third, there might be an age-related decline in learning capacity. Although age-related decline is generally considered to start gradually in the sixties or later (Schaie and Willis, 2013), fluid intellectual abilities – i.e. working memory, abstract reasoning and processing novel information – tend to decline earlier than other intellectual abilities (Kanfer and Ackerman, 2004). Fluid intellectual capacity is central in this program focusing on math and working with technology-based methods that are new to many participants.

So far, the heterogeneity analysis estimating Equation (3.3) is restricted to a linear model. We continue the analysis now by estimating the effect for several subgroups: For the bottom, middle and top terciles of baseline ability and for three age groups. On the one hand, this allows us to quantify the magnitude of the effect for subgroups. On the other hand, this approach reveals insights that do not become evident using the linear estimation models in Table 3.5. Figure 3.4 shows the results of these subgroup analysis. We find that the program effect is insignificant for the bottom tercile at the baseline (0.14σ, p-value=0.325).[17] However, there is a positive and significant effect for the middle and top tercile at the baseline. The effect is higher

[15] Note that covariates are demeaned such that the treatment effect does not vary when the interaction term is included.

[16] The results with the percentage score as dependent variable are presented in Table A.3 in Appendix A.1.4.

[17] Given the sample size, we are constrained by low statistical power here and the magnitude of the effect is with 0.14σ relatively large. We therefore cannot conclude that there is no impact on teachers of the bottom tercile.

Table 3.5: Effect heterogeneity along baseline ability, gender and age

Covariates:	Baseline Score		Female		Age	
Dep. var.: Std. Math Score	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.28***	0.28***	0.28***	0.28***	0.28***	0.28***
	(0.08)	(0.08)	(0.07)	(0.08)	(0.07)	(0.08)
Covariate	0.89***	0.84***	0.23	0.42	0.00	−0.01*
	(0.10)	(0.11)	(0.31)	(0.34)	(0.01)	(0.01)
Treatment x Covariate	0.03	0.03	−0.35**	−0.39**	−0.02*	−0.02*
	(0.06)	(0.07)	(0.15)	(0.16)	(0.01)	(0.01)
Adjusted R^2	0.80	0.80	0.81	0.81	0.81	0.81
Observations	164	164	164	164	164	164
Baseline Math Score	Yes	Yes	Yes	Yes	Yes	Yes
Additional Controls	No	Yes	No	Yes	No	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Additional controls include teacher level and school level cotrols. Robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

for the middle tercile (0.38σ) than for the top tercile (0.32σ). While point estimates suggest that the program is less effective for teachers with the lowest math ability at the baseline, the difference between the bottom tercile at the baseline and the other two terciles is not significant. However, an effect difference of 0.24σ between the bottom and middle tercile is considerable and the reason why this difference is not significant is likely to be a lack of power due to small subsample sizes. One reason why the program was less effective for the bottom tercile might be that a self-study based training requires a minimum understanding of the most basic concepts in order to be successful. However, the tutorial videos are designed to explain basic concepts from ground up. Further, CAL for students proved to be similarly effective in math for low- and high-performing students – even if there is no teacher who may support students if they do not understand the tutorial videos (see Muralidharan, Singh and Ganimian, 2019, or Chapter 2). Another explanation might be that our study plan started with modules on basic concepts but then gradually moves to more complex concepts of grades five and six. Teachers of the bottom tercile might need more time to fully understand the most basic concepts before they move on to more complex contents. Therefore, tailor-made study plans for each teacher or each group and self-paced learning – as it has proved to be crucial for students using the CAL software – might be important to make sure that all teachers benefit from the program.

Figure 3.4 further shows that the effect is most pronounced for the youngest group of teachers with 0.52σ (19 percent increase). The program also increases the math content knowledge of teachers in the middle age group by 0.28σ (p-value=0.01), while it has no effect on the oldest age group of teachers (0.03σ and p-value=0.879). Note, however, that only the effect difference between the youngest and the oldest group is statistically significant (p-value=0.02).[18]

Summing up, the program appears to be less effective for teachers with low baseline math

[18]See Figure A.2 in Appendix A.1.4 for results based on percentage score as outcome variable.

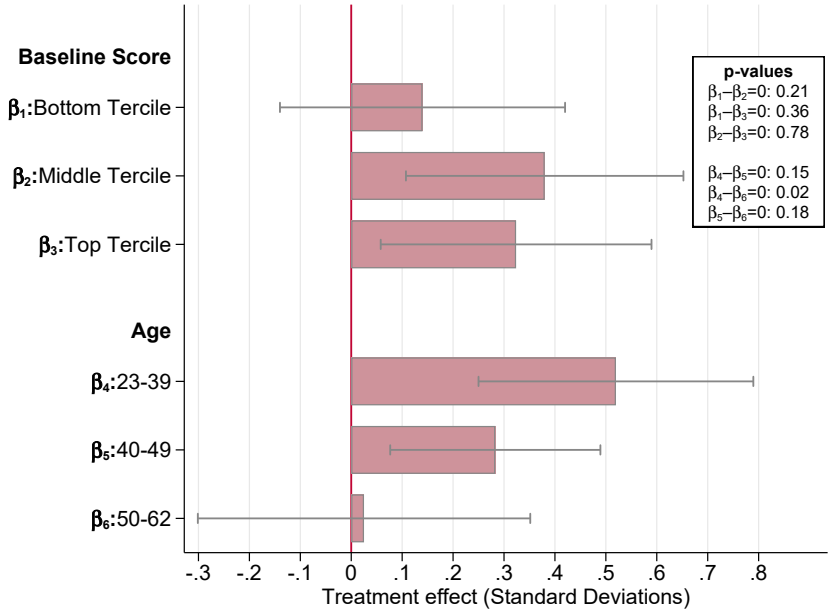


Figure 3.4: Effect heterogeneity by baseline score and and age
Note: Teacher and school level controls as well as strata fixed effects are included in all estimations. 95% confidence spikes are added.

ability and not effective for teachers belonging to the oldest age group (50 to 62 years old). The program is most effective for young teachers, male teachers or teachers who did not start from a very low ability level at the beginning of the program. Further, the program is most effective for more complex contents with which teachers struggled most in the baseline assessment (contents of grade five and six).

3.5 Discussion: Effect on Students and Cost-Effectiveness

Our results show that the program has a significant, positive effect on the math content knowledge of teachers. Ultimately, however, we are interested in the effect on students' learning progress and in the relative effectiveness of the program. To examine the relative effectiveness of a teacher intervention we need to know *(i)* the effect on teachers, *(ii)* how much of the effect teachers transmit to their students, *(iii)* how persistent the effect on teachers is and *(iv)* the costs of the program. While we discussed the program's effect on teachers' content knowledge in Section 3.4, the goal of this section is to discuss the remaining three parameters in the causal chain determining the cost-effectiveness of this intervention. Finally, we compare the relative cost-effectiveness of CAL for teachers with CAL for students from Chapter 2.

Estimated Effect on Students. Brunetti et al. (2020) examine the relation between teachers' math skills and their students' learning gains over a period of eight months in the very same context. Depending on the model specification they find that a 1σ increase in teacher knowledge is associated with a student learning gain of 0.09σ to 0.13σ. This matches well with a transmission

parameter of 0.09 reported by Metzler and Woessmann (2012) based on Peruvian data, and Bau and Das (2020) examining the impact of teacher content knowledge in Pakistan. These remarkably closely aligned estimates suggest that the effect on teachers of 0.28σ found in this study would correspond to an increase of students' learning gain of roughly 0.025σ to 0.035σ. This would be a very modest effect on students' math skills. Even if we take the effect of 0.52σ on young teachers under 40 as a reference point, we would expect a relatively small effect of 0.047σ on students. However, young teachers will instruct hundreds or even thousands of students in their career. Therefore, the overall effect might still be considerable, if the content knowledge gain of teachers is persistent.[19]

Persistence of Effect. In a next step, we discuss the persistence of the effect on teachers and the potential future stream of benefits of the five-month program. This is important for the evaluation of the cost-effectiveness, since programs addressing teachers' content knowledge or teaching abilities will have an impact on future cohorts of children, while many other interventions, e.g. CAL for students, will hardly affect future cohorts. Thus, the relative effectiveness depends largely on the persistence of the effect on teachers.

In contrast to the transmission between teacher content knowledge and student learning outcomes, the evidence base regarding the persistence of teacher training programs is still very scarce.[20] One notable exception is Cilliers et al. (2020), who evaluate the persistence of two teacher training programs focusing on pedagogy. They find that there is persistence in teachers' pedagogical knowledge one year after the program has ended. Furthermore, Cilliers et al. (2020) tested students in the same year their teacher was treated and then they tested an new cohort of students one year after the treatment and found that the effect on the new cohort was cut in half. However, these findings are based on a pedagogical program and depreciation rates of content knowledge and pedagogical knowledge might be very different. Therefore, they might not be applicable to our case.

Program Costs. The costs of the program are calculated according to the guidelines described by Dhaliwal et al. (2014) and are consistent with the cost calculations of the CAL for students intervention in Chapter 2. The bulk of costs comes from incentives paid to teachers (52 percent) and salaries paid to the NGO staff (24 percent). Furthermore, the NGO lent 87 computers to the

[19] The aforementioned studies are only to a limited extent comparable to our study. First, the previously mentioned studies analyze differences in the content knowledge of teachers from a given distribution and control for teacher characteristics with fixed effects or covariates. In this study, the content knowledge of teachers is raised to a higher level and thus, not differences within a given distribution but a shift of teachers' content knowledge is evaluated. Second, teachers' content knowledge is raised to a higher level via a training program which might affect teaching skills or behavior in other ways. For example, the training might affect teachers' motivation to teach or the learning videos might improve their pedagogical skills. Therefore, the effect of this teacher training on students' learning achievements can only be determined more exactly with a student exam (and effects on motivation or pedagogical skills of teachers with classroom observations). This student exam and monitoring of classes were planned to be conducted in 2020. However, due to the COVID-19 pandemic, schools in El Salvador were closed and neither the student exam nor the monitoring could be realized. Therefore, we rely in this discussion on the findings based on a value-added approach using panel data.

[20] A follow-up teacher content knowledge assessment with teachers participating in this study is planned at the end of 2020 resp. early 2021 to shed more light on this matter.

teachers to work on the modules. The costs of these computers are calculated according to the same standards as for the CAL for students program in Chapter 2: We assume costs of 200 USD per station and an average of five years of usage time. Following this procedure, total computer costs amount to 13 percent of the program costs. Total costs per student are 12 USD in the first year (compared to 43 USD to 56 USD for CAL for students in Chapter 2).

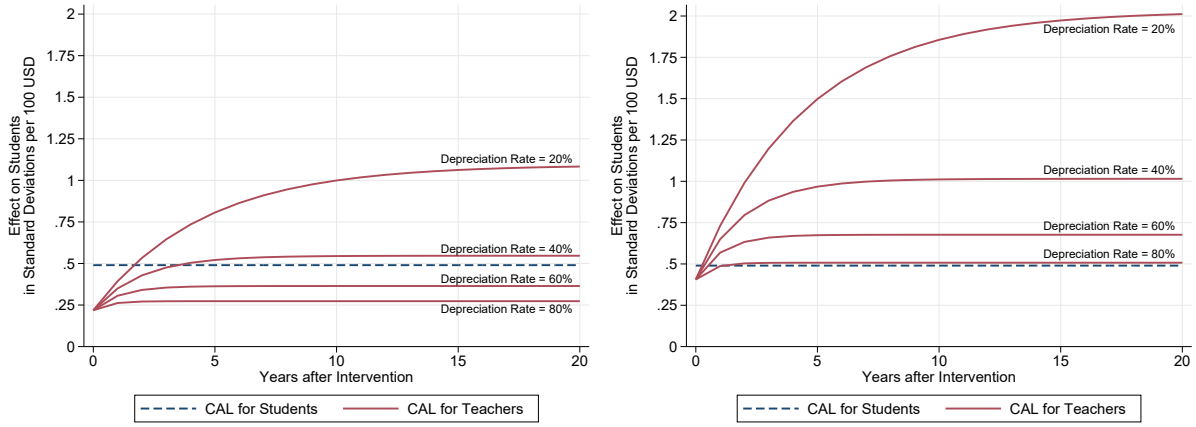
Relative Cost-Effectiveness. Based on the estimates from Section 3.4 and the above considerations on the effect's transmission on students and its persistence, we aim to compare the cost-effectiveness of CAL for teachers to the cost-effectiveness of CAL for students (see Chapter 2 of this thesis). We first discuss the cost-effectiveness during the first year of the intervention and next examine the cost-effectiveness considering also benefits of future cohorts of students.

To calculate the cost-effectiveness of CAL for teachers, we assume a transmission of the treatment effect on teachers to their students of 0.09. Consequently, an effect of 0.28σ on teachers corresponds to an effect of 0.025σ on students. Further, we know that the 87 program teachers instruct 2590 students in math in the first year. Thus, with total costs of roughly 30,000 USD the program is expected to yield a 0.22σ increase in students' math score per 100 USD spent. Investing the same amount of money in CAL lessons for students yields an increase in math scores of 0.43σ to 0.49σ (see Chapter 2). Hence, in the first year the cost-effectiveness of CAL for teachers appears to be lower than the one of the CAL for students program. When we consider the effect on young teachers below 40 years of age, we find an effect on teachers of 0.52σ. Taking this effect as an upper bound of the CAL for teachers intervention, investing 100 USD yields an increase of 0.41σ in students' math score, which is comparable to the cost-effectiveness of the CAL for students program.

To consider future cohorts that might benefit from the program, we need to make two additional assumptions: the number of years that treatment teachers are expected to teach after the intervention and the persistence of the effect on teachers. As treatment teachers are on average 44 years old, we assume that they teach 21 more years (although many teachers in El Salvador work beyond the retirement age of 65). It gets more complicated when we attempt to make a reasonable assumption about the depreciation rate of the effect on teachers. Since we cannot draw on directly comparable estimates on the persistence of the effect, we assume four different scenarios with the annual depreciation rate varying between 20 percent and 80 percent; the yearly depreciation rate of the CAL for students program is assumed to be 100 percent, since this program does not affect future cohorts absent additional investments.

Figure 3.5 shows the cost-effectiveness of CAL for teachers and CAL for students over 21 years. It plots the cost-effectiveness of the benchmark results with an effect of 0.28σ on teachers (Figure 3.5a) as well as an upper bound effect of 0.52σ on young teachers (Figure 3.5b). Because of the assumed depreciation rate of 100 percent, the cost-effectiveness of CAL for students remains the same for all years.[21] Considering the cost-effectiveness of CAL for teachers, we find that the rate of depreciation is decisive. In the benchmark scenario a yearly depreciation rate of 20 percent

[21] Figure 3.5 shows the upper bound cost-effectiveness of 0.49σ per 100 USD for CAL for students.



(a) Benchmark with all teachers

(b) Upper bound with young teachers only

Figure 3.5: Cost-Effectiveness of CAL for teachers vs. CAL for students
Note: Figure 3.5a is based on benchmark results assuming an effect of 0.28σ on teachers. Figure 3.5a is based on results considering only young teachers below 40 years of age assuming an effect of 0.52σ on teachers. In both figures we assume a reach of 2590 students in each year and a transmission rate of 0.09 to students for the CAL for teachers program.

would amount to an increase of students' math score of 1.08σ per 100 USD 20 years after the completion of the training. Under this scenario, CAL for teachers would already be more cost-effective than CAL for students two years after the intervention. In a scenario of a depreciation rate of 40 percent, CAL for teachers would also be more cost-effective than CAL for students four and more years after the intervention. However, in scenarios of higher depreciation rates CAL for teachers would remain less cost-effective than the CAL version for students. In the upper bound scenario, assuming that only young teachers below 40 years of age are accepted into the program, CAL for teachers is more cost-effective than CAL for students from one year after the intervention onward if the depreciation rate is 80 percent or lower. With a depreciation rate of 80 percent, CAL for teachers would be as cost-effective as CAL for students one year after the intervention (0.49σ per 100 USD on students) and then remain on a similar level for later years.

These calculations show how important it is to take benefits of future cohorts into account when comparing the cost-effectiveness of different programs. While benchmark estimations show that CAL for teachers is considerably less cost-effective than CAL for students in the first year after the intervention, CAL for teachers may become substantially more cost-effective when considering the benefits accruing to future cohorts of students. However, these calculations should be interpreted with caution due to several limitations. First, we need to make assumptions for the transmission of the effect to students and the persistence of the effect. Second, the effect of CAL for teachers on a single student is very small (0.025σ to 0.035σ) and the CAL for teachers intervention becomes more cost-effective because the program reaches many students at a low cost. It is not clear whether it is more desirable to get such a small (and probably not measurable) effect on a vast number of students or whether it is more desirable to get a larger effect on a smaller number of students.

3.6 Conclusion

Teachers are a central component of an effective public schooling system. However, there is growing evidence that teachers are insufficiently prepared for their teaching responsibilities. In particular, it has recently been found in several countries that teachers do not understand the material they are supposed to teach their students in primary school (Bold et al., 2017a; Sinha, Banerji and Wadhwa, 2016; Brunetti et al., 2020). Accordingly, it is essential to identify ways of efficiently improving the content knowledge of teachers. This paper investigates the impact of a novel teacher training approach: Instead of providing CAL lessons to students, CAL lessons are offered to teachers.

We find that this training based on computer-assisted self-studying significantly improved teachers' math skills. After five months of training teachers score 0.28σ or 5.45 percentage points higher than their peers from the control group. The training is most effective in the subtopic and grade levels where the teachers scored particularly poorly in the baseline assessment (i.e. contents of grade five and six). Furthermore, the program is most effective for young teachers under the age of 40 (0.52σ). At the same time, it does not appear to be effective to improve the math skills of teachers above the age of 50 years. Hence, it would make sense to focus on young teachers for two reasons: First, the effect is larger for younger teachers and second, young teachers will be instructing for more years and will therefore teach more students.

Our results show how important it is to take benefits of future cohorts of students as well into account when discussing the cost-effectiveness of programs. While the benchmark results suggests that CAL for teachers is less cost-effective than CAL for students in the first year, CAL for teachers may become substantially more cost-effective than the version for students when considering as well the costs and benefits of later years. Since teacher trainings are costly and future benefits are often not considered, the cost-effectiveness of such programs is often underestimated. Our estimations depend on the assumed persistence of the effect. Therefore, gaining a better understanding about the impact persistence of teacher trainings and how effects of interventions can better be sustained seems to be important and is therefore an interesting question for future research.

A. Appendices

A.1 Appendix: Additional Analysis

A.1.1 Attrition

A total of 11 teachers (6.3 percent) did not take part in the endline assessment. In Table A.1 we examine whether this attrition is correlated with the treatment assignment. Results in columns (1) to (3) are based OLS estimations and the result in column (4) is based on a Logit model.[1] We find that attrition is not correlated with the treatment status.

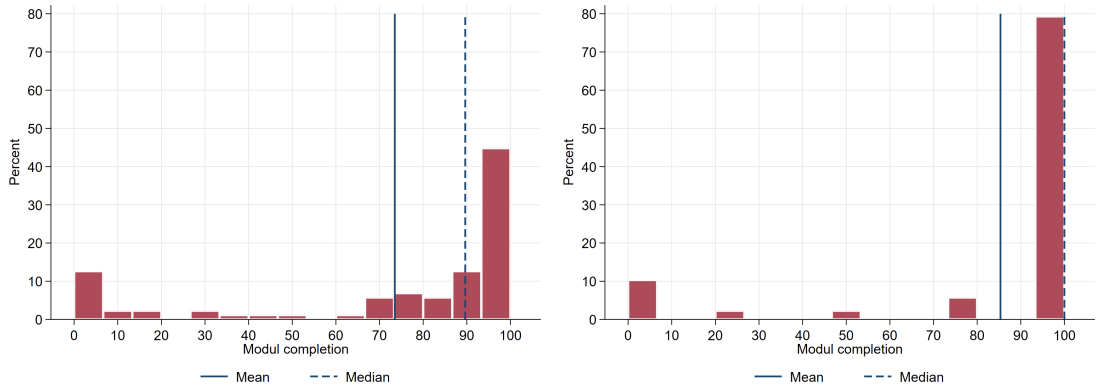
Table A.1: Attrition

	OLS			Logit
Dep. Var.: Attrition at Endline	(1)	(2)	(3)	(4)
Treatment	0.012	0.011	0.023	0.207
	(0.037)	(0.039)	(0.040)	(0.627)
Baseline Score		−0.003*	−0.004*	
		(0.002)	(0.002)	
Adjusted R^2	0.00	0.00	0.00	–
Pseudo R^2	–	–	–	0.00
Observations	175	175	175	175
Teacher Controls	No	No	Yes	No
School Controls	No	No	Yes	No
Stratum FE	No	Yes	Yes	No

Notes: Robust standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

[1] The reason why no controls nor stratum fixed effects are included in the Logit model is that with only 11 teachers absent at the endline, the outcome variable does not vary in some strata or covariates. Consequently, a large number of observations would be dropped in the Logit model when controlling for other covariates.

A.1.2 Program Compliance



(a) Completion rate of modules (b) Attendance rate of meetings

Figure A.1: Compliance of teachers with the program
Note: The overall compliance rate is the weighted average of the module completion rate in Figure A.1a and the attendance rate in Figure A.1b with an average of 75 percent and a median of 91 percent.

A.1.3 Robustness without Outlier

Table A.2: ITT-Estimates on the effects of the program on teachers' math scores: Outlier in control group excluded

Dependent variable:	Percent Correct			Std. Math Score		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	5.90***	5.89***	6.04***	0.30***	0.30***	0.31***
	(1.37)	(1.40)	(1.41)	(0.07)	(0.07)	(0.07)
Baseline Score	0.90***	0.84***	0.84***	0.90***	0.84***	0.84***
	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)	(0.10)
Adjusted R^2	0.82	0.83	0.83	0.82	0.83	0.83
Observations	163	163	163	163	163	163
Teacher Controls[a]	No	Yes	Yes	No	Yes	Yes
School Controls[b]	No	No	Yes	No	No	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: *a* Teacher level controls correspond to variables in panel B in Table 2.1 (except *Absence at Endline*).
b School level controls correspond to variables in panel C in Table 2.1. Robust standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

A.1.4 Heterogeneity

Table A.3: Effect heterogeneity along baseline ability, gender and age

Covariates:	Baseline Score		Female		Age	
Dep. var.: Percent Correct	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	5.35***	5.41***	5.35***	5.39***	5.40***	5.37***
	(1.47)	(1.54)	(1.44)	(1.49)	(1.44)	(1.51)
Covariate	0.89***	0.84***	4.42	8.08	−0.02	−0.28*
	(0.10)	(0.11)	(5.97)	(6.68)	(0.13)	(0.16)
Treatment x Covariate	0.03	0.03	−6.72**	−7.50**	−0.35*	−0.33*
	(0.06)	(0.07)	(2.98)	(3.04)	(0.18)	(0.17)
Adjusted R^2	0.80	0.80	0.81	0.81	0.81	0.81
Observations	164	164	164	164	164	164
Baseline Math Score	Yes	Yes	Yes	Yes	Yes	Yes
Additional Controls	No	Yes	No	Yes	No	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Additional controls include teacher and school level controls. Robust standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

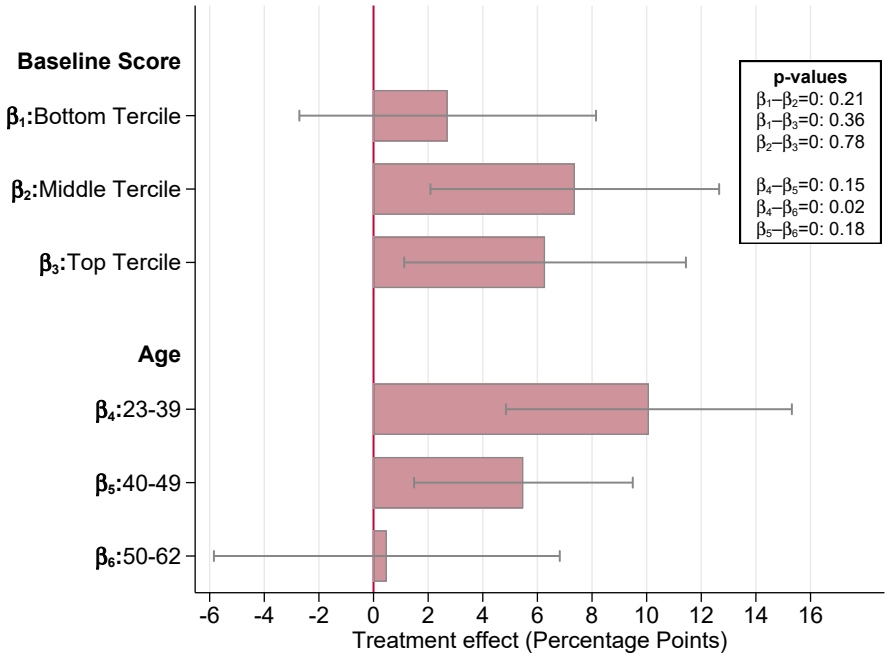
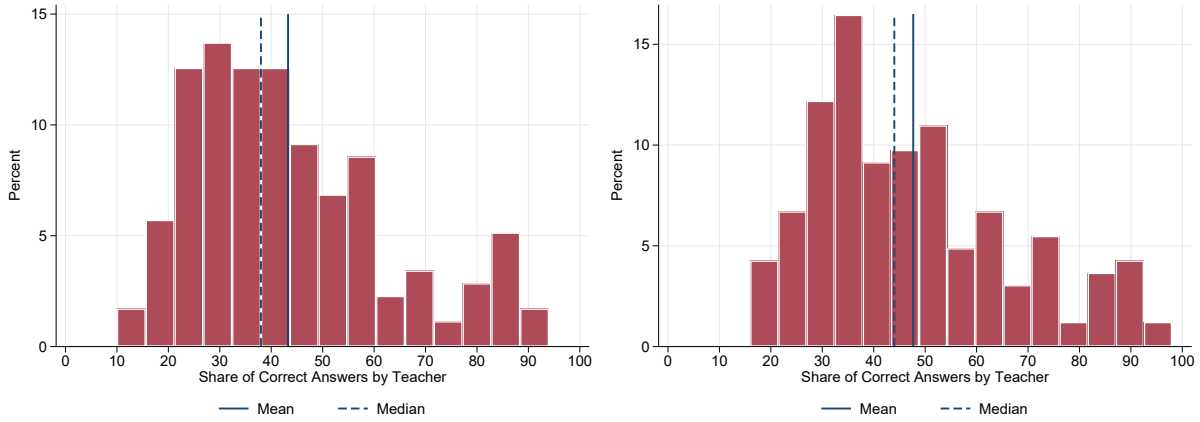


Figure A.2: Effect heterogeneity by baseline score and age

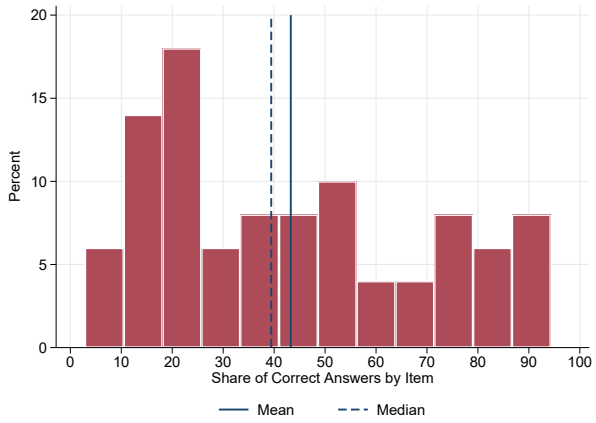
A.2 Appendix: Assessment Diagnostics

Figures A.3a and A.3b present the distribution of correct answers by teacher for the base- and endline assessment. Similarly, Figures A.3c and A.3d display the percentage of correct answers by item for both test waves. These figures show that there are no floor nor ceiling effects. Teachers were able to answer at least 10 percent of the items in the baseline and 16 percent in the endline. On the other hand, no teacher scored 100 percent correct answers in any of the waves. Further, there is no item that was not answered correctly by anyone (minimum of correct answers is 3 percent in baseline and 15 percent in endline) or an item which was solved successfully by all teachers (maximum of correct answers is 94 percent in baseline and 98 percent in endline).

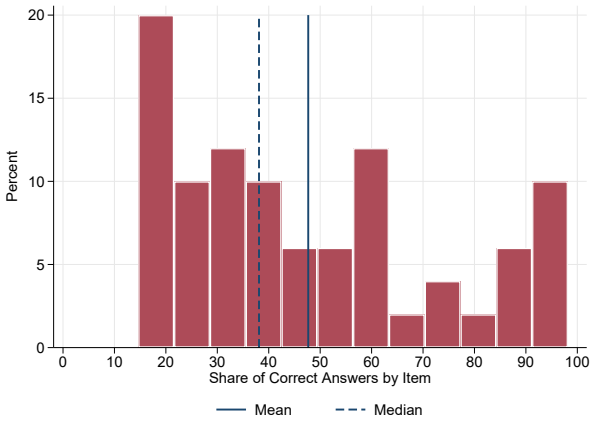


(a) Baseline assessment by teacher

(b) Endline assessment by teacher



(c) Baseline assessment by item



(d) Endline assessment by item

Figure A.3: Share of correct answers across teachers and items

A.3 Appendix: Review of Teacher-Oriented Interventions

In total, we found 31 studies that estimate the impact of teacher-oriented interventions on teacher, student or classroom outcomes in low- and middle-income countries with experimental or quasi-experimental methods. With teacher-oriented interventions we mean programs that at least have a small teacher training component. This overview further only includes studies addressing preschools/Kindergarten, primary schools or secondary schools.

We started our literature search by examining different review articles on professional teacher development or related topics. Most important sources are the review articles by Snilstveit et al. (2015) and Popova et al. (2018) but as well the handbook chapters by Muralidharan (2017) and Glewwe and Muralidharan (2016). Other sources used for this literature review are Ganimian and Murnane (2016), McEwan (2015), Conn (2017), Kraft, Blazar and Hogan (2018) as well as the report on education by the World Bank (2018). Furthermore, we searched the AEA RCT registry with the keyword "*Teacher Training*". Based on the articles found in these first two steps, we conducted a thorough internet search. In this internet search we mainly added more recent articles up to early 2020.

Of the 31 studies, 22 studies do not have a content knowledge training component. These 22 studies are briefly summarized in Table A.4 and labeled as *pedagogical* studies. Note, however, that the focus of these studies is not necessarily on pedagogical training. Some studies are mainly providing teaching materials and include only a small component of pedagogical training. However, these studies have in common that the trainings do not include a content knowledge component. The remaining nine studies that target at least gradually teachers' content knowledge are summarized in Table A.5. These studies evaluate mostly multifaceted interventions that do not focus on content knowledge. Therefore, we label these studies as *mixed* teacher training studies. Two of these nine studies have a main focus on teachers' content knowledge, but are still mixed approaches since they involve as well pedagogical aspects. One study is a follow-up study examining the persistence of the impact of a program (Cilliers et al., 2019, 2020).

In the two overview tables we document the evaluation method, the region in which the intervention took place and the sample size of the study, the grade level, the main intervention and the type of teacher training, as well as the school subject targeted, the intensity and duration and the main insights of the study. Because most studies evaluate multifaceted interventions, we try to summarize the most important components with a clear focus on the teacher training. The type of teacher training is either labelled as "pedagogical" or "pedagogical and content knowledge". Since studies examine a broad variety of different outcomes and use different measurement units, we briefly summarize main findings, but refrain from always indicating the exact effect size (e.g. in standard deviations).

Table A.4: Literature Review of Pedagogical Teacher Trainings (I)

Study	Method	Sample & Region	Grade Level	Main Intervention	Teacher Training	Subject	Intensity & Duration	Main Insight
Aber et al. (2017)	RCT	DR Congo 64 schools 4465 students	2 & 4	1) Integrated teacher resource material, 2) collaborative school-based teacher learning circles	Pedagogical	not specified	Weekly meeting, 1 year	Marginally significant positive impacts on children's reading and geometry scores but not on their arithmetic scores. The intervention had the largest impacts on math scores for language minority children and in low-performing schools
Albornoz et al. (2018)	RCT	Argentina 70 schools 3000 students	7	Providing a structured curriculum unit, teacher training and weekly coaching	Pedagogical	Science	12 weeks	Substantial learning gains of 0.55–0.64σ for students whose teachers were trained using structured curriculum units, as well as for those whose teachers received coaching
Banerjee et al. (2016)	RCT	India 4 different evaluations	Primary school	Multiple interventions which implemented a core pedagogical approach, Teaching at the Right Level (with differences in delivery method, duration, and location)	Pedagogical	Language & math	Duration and intensity differ among evaluations	Gains in language and math between 0.15σ and 0.70σ for students
Bassi, Meghir and Reynoso (2016)	RCT	Chile 843 schools	4	Targeted educational policy providing technical and pedagogical support (scaffolding approach)	Pedagogical	Language, math & science	6-week cycle	Overall, the program improves Reading test scores of students by about 0.1σ in the first year and by 0.09–0.13σ in the second year in all subjects
Berlinski and Busso (2013)	RCT	85 schools 4800 students	Secondary school	4 Treatments: 1) new curriculum design (NCD); 2) NCD & interactive whiteboard; 3) NCD & computer lab; 4) NCD & a laptop for every child	Pedagogical	Math	4 weeks of 10 hours teacher training	Students in the control group learned significantly more than those who receive treatment

Cont. of Table A.4: Literature Review of Pedagogical Teacher Trainings (II)

Study	Method	Sample & Region	Grade Level	Main Intervention	Teacher Training	Subject	Intensity & Duration	Main Insight
Blimpo et al. (2019)	RCT	Gambia 53 centers	Preschool	Two treatments: 1) Improving access to early childhood development centers (ECD) and 2) teacher training	Pedagogical	not specified	3 sessions of 5–8 days	No evidence that either intervention improved average levels of child development. Evidence from first treatment that it led to declines in child development among children from less disadvantaged households
Fleisch et al. (2016)	RDD	South Africa 870 schools	1–3	Daily lesson plans, high-quality learning and teaching materials, and instructional coaching linked to "just-in-time" training	Pedagogical	Numeracy (literacy part of intervention but not evaluated)	Coaching once per month over 3 years	Increase of average numeracy scores of students between 0.35σ and 0.77σ depending on grade (LATE)
He, Linden and MacLeod (2009)	RCT	India 3 different RCTs	Primary & preschool	Teachers use story books, flash cards for word and letter recognition, and barakhadi charts to instruct children	Pedagogical	Literacy	$\sim$6 weeks	The program delivers gains of 0.12–0.70σ on student level
Jukes et al. (2016)	RCT	Kenya 101 schools 2539 students	1 & 2	Lesson plans, initial workshop to create instructional materials, problem-solving workshop, refresher training, support for teachers	Pedagogical	Literacy	Two workshops of a few days and support via text messages during two years	There was more instruction with written text and more focus on letters and sounds. There was a positive impact on three of four primary measures of children's literacy after two years, and school dropout reduced
Kerwin and Thornton (2020)	RCT	Uganda 38 schools 1481 students	1	Mother-tongue-first early-primary literacy program providing material inputs, high-quality teacher training, and support	Pedagogical	Literacy	3 intensive, residential teacher-training sessions and 6 in-service training workshops on Saturdays	Full program raised reading scores by 0.64σ and writing scores by 0.45σ. Reduced-cost version instead yields statistically-insignificant reading gains and some large negative effects (-0.33σ) on advanced writing

Cont. of Table A.4: Literature Review of Pedagogical Teacher Trainings (III)

Study	Method	Sample & Region	Grade Level	Main Intervention	Teacher Training	Subject	Intensity & Duration	Main Insight
Leme et al. (2012)	Diff-in-Diff	Brazil 393 municipalities	4 & 8	Introducing structured teaching methods and providing teaching materials	Pedagogical	Language & Math	Instructions every two or six months	Treatment group students perform significantly better in Portuguese and mathematics. No difference in passing rates
Lucas et al. (2014)	RCT	Kenya/Uganda 112 schools 7000 students	Primary school	Training of teachers, principals and school management committees (scaffolding approach); supply of materials and mentoring of teachers	Pedagogical	Literacy	$\sim$1 year	Written literacy of students increased by 0.2σ in Uganda, while no significant effect was found in Kenya. Oral literacy increased by 0.18σ in Uganda and by 0.08σ Kenya
Nonoyama-Tarum and Bredenberg (2009)	RCT	Cambodia 20 schools 25,000 students	1	"Bridging" curriculum, teacher training, regular monitoring and physical upgrading of classrooms	Pedagogical	not specified	14 days	Treatment students performed significantly better in school readiness skills and achievement of formal curriculum and maintained their learning advantage after a year
Özler et al. (2016)	RCT	Malawi 199 child care centers	Preschool	4 treatments (cumulative): 1) Provision of learning materials (control group), 2) teacher training and mentoring, 3) monthly stipend for teachers, 4) parenting education program	Pedagogical	Language	5 weeks of residential training	Children in the integrated intervention arm (teacher training and parenting) had significantly higher scores in measures of language and socio-emotional development than children in centers receiving teacher training alone at the 18-month follow-up. No effects on child assessments at the 36-month follow-up
Pallante and Kim (2013)	RCT	Chile 22 classrooms 617 students	Kinder-garten & 1	Teacher workshops and literacy coaching	Pedagogical	Literacy	4 workshops, 1 year	Positive effects on growth rates in letter naming, word reading, vocabulary and phonemic segmentation fluency. Effect sizes ranged from small *Cohen's* d=0.18 to fairly large *Cohen's* d=0.7

Cont. of Table A.4: Literatrue Review of Pedagogical Teacher Trainings (IV)

Study	Method	Sample & Region	Grade Level	Main Intervention	Teacher Training	Subject	Intensity & Duration	Main Insight
Piper and Korda (2011)	RCT	Liberia 400 schools 4000 students	1 & 2	Teaching methods, new material, and professional development and coaching	Pedagogical	Language, numeracy	22 months	Significant positive average effect 0.42σ for English, 0.35σ for local language and 0.19σ for numeracy
Piper et al. (2018)	RCT	Kenya 847 schools 3309 students	1–2	3 treatments (cumulative): 1) Teacher professional development (PD), instructional support and coaching; 2) add. revised student books; 3) add. structured lesson plans	Pedagogical	Literacy and numeracy	10 days PD per year, 15 days of PD for curriculum support and coaching	Treatment 2) and 3) had statistically significant positive learning impacts, treatment 1) did not. Treatment 3) was most cost-effective
Sailors et al. (2014)	RCT	Malawi 42 schools	1–3	Complementary teaching and learning materials, workshops, and directive coaching	Pedagogical	Literacy	Three 3-day workshops, intervention took 5 months in total	Treatment teachers were significantly more comfortable with languages of instruction and their teaching ability, beliefs about learning materials and about the culture of reading in communities
Spratt, King and Bulat (2013)	RCT	Mali 100 schools 9000 students	1 & 2	Teacher training that introduces the Read-Learn-Lead program as well as provision of supportive material	Pedagogical	Language/ Literacy	2 years	Significant positive effect on children's learning skills. However, the authors conclude that those skills are still exceedingly low even after two years of engagement with the program. The average third grader in a treatment school could read only 11 words of connected text in one minute at endline
Tan, Lane and Lassibille (1999)	RCT	Philippines 30 schools 3000 students	1–5	Training course to use supplied pedagogical material, which aims at pacing teaching according to the abilities of students	Pedagogical	Language & math	1 week training course for teachers	The combination of multi-level learning materials and parent-teacher partnerships appears to be the most cost-effective

Cont. of Table A.4: Literatrue Review of Pedagogical Teacher Trainings (V)

Study	**Method**	**Sample & Region**	**Grade Level**	**Main Intervention**	**Teacher Training**	**Subject**	**Intensity & Duration**	**Main Insight**
Wolf (2019), Wolf et al. (2019)	RCT	Ghana 240 schools 3345 students	Preschool	Two treatments: 1) teacher training and coaching; 2) additional parental-awareness meetings	Pedagogical	Literacy & numeracy	8 days of training workshops, 6 coaching visits	After year 3: Positive impacts of treatment 1) on literacy, and negative impacts of treatment 2) on numeracy. Moderate impacts found on professional well-being and classroom quality
Yoshikawa et al. (2015)	RCT	Peru 64 schools 1876 students	(pre-) Kinder-garten	Workshops & in-class coaching (and coordination health services)	Pedagogical	Language/ literacy	4 weekly activities, 2 years	Medium to large impacts on classroom organization; no significant impacts on child outcomes, i.e. language/ literacy

Table A.5: Literature Review of Mixed Teacher Trainings (I)

Study	Method	Sample & Region	Grade Level	Main Intervention	Teacher Training	Subject	Intensity & Duration	Main Insight
Bando and Li (2014)	RCT	Mexico 144 teachers	7–12	Intensive English instruction and pedagogical training	Content (80%) and pedagogical knowledge (20%)	English	100 hours in 10 days	Trained teachers improved their English by 0.35σ in the short run and employed more creative pedagogical techniques. Additionally, they spent more class time speaking English. As a result, students improved their English by around 0.16σ
Beuermann et al. (2013)	RCT	Peru 106 schools 2771 students	3	Introduce student-centered methodologies, interactive workshops to develop content knowledge and teacher tutoring	Pedagogical and content knowledge	Science	20–60 hours, 5 month	Positive and significant improvements of 0.18σ in students' test scores
Cilliers et al. (2019)	RCT	South Africa 230 schools 3000 students	1–2	2 treatments: 1) Short, intensive training at central venue and 2) monthly coaching visits	Pedagogical and content knowledge (treatment 2)	Reading	2-days teacher training twice in a year and coaching during the year (37 hours)	After two years of exposure to the program, students' reading proficiency increased by 0.12 and 0.24σ if their teachers received training or coaching, respectively
Cilliers et al. (2020)	RCT	South Africa 180 schools	1–3	Two treatment arms: 1) Centralized training, 2) On-site coaching, lesson plans and learning material for both arms	Pedagogical	Literacy	1) Two-day training twice a year, 2) Monthly coaching visits	Positive effect on students in both treatment arms; Effect is cut in half for new cohort of students
Johnson et al. (2019)	RCT	Rwanda 236 teachers	1–4	Teacher professional development aiming to improve teachers' instructional practices and literacy content knowledge	Pedagogical and content knowledge (pedagogical content knowledge)	Literacy	9 sessions at regular intervals during school year lasting 4–5 hours	Teachers in sectors assigned to receive professional development had significantly higher levels of early literacy pedagogical content knowledge than teachers in control sectors (0.56σ), and they reported using significantly more research-based literacy pedagogical practices

Cont. of Table A.5: Literatrue Review of Mixed Teacher Trainings (II)

Study	**Method**	**Sample & Region**	**Grade Level**	**Main Intervention**	**Teacher Training**	**Subject**	**Intensity & Duration**	**Main Insight**
Mouton (1995)	RCT	South Africa 48 teachers 2009 students	5	Introduction of communicative teaching method and content training (English Operacy Program)	Pedagogical & content knowledge	English	3 weeks of teacher training	Significant improvements in performance of the students, especially in math and social studies. Program teachers showed more confidence and were more relaxed and enthusiastic compared to their counterparts in the control group
Pournara et al. (2015)	Diff-in-Diff	South Africa 5 schools 800 students	Secondary schools	Teacher training with a focus on mathematics-for-teaching	Content Knowledge (75%) & Pedagogical (25%)	Math	8 two-day sessions, 1 year	Effect of *Cohen's d* of 0.17 on student level, which is equivalent to 2 months' additional progress
San Antonio, Morales and Moral (2011)	RCT	Philippines 50 teachers	6	Module-based professional development: 1)–3) Modules focusing on pedagogy, 4) & 5) mathematics	Pedagogical & content knowledge	Math	5 weekly modules	No significant effect on teachers' commitment levels or pupils' mathematics proficiency levels
Zhang et al. (2013)	RCT	China 87 teachers 8387 students	3 & 4	In-service intensive English training program for teachers to improve knowledge & pedagogy	Pedagogical & content knowledge	English	3-weeks daily training	No significant effect on teacher or student English knowledge

4 The Effect of a Second Home Construction Ban on Real Estate Prices: Quasi-Experimental Evidence Using the Synthetic Control Method*

4.1 Introduction

Housing markets are typically heavily regulated. However, the causal effects of these regulations are usually hard to estimate. Therefore, it often remains unknown whether the effect caused by a regulation is the effect regulators were seeking. Currently, there are worldwide efforts for such regulations that attempt to restrict the construction or purchase of second homes by non-local buyers. The stock of second homes has risen sharply in a number of countries such as the USA, United Kingdom, France, Switzerland or China.[1] This strong raise of non-local demand for residential properties leads to growing resistance against second home investors. One concern is that this demand for second homes leads to price increases, which the local population can no longer afford. Another concern is that second homes aggravate urban sprawl. Thus, politicians are seeking regulations that limit the second home construction without entailing any unintended effects on local real estate markets.

An example of such a regulation is the the so-called Second Home Initiative (SHI) in Switzerland. On March 11, 2012, Swiss citizens accepted this very drastic regulation in a popular vote. The SHI bans the construction of second homes[2] in all municipalities with a share of 20 percent of second homes or more. Accordingly, homes built after the vote in 2012 cannot be used as second homes at any point in the future (Federal Act on Second Homes of 2015). Approximately one out of five municipalities in Switzerland has a second home share above the limit of 20 percent, and 17 percent of all homes are second homes. Thus, second homes are popular in Switzerland and play an important role in the real estate market. In some regions, second homes are even the main driver of the real estate market and have a notable impact on local economies in general. Particularly in the Alps, almost all municipalities are affected by the SHI (see Figure 4.1) and second home shares of 50 percent and more are common in touristic municipalities.

[1] See Hilber and Schöni (2020) for an overview.

[2] Second homes are broadly defined as homes not permanently used by persons who are either registered as permanent residents in this specific municipality or living in this municipality for work or educational reasons during the working week (Ordinance of Second Homes of 2012). Hence, second homes are mostly used as holiday or investment homes.

*This chapter partially builds on my MA-Thesis.

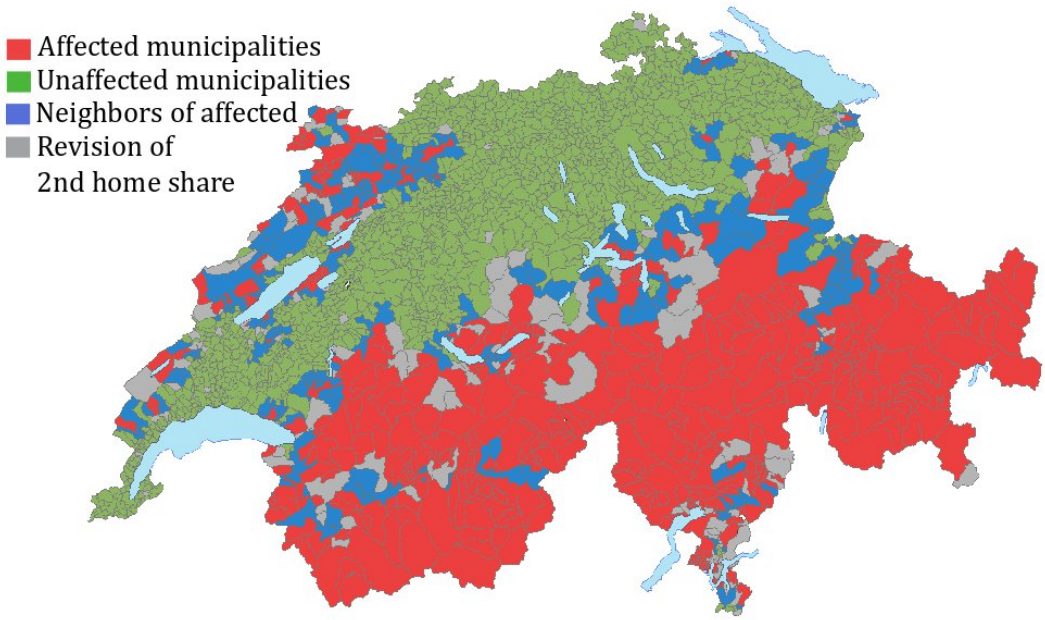


Figure 4.1: Municipalities of Switzerland marked as affected or unaffected municipalities, neighbors of affected municipalities and municipalities that asked for a revision of their official second home share.

For instance, the SHI affects more than 70 percent of all municipalities in the biggest mountain cantons, Graubünden and Valais. Therefore, the SHI was expected to cause major distortions in regional real estate markets and – as will be elaborated later – perhaps even to present some drawbacks for local economies. The goal of this paper is to estimate the (unintended) causal effects of this intervention on real estate prices.

As mentioned before, it is often very difficult to isolate the effect of such regulations. Since not all municipalities are affected by the SHI, the SHI offers a unique quasi-experimental research design to estimate the causal effect of the intervention. In 2012, 458 municipalities out of 2352 held a share of more than 20 percent of second homes. This enables a separation of municipalities into an unaffected control (less than 20 percent share of second homes) and an affected treatment group (all others). The possibility of separating municipalities into treatment and control groups might upon first consideration suggest applying a difference-in-differences estimation (DD). However, DD is based on the assumption that the outcome variables of the control and treatment groups have parallel trends in the pre-intervention period (Angrist and Pischke, 2008), which is often not the situation.

In this case, almost all affected municipalities are rather small and remote towns located in the Alps, while most unaffected municipalities are located in the densely populated Swiss Mittelland dominated by major urban centers of the country (see Figure 4.1). Thus, DD would roughly compare the real estate market of the Alps with a real estate market dominated by

major urban centers of the Swiss Mittelland. The average market structures of these two regions are not comparable. Therefore, the key assumption of parallel pre-intervention trends does not appear to be credible in the case of the SHI. In a first step, I demonstrate with DD placebo tests, the inclusion of group-specific time trends and causality tests in the spirit of Granger that the parallel trend assumption is unlikely to hold in this context (see Section 4.6.1). Hence, a *first challenge* of the paper is to find an identification strategy that handles this problem.

The synthetic control method (SCM), pioneered by Abadie and Gardeazabal (2003) and Abadie, Diamond and Hainmueller (2010) relaxes the parallel trend assumption. Although treatment and control municipalities are different on average, there exist control municipalities, which are similar to the treated municipalities in the Alps. I.e. almost all affected municipalities are located in the Alps, but not all municipalities in the Alps (or municipalities comparable to municipalities in the Alps) are treated. Hence, there are municipalities in the donor pool that are fairly similar to affected municipalities. The SCM assigns higher weights to control units that are similar and a weight of zero to control units that are very different from the treatment unit instead of comparing plain averages. For that reason, the SCM is applied in this paper. While the classic SCM literature usually deals with one or only a few treatment units, the SHI affects numerous local and heterogeneous housing markets (i.e., municipalities). A *second challenge* is, accordingly, to adapt the classic SCM to a setup with numerous treatment units.[3]

I compute a synthetic control for each treated unit, re-weight the gaps between treatment units and synthetic controls and aggregate them to compute the overall effect of the SHI, comparable to Acemoglu et al. (2016) and Kreif et al. (2015). Above all, this extension allows more precise inferences to be drawn by applying almost arbitrarily many placebo permutations, while the number of permutations and therefore, the power of statistical significance is limited in the classic synthetic control approach (e.g., Abadie, Diamond and Hainmueller, 2010). Insufficient power to detect statistical significance is commonly considered to be a weakness of classic SCM. The SCM with multiple treatments applied in this paper is able to overcome the issue of insufficient statistical power. I take this opportunity of multiple treatment units to introduce an innovative way to compute precise statistical significance levels.

This paper connects to an emerging strand of literature studying the relationship between non-local demand and local real estate markets. Favilukis and Van Nieuwerburgh (2017) and Chao and Yu (2014) develop theoretical models to demonstrate how non-local demand decreases the affordability of residential properties for local buyers. Favilukis and Van Nieuwerburgh (2017) calibrate the model for typical US metropolitan areas and find that rents increase by 19 percent and housing prices by 10 percent if non-local investors buy 10 percent of a city's housing supply. Badarinza and Ramadorai (2018) are using political shocks in a source country as exogenous instrument to address the question whether foreign capital is responsible for real estate price movements. They find that non-local demand strongly affects London's housing prices. Similarly, Cvijanovic and Spaenjers (2015), Chinco and Mayer (2015) and Sá (2016)

[3]Chen, Jain and Yang (2019) conduct a literature review for studies using the SCM with multiple treatment units. Including this study they have found 7 very recent studies using SCM with multiple treatment units. While I am dealing with 89 treatment units in the final data, all studies but one deal with less than 30 treatment units.

all find evidence that non-local demand is linked with higher local housing prices in different contexts. Thus, this strand of literature demonstrates the existence of a relationship between non-local demand and rising housing prices as well as a corresponding decrease in affordability of residential properties for local buyers. These results show that the concerns that real estate is no longer affordable for local buyers due to non-local demand may well be justified. Furthermore, the emergence and growth of this strand of literature in recent years shows the relevance of the topic. However, evidence on the effects of interventions that aim at restricting non-local demand in order to stabilize prices and stop urban sprawl is scarce.

There is a series of studies examining non-local buyer restrictions in China. Du and Zhang (2015) find that purchase restrictions in Beijing reduced the annual growth rate of real estate prices by 7.69 percent, while they find only smaller or no effects for the trial property tax rate in Chongqing and Shanghai. Yan and Ouyang (2018) find as well a substantial negative effect of house-sale restrictions on housing prices. Somerville, Wang and Yang (2020) exploit within city instead of inter-city variation on a purchase restriction and find no relative changes in housing prices between restricted and unrestricted areas. In general, all studies find much larger declines in volume than in prices. In contrast to the SHI in Switzerland, these interventions in China restrict the demand (an individual can only buy a fixed number of residential properties) instead of the supply. Consequently, the mechanism of the SHI is different from the Chinese purchase restrictions: The aim of the SHI is that free building land is only available for local people and thus all new buildings are no longer subject to non-local demand. Moreover, the focus of the initiators of the SHI is also very much on preventing splinter development.

In simultaneous and independent work, Hilber and Schöni (2020) examine as well the effects of the SHI. They estimate the short-run effect of the SHI on real estate prices using a different method than I do – the DD method. Hilber and Schöni (2020) pool the years 2010 and 2011 to obtain a pre-intervention period and the years 2013 and 2014 to obtain a single post-intervention period, while they drop the year of intervention, 2012. I use a longer time period from the year 2000 until 2018 to estimate the yearly effect in the longer-run instead and I am able to estimate whether the effects varies over time. Hilber and Schöni (2020) find a strong negative and significant effect on primary housing prices of approximately -15 percent in the pooled period of the second and third year after the vote (2013 and 2014).

Applying the SCM approach, I find no effect on prices in the first and second year after the intervention (2012 to 2013). However, I find a strong negative effect on prices of between -10 percent and -19 percent compared to the counterfactual in the third, fourth and fifth year after the vote (2014 to 2016). These negative effects are significant at a 99 percent level. Prices in affected municipalities remain below the counterfactual prices in the following years (2017 to 2018). However, it depends on the specification whether price differences in 2017 and 2018 are significantly different from zero. The analysis of the impact channels suggests that the second home initiative has led to lower demand through adverse effects on economic activity and legal uncertainty. The effect found by Hilber and Schöni (2020) in the pooled second and third year after the SHI is comparable to the effect of -10 percent to -19 percent I found for the third to

fifth year after the SHI. However, I do not find any effect in the second year after intervention. Further, I show indirectly that the effect on second homes must be negative and very similar to the effect on pre-law first homes not affected by the SHI, while Hilber and Schöni (2020) find a positive price effect on second homes.

This study makes mainly two contributions to the literature. *First*, it adds to the scarce evidence on the effects of interventions that aim at restricting the global emergence of second homes. Worldwide there is increasing resistance to second homes. This is leading increasingly to regulations designed to limit the potentially negative effects of second homes on the local population. This study contributes to understanding the (unexpected) effects of such regulations, which is important for implementing more targeted and effective solutions in the future.

Second, this study is among the first studies to implement the SCM with multiple treatments units and extends the classic SCM in order to compute (more) precise statistical inference. Acemoglu et al. (2016) and Kreif et al. (2015) apply comparable techniques to compute statistical inference. This study shows that the SCM might be a very attractive method to examine interventions in real estate economics with numerous and heterogeneous housing markets.

4.2 Background of the Second Home Initiative

In 2006, a committee called Helvetia Nostra started to collect signatures for the SHI. In 2007, Helvetia Nostra handed in more than 100,000 signatures to the Federal Chancellery and in January 2008, the Federal Chancellery validated those signatures and the Federal Council authorized the initiative. In 2011, the parliament followed the Council's decision. Consequently, Swiss citizens voted on the SHI in March 2012.[4] The main goals of the initiators of the SHI are to protect the landscape, stop splinter development and keep housing affordable for locals.

Most major political parties, most known economic organizations, the Federal Council, and parliament clearly recommended declining the SHI.[5] It is important to know that only a small minority of all popular initiatives held in Switzerland are accepted. Up to April 2020: only 22 of 217 initiatives have been accepted by popular vote.[6] Because of this broad resistance in politics and economics and the general tendency of initiatives to be turned down, most opponents of the initiative were quite confident that the SHI would be declined and, thus, did not start a vigorous campaign against the SHI.[7] In March 2012, a very narrow majority of 50.6 percent of all voters accepted the SHI. Although surveys predicted a tight race, the result was a surprise for most observers (as placebo studies in Figure 4.5 or submissions of construction permits in Figure 4.2 confirm).

The SHI was applied immediately after the vote in March 2012 (Ordinance on Second Homes 2012). Hence, the Federal Court declared all building permits for second homes in affected

[4] see Swiss Federal Chancellery for more information, URL: `www.bk.admin.ch/`
[5] see "Swiss Parliament" for more information, URL: `www.parlament.ch/`
[6] see Federal Statistical Office (FSO), URL: `www.bfs.admin.ch/`
[7] see Dulio, Claudio, "Die Zweitwohnungsinitiative unterschätzt", *Neue Zürcher Zeitung*, February 24, 2012.

municipalities submitted after the vote on March 11, 2012, invalid in retrospect. I.e. building permits for second homes submitted after the SHI can be prevented by objections. Although the Swiss government elaborated a provisional ordinance corresponding to the SHI in August 2012, it took almost three years for the parliament to work out the law. Parliament accepted the definitive law in March 2015 and began enforcing it on January 1st 2016. The ordinance of 2012 and the ultimate law of 2015 differ in some points, but they remained the same at their cores (see Section 4.3). Nevertheless, the vote in favor of the SHI in 2012 meant an immediate building freeze for second homes in affected municipalities and, hence, a sharp cut in supply.

4.3 Conceptual Framework and Impact Channels

As mentioned in the introduction, the SHI is a drastic intervention and affects particularly municipalities in the Alps regions (see Figure 4.1). Predicted effects can be separated into *direct* and *indirect* effects. Direct effects suggest an increase in prices, while indirect effects propose a negative effect on prices. In the next subsections, the mechanisms of direct and indirect effects are discussed. But first, a closer look at the law is required (Federal Act on Second Homes of 2015, Ordinance on Second Homes of 2012) to understand the impact channels.

The second home law separates the real estate market into two different categories:

1. *New first homes*: First homes, whose construction was still permitted after the vote in 2012. These homes can no longer be used as second homes. Their use is severely restricted.

2. *Second homes* and *pre-law first homes*: Homes declared as second homes can arbitrarily be used as first or second homes. Their use is unrestricted in the future. Pre-law first homes are homes that were either built or whose construction was permitted before the vote in 2012. These homes can arbitrarily be used, sold or even rebuilt and enlarged as first or second homes according to the law of 2015. However, according to the ordinance of 2012 (in effect until 2015), pre-law first homes can only be sold as second homes under the condition that the pre-law first home is not replaced by a new first home in the same municipality. This article in the ordinance of 2012 was very easy to avoid and was dropped in the final law. Hence, the use of pre-law first homes is virtually unrestricted by the SHI and legally equivalent to second homes. Because second homes and pre-law first homes are unrestricted in their use, I assume that these types of homes are substitutes. There is no reason why these two types of homes should vary substantially in layout. If this should be the case, owners always have the possibility to renovate, alter the layout or even rebuild and enlarge properties. Finally, since almost all affected municipalities are small towns in the Alps with highly restricted building zones, first and second homes are not located in different quarters (especially in high-amenity places with typical second home rates of more than 50 percent). Unfortunately, the Swiss Real Estate Datapool (SRED) data used in this paper does not allow to reasonably estimate the effect on first and second homes separately, because the information on the second home status of houses is not properly

collected and too deficient to deliver reasonable results (see Section 4.6.3 for a discussion). However, estimates in Section 4.6.3 suggest that the SHI affects first and second homes similarly. This finding supports the argument that pre-law first and second homes are rather close substitutes.

4.3.1 Direct Effects

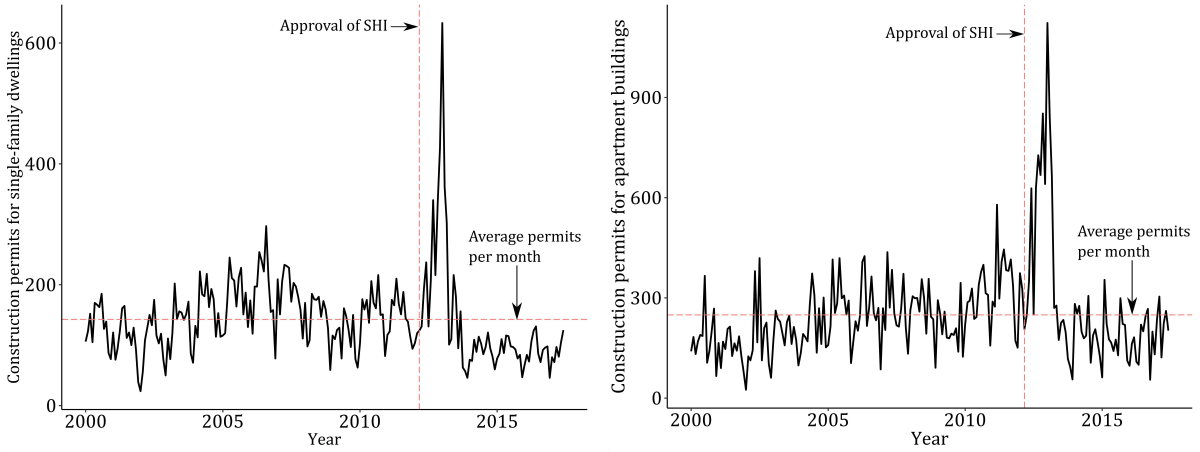
First, the *direct effects* of the SHI on demand and supply for real estate are considered. The separation in homes restricted (new first homes) and unrestricted in use (all others) leads to different expected consequences for different groups of houses in affected municipalities. The price of new first homes should decrease because they can no longer be sold as second homes. Especially in tourist regions, real estate prices are driven by non-local buyers with a high willingness to pay (Kaufmann and Rieder, 2012). Since the option to sell new first homes as second homes to international buyers with a high willingness to pay is gone, a portion of the demand is gone. By contrast, there is no construction ban for new first homes and, therefore, supply can be extended if needed.

On the other hand, is the use of second homes and pre-law first homes unrestricted, but the building freeze has caused a cut in supply. Non-local potential buyers have a high willingness to pay for second homes, especially given the shortage of supply (Kaufmann and Rieder, 2012). The cut in supply by law should cause an increase in the prices of these homes given a stable demand. Therefore, a separation of the market between second homes and pre-law first homes with increasing prices and new first homes with decreasing prices is expected. Since at least 92 percent[8] of all post-intervention transactions in treated municipalities involved pre-law first homes or second homes, the overall direct price effect of the SHI should increase housing prices in affected municipalities.

Although the SHI was supposed to cut the supply in affected municipalities, practitioners often argue that the SHI caused a last-minute glut of new construction project submissions: Many land owners became aware that they cannot build second homes on their land anymore as soon as the new second home law is enforced. Therefore, they applied for last-minute building permits shortly after the SHI vote in 2012. Even though these building applications submitted after the vote should not have been approved, there was a glut of building applications accepted right after the SHI vote took place. The number of approved building permits right after the vote in March 2012 was more than three times higher than the long-term average (see Figure 4.2). Two years after this panic-stricken submission of construction projects, the corresponding apartments came on the market and provided a one-time boost to supply. Although objections[9] prevented all the planned objects from being built, the initiative nonetheless might have sparked

[8] All first homes with construction finalized in 2013 or later are here considered new first homes. This is a very conservative estimate because only homes that received a construction permit after March 2012 are actually new first homes. A considerable number of homes finalized in 2013 or later may have received their building permit before the vote in March 2012.

[9] The Federal Court declared that buildings permits for applications submitted after the SHI vote are invalid. Hence, objections against these permits prevent the construction of such buildings.



(a) Permits for single-family dwellings (b) Permits for apartment buildings

Figure 4.2: Monthly construction permits in affected municipalities (*Source*: Baublatt/Credit Suisse).

a short-term boost in supply and lead to a delayed fall in prices when these newly built homes came on the market.[10] In Section 4.7, I check whether there is any evidence for this impact channel.

4.3.2 Indirect Effects

Adverse Effects on Local Economies

So far, the direct effects of the SHI on demand and supply have been considered. In a second step, the *indirect effects* of the SHI on prices will be discussed. As mentioned in the introduction, second homes are of great economic importance in affected regions. Not only do they guarantee many jobs in the locally very vital construction sector,[11] but they are also an important source of income for hotels and mountain railway companies. Hotels, mountain railways and other companies involved in the tourism industry sold second homes or land to cross-subsidize investments in infrastructure. Since this ability to cross-subsidize is gone, tourism resorts might no longer be able to maintain their costly (touristic) infrastructure (Codoni and Grob, 2013; Kaufmann and Rieder, 2012). If tourism infrastructure is worsening, tourism demand and thereby demand for second homes will decrease. Because of the decrease in tourism demand and lower economic activity, the tax income of affected municipalities will drop further, and municipalities might face additional difficulties maintaining their infrastructure. Since taxes are paid on the

[10] See for example Kohler, Franziska, "Wo die Wohnungspreise in den Bergen jetzt tiefer sind", *Tages-Anzeiger*, December 11, 2016 or Martel, Andrea, "Kein Run mehr auf Ferienwohnungen", *Neue Zürcher Zeitung*, July 17, 2017 for a discussion of the one-time boost in supply due to the SHI.

[11] In 2015, 8.1 percent of the labor force in the mountain cantons (Uri, Obwalden, Nidwalden, Glarus, Graubünden, Ticino and Valais) was employed in the construction sector. The Swiss average share of the labor force employed in the construction sector is 6.5 percent. In the two cantons with the highest share of affected municipalities, 8.8 percent (Graubünden) and 8.7 percent (Valais) of the labor force is employed in the construction sector (see FSO Structural Survey 2015).

primary residence, second home municipalities have by definition a small tax base compared to the number of houses and face high infrastructure costs (e.g., ski lifts). Therefore, municipalities affected by the SHI are especially vulnerable to such tax income reductions. In consequence, some municipalities were forced to introduce a second home tax in order to be able to maintain their ski lifts and other infrastructure projects. The introduction of second home taxes increases the implicit price of a second home. Therefore, the SHI might have reduced local and non-local demand in affected municipalities, which leads to a decline in prices.

Hilber and Schöni (2020) develop a formal model that explores the housing and labor market impacts of a ban on second homes in a general dynamic equilibrium setting and formalizes some of the arguments mentioned above. The predictions of the model crucially depend on whether pre-law first and second homes are poor substitutes or not. The effect of the SHI on real estate prices is ambiguous, if first and second homes are close substitutes: There is a negative effect on local wages that decreases the aggregate demand for housing on the one hand side, while the SHI causes a cut in supply of pre-law first homes and second homes on the other hand side. If first homes and second homes are poor substitutes the model predicts a decrease in first home prices and an increase in second home prices. As discussed above, I assume pre-law first and second homes to be close substitutes. I show that the effect on first and second homes is very similar (see Section 4.6.3), what supports the assumption that pre-law first homes and second homes are rather close substitutes.

Legal Uncertainty and the Lock-In Effect

Another very important point is that the SHI created insecurity in the local real estate markets. As stated in Section 4.2, after the vote, it took parliament three years to reach an agreement on the final second home law in 2015. Because it was not clear what the final law would look like, many market players might have been more conservative when selling or buying real estate in affected municipalities. This causes a decrease in transactions. Hence, the legal uncertainty might cause a "lock-in" effect, where second homeowners do not sell their homes if there is no immediate need to do so. Furthermore, homeowners know that it will be difficult to buy a second home in an affected municipality in the future due to the building freeze and do not sell their properties at all. With this in mind, homeowners might delay transactions in the hope that in the longer run, prices will increase due to the building freeze. This lock-in effect might lead to a situation in which primarily those in need of liquidity sell their homes. Therefore, houses sold after the vote in 2012 might be of lower quality on average. Hence, average housing prices decrease because the hedonic characteristics of houses changed.

In summary, it remains unclear whether the SHI is supposed to increase or decrease real estate prices, since there is a price increasing direct effect and opposing indirect effect via a drawback on local economies or a lock-in effect. Furthermore, there might even be a negative effect due to a one-time surge in supply. Which of these effects dominates needs to be clarified empirically.

4.4 Empirical Strategy

Figure 4.1 demonstrates where the affected municipalities are located. Almost all treated municipalities are found in the Alps and barely any are in the Swiss Mittelland. Hence, as argued above and substantiated in Section 4.6.1, a DD approach does not appear to be suitable in this case. Therefore, the SCM developed by Abadie and Gardeazabal (2003) and Abadie, Diamond and Hainmueller (2010, 2015) is chosen. The point is that almost all affected municipalities are located in the Alps, but not all municipalities in the Alps are affected. Further, there exist as well other control municipalities which are not located in the Alps that are comparable to treatment municipalities. Hence, there are still control municipalities in the donor pool that are similar to affected municipalities. The SCM attempts to construct an optimal synthetic control by assigning different weights to different control units. In the next subsections, this data-driven procedure is presented.

4.4.1 Classic Synthetic Control Method

The starting point is the classic SCM following Abadie and Gardeazabal (2003) and Abadie, Diamond and Hainmueller (2010, 2015). Suppose that our dataset contains J+1 different units (in this case, units are municipalities). One of these units (j=1) is the treatment unit, all other J units (j=2,..., J+1) are control units. Furthermore, the dataset contains T periods. T_0 of these periods are pre-treatment periods (t=1, 2,..., T_0), and T_1 periods are post-treatment periods (T_0+1,..., T). $\mathbf{W}$ is a (Jx1) vector $\mathbf{W}$=(w_2,..., w_{J+1}) of non-negative weights, such that $w_j \geq 0$ for all j and $w_2 + w_3 + \ldots + w_{J+1} = 1$. Each scalar w_j of this vector represents the weight of one control unit. The weights $\mathbf{W}$ shall be chosen to ensure that the synthetic control closely resembles the treatment unit before the intervention. $\mathbf{X}_1$ is a (Kx1) vector of K pre-treatment characteristics of treatment unit j=1 including the pre-treatment outcome variable. $\mathbf{X}_0$ is a (KxJ) matrix containing K pre-treatment characteristics of J control units. Further, $\mathbf{V}$ is a diagonal matrix containing the relative importance of each of these pre-treatment predictors. The goal is to find the $\mathbf{W}^*$ that assigns the optimal weight to every control municipality in order to minimize the pre-intervention distance metric of real estate characteristics, including the pre-intervention outcome of the treatment unit and the synthetic control of one year. To find $\mathbf{W}^*$, the problem

$$\underset{W \in \omega}{min}(\mathbf{X}_1 - \mathbf{X}_0\mathbf{W})'\mathbf{V}(\mathbf{X}_1 - \mathbf{X}_0\mathbf{W}) \tag{4.1}$$

must be solved, where $\omega = \{(w_2, w_3, ..., w_{J+1})\}$ is subject to $w_j \geq 0$, and $w_2 + w_3 + \ldots + w_{J+1} = 1$. The $\mathbf{W}^*$ that solves (4.1) is the vector of weights, which gives each control unit a weight such that the synthetic control best resembles the treated unit in the pre-intervention period. In this paper, a data-driven approach is applied to select an optimal $\mathbf{V}^*$ that minimizes the root mean squared error (RMSE) of the outcome variable in the pre-intervention period, as done in Abadie and Gardeazabal (2003) and Abadie, Diamond and Hainmueller (2010, 2015).

As soon as I obtain $\mathbf{W}^*$, the synthetic control can be computed:

$$\hat{Y}_{1t}^{SC} = \sum_{j=2}^{J+1} w_j^* * Y_{jt} \tag{4.2}$$

Y_{jt} is the outcome variable (i.e., the real estate price or number of transactions) in municipality j and time period t. $\widehat{Y}_{1t}^{SC}$ is the synthetic control and is supposed to be the counterfactual of the treatment unit. The gap between the treatment unit and the synthetic control is

$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* * Y_{jt} = Y_{1t} - \hat{Y}_{1t}^{SC}. \tag{4.3}$$

Since $\mathbf{W}^*$ is minimizing the pre-intervention distance metric of real estate characteristics between the treatment unit and the synthetic control, pre-intervention gaps should be close to zero, i.e., $\widehat{\alpha}_{1t} \approx 0, \forall t \leq T_0$. Because only the treatment unit receives the intervention, the post-intervention gaps are supposed to be the causal effect of the intervention.

4.4.2 Multiple Treatment Units and Exact Inference with Permutation Tests

As mentioned in the introduction, I need to extend the classic SCM. The classic SCM only deals with one treatment unit. In this paper, 89 treatment units have to be considered. There is not much literature dealing with that many treatment units. Abadie, Diamond and Hainmueller (2010) suggest to simply aggregate the treated units into a single treated unit. Billmeier and Nannicini (2013) deal with several treatment units but look at effects for each treatment unit separately. Cavallo et al. (2013) also deal with several treated units, when estimating the causal effect of different natural disasters in different countries on GDP development. However, Cavallo et al. (2013) do not attempt to estimate one single intervention's overall effect on different units but many different treatments on many different units. Nevertheless, Cavallo et al. (2013)'s approach is comparable to that applied in this paper. Meanwhile, Kreif et al. (2015) and Acemoglu et al. (2016) come up with an approach closely related to the approach I am applying in this paper. They compute a synthetic control for each treatment unit and then calculate an aggregate effect.

The following approach is applied in this paper. Instead of having one treatment unit j=1, I have J_0 treatment units and J_1 control units (where $J_1 \gg J_0$). First, I compute a synthetic control and the corresponding gaps, $\hat{\alpha}_{jt}$, for all J_0 treatment units separately, as in the classic SCM. Hence, I obtain J_0 different gaps per year. To obtain an average effect per year, I compute the average of all J_0 gaps re-weighted by the number of transactions per year and municipality, L_{jt}:

$$\overline{\alpha}_t = \frac{\sum_{j=1}^{J_0} \hat{\alpha}_{jt} * L_{jt}}{\sum_{j=1}^{J_0} L_{jt}} \tag{4.4}$$

I weight the gaps to give municipalities with numerous transactions per year a higher weight

than small municipalities with only a few transactions per year. Based on these weighted gaps $\overline{\alpha}_t$, I can reconstruct an aggregate weighted synthetic control $\hat{Y}_t^{SC} = \overline{Y}_t - \overline{\alpha}_t$, where $\overline{Y}_t$ is the average price of the outcome variable of treatment units weighted by the number of transactions. Hence, by re-weighting the single gaps, I am able to construct aggregate gaps and an aggregate synthetic control.

I further compute the ratio of the post- and the pre-intervention RMSE of the treatment units and the synthetic control:

$$RMSE\ ratio = \sqrt{\frac{(\sum_{t=T_0+1}^{T} \overline{\alpha}_t)^2}{T_1}} / \sqrt{\frac{(\sum_{t=0}^{T_0} \overline{\alpha}_t)^2}{T_0}}, \tag{4.5}$$

where the time periods (t=0,1,..., T_0) are pre-intervention and periods (t=T_0,..., T) post-intervention years. The ratio of the post- and pre-intervention RMSE should be larger than 1, since the gaps $\overline{\alpha}_t$ are supposed to be substantially larger in the post-intervention period than in the pre-intervention period. The ratio of post- and pre-intervention RMSE is an important indicator, since it reflects the magnitude of the causal effect relative to pre-intervention fit. The worse the pre-intervention fit (the larger the gaps), the higher are the expected gaps in the post-intervention period. The RMSE-ratio takes the pre-intervention fit into account when assessing the magnitude of the intervention effect (see Abadie, Diamond and Hainmueller, 2010). To evaluate the statistical significance of the results in the next step, the RMSE-ratio is required.

Following Abadie, Diamond and Hainmueller (2010), I assess the significance of the estimates by conducting a series of placebo studies. The difference from Abadie, Diamond and Hainmueller (2010) is that my treatment group consists of J_0 treatment units instead of one treatment unit. This renders an opportunity to extend the approach of Abadie, Diamond and Hainmueller (2010) slightly. I construct a placebo group consisting of J_0 randomly chosen control units and apply the SCM used to estimate the actual treatment effect of the SHI to this placebo group. Because none of the placebo units received the treatment, the pre- and post-intervention gaps should be similar. Therefore, the placebo RMSE-ratio should be close to 1 or at least smaller than the treatment RMSE-ratio. I can iterate this placebo study almost an arbitrary number of times (say N times) with different randomly chosen groups of J_0 control units, because $J_1 \gg J_0$. This renders it possible to obtain far more statistical power than in the classic SCM. This iterative placebo procedure provides me with a distribution of RMSE-ratios for municipalities that never received treatment. This distribution can be used to compute the statistical significance of our treatment effect estimation by computing the corresponding p-value. The p-value reflects the probability of obtaining a placebo RMSE-ratio larger or equal than the treatment RMSE-ratio:

$$p - value = Pr(ratio_n^{plac} > ratio^{treat} | H_0) = \frac{\sum_{n=1}^{N} I(ratio_n^{plac} > ratio^{treat})}{N}, \tag{4.6}$$

where N is the number of iterations, H_0 is the null hypothesis that the SHI has no effect on prices or transactions, $ratio_n^{plac}$ is the RMSE-ratio of placebo iteration n, and $ratio^{treat}$ is the original treatment RMSE-ratio.

Multiple treatment units additionally allow confidence intervals (CI) to be computed as in Acemoglu et al. (2016). When conducting N placebo studies as described above, I obtain N weighted placebo gaps, $\overline{\alpha}_{nt}$, per year. I then compute the standard deviation of these N gaps. Using the standard deviation, CI can easily be computed using average treatment prices as the basis.

4.5 Data and Descriptive Statistics

Swiss Real Estate Datapool (SRED)[12] collects the data used in this paper. SRED is an association founded by the three Swiss banks UBS, Credit Suisse and Zürcher Kantonalbank. The SRED dataset contains information on real estate transactions executed by these three banks between 2000 and 2018. It contains information for more than 240,000 transactions completed during this period and includes transaction prices as well as other relevant attributes.

To define whether a transaction takes place in a treatment or a control unit, we need to know the second home share of each municipality. The Swiss Federal Spatial Development Office (ARE) provides the official second home share per municipality. However, the second home share of municipalities was unknown before the vote in 2012. Therefore, the ARE had to estimate these second home shares. Municipalities had the opportunity to ask for a revision of ARE's second home share estimation. When these municipalities were able to prove that their second home share was lower than that estimated by ARE, the ARE adjusted its original estimation. I dropped all municipalities that asked for such a revision of their estimated second home share because market players did not know whether municipalities that asked for revision end up as treated or untreated municipalities, until ARE accepts or rejects the objection (see Figure 4.1 for municipalities that revised the original second home estimation). Furthermore, I use the administrative data of the "Gebäude- und Wohnungregister" (GWR) provided by the FSO to estimate the effect of the SHI on the housing stock and data on unemployment provided by the State Secretariat for Economic Affairs (SECO) to estimate the effect of the SHI on unemployment rates. Finally, the "Baublatt" in cooperation with the Credit Suisse collects all building permits on a monthly basis. I use these data in order to see if there actually was an increase in building permits right after the vote (Figure 4.2).

In order to apply the SCM I only keep municipalities with at least one transaction in every year. Furthermore, I dropped all municipalities with a second home share between 18 percent and 20 percent. This is done because market players might foresee that these municipalities are going to cross the threshold of 20 percent of second homes in the near future. Therefore, the SHI might affect these municipalities to a certain degree. Sensitivity calculations considering second homes build since 2012 show that only a few of the municipalities with a second home share between 18 percent and 20 percent are at risk of belonging to the treated group in the near future. Therefore, it is rather conservative to drop all municipalities in this bandwidth. The original dataset in total contains transactions in 2209 municipalities. After the data preparations

[12] See `https://www.sred.ch/`

Table 4.1: Summary statistics, averages of transactions by treatment and control group.

	Control group		Treatment group	
	Pre-interv.	Post-interv.	Pre-interv.	Post-interv.
Transaction price[a]	761,246	1,002,434	628,214	785,683
Transactions [b]	66.2	53.7	34.6	22.3
Number of rooms	4.78	4.35	3.79	3.60
Plumbing units	2.05	2.01	1.82	1.76
Number of garages	1.1	0.83	0.95	0.71
Micro-location[c,d]	2.87	2.71	3.08	2.93
Quality[d]	2.87	2.87	2.84	2.59
State[d]	3.17	2.99	2.88	2.56
Year of construction	1983	1988	1984	1984

Notes: *a* In Swiss Francs (CHF)
b Per year and municipality
c Micro-location is the quality of the location of a property within the municipality
d Values between 1 (=poor) and 4 (=very good)

mentioned above, 613 or approximately 30 percent of these 2209 municipalities remain in the final dataset. It is important to know that almost all of the removed municipalities were dropped because no transaction took place in one or more years. This means that mainly very small municipalities were dropped. Therefore, the final dataset of 613 municipalities contains 186,508 or approximately 77 percent of all transactions in the original dataset.

Because the vote on the SHI took place in mid-March 2012, I prepared the data so that every year starts in the second quarter and ends after the first quarter of the following year.[13] Hence, the post-intervention period (2012 to 2018) starts in April 2012, right after the vote took place.

A summary of the most important housing characteristics is presented in Table 4.1. A typical dwelling in an unaffected municipality is more expensive and bigger when it comes to the number of rooms, plumbing units or garages than a typical dwelling in affected municipalities. Furthermore, housing markets in affected municipalities, with 35 yearly transactions in the pre-treatment period, are clearly smaller than those in the unaffected municipalities, with 66 transactions in the same period. This underlines that the housing markets of the two groups on average differ clearly in size and characteristics. In both types of municipalities, the number of transactions and most other indicators such as the number of rooms decrease over time. Figure A.1 presents the price development in control and treatment municipalities. Prices in treated municipalities are stagnating after the vote in 2012, while prices in unaffected municipalities continue to increase.

[13] For instance, the year 2000 starts in April 2000 and ends in end-March 2001.

4.6 Results

4.6.1 Robustness Checks for the DD Approach

As discussed in Section 4.1, municipalities affected by the SHI are placed in the Alps and unaffected municipalities are mostly located in the Swiss Mittelland. Therefore, the average control municipality does not seem to be a suitable counterfactual for affected municipalities and the parallel trend assumption of the DD strategy is likely to be violated. While it is not possible to directly test the parallel trends assumption, there are several checks that are able to narrow down whether the key identification assumption holds or not. In this section, all tests commonly mentioned in the literature (see e.g. Angrist and Pischke, 2008; Wing, Simon and Bello-Gomez, 2018, for an overview) are applied to further investigate whether the parallel trends assumption holds in the case of the SHI context.[14]

One possible check is to estimate the effects of placebo interventions before the actual SHI vote took place. The placebo DD estimations assume, for example, 2007 or any other pre-intervention year, to be the year of the SHI vote. Then, the DD method is used to test whether any significant effect of this placebo vote is found in the remaining pre-intervention period 2007 to 2011 (the actual vote took place in 2012). Because there was no vote in 2007, there should be no significant effect for the 2007 to 2011 period. If there are significant effects of such placebo interventions, the common trend assumption is not credible. I conducted such placebo tests for the pre-intervention years 2006, 2007, 2008, 2009, 2010. Results in Tables A.1 and A.2 in Appendix A.2 show that these placebo intervention effects are significant and, hence, that the DD is not likely to be a suitable identification strategy for the SHI context.[15]

Further, there are two common checks for the validity of the common trend assumption, the inclusion of group-specific time trends and causality tests in the spirit of Granger. If the group-specific time trends are significantly different from each other or if their inclusion notably changes the estimated effect of the intervention, this is discouraging for the parallel trends assumption (see Besley and Burgess, 2004, for an application). As shown in Tables A.3 and A.4 in Appendix A.3 the difference in group-specific time trends is significant and the estimated treatment effect changes clearly when different group-specific time trends are allowed. Allowing treatment and control group to follow different trends changes the effect of the SHI from a significant negative effect to a (significant) positive effect when using transaction-level data (see Table A.3). When aggregating the data on municipality level, as well a strong change of the estimated effect is found (see Table A.4). Hence, the inclusion of group-specific time trends shows in a revealing way that the parallel trends assumption is likely to be violated.

[14] These tests are all quite closely related to each other and test very similar properties. Nevertheless, I conduct all tests in order to provide a complete picture.

[15] Once I ran the DD placebo tests with transaction level data in order to keep as much information as possible in the data (see Table A.1). However, in order to conduct the SCM method I have to aggregate the observations on municipality level and make sure that the data is balanced, i.e. make sure that every municipality has at least one observation per year. Hence, I ran additional DD placebo tests with data containing municipality level observations, which is the same as the data used in the SCM (see Table A.2).

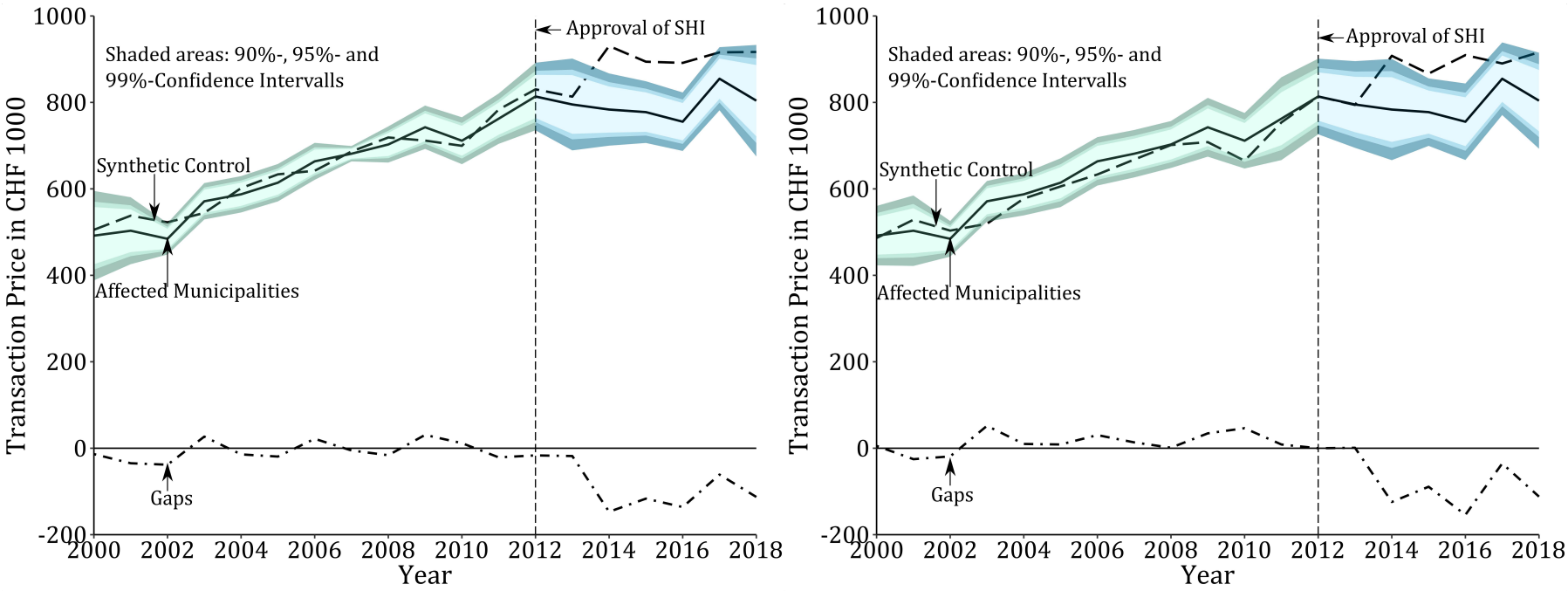
Finally, I compute lags and leads of the effect similar as in a Granger causality test. The idea is to check, whether past interventions predict outcomes while future interventions do not (see Angrist and Pischke, 2008 or for an application Autor, 2003). Table A.5 in Appendix A.4 shows that effects in several of the five years before the SHI took place are significantly different from zero for transaction-level data (see columns 1 and 2). Effects in the years before the SHI took place are as well significantly different form zero when using municipality level data (see columns 3 and 4).

In conclusion, the validity of the common trend assumption was already doubted, because treatment group municipalities are located in the Alps and are mostly remote municipalities. Control municipalities are partly as well located in the Alps, but the control group is clearly dominated by the densely populated Mittelland region including all Swiss major cities. All three checks conducted in this section, placebo tests, inclusion of group-specific time trends and causality test in the spirit of Granger, indicate that the common trend assumption is violated in the context of the SHI. Therefore, DD seems not to be suitable to estimate the effect of the SHI. For this reason, the SCM is applied in the following Sections.

4.6.2 Main Results Using the Synthetic Control Method

In this section, results of the SCM as illustrated in Section 4.4 are presented. As mentioned in Section 4.4.1, I minimize the distance metric of real estate characteristics between the treatment unit and control municipalities. Real estate characteristics include the number of transactions per year, number of rooms, plumbing units, garages, micro-location, quality and state of the property as well as the pre-intervention outcome variable data point of one year, i.e. the transaction price of 2007. Kaul et al. (2017) point out that using too many pre-intervention outcomes could cause a bias. Therefore, I run the SCM two times: First, I include only the price level of 2007, and second, I do not include any pre-intervention outcome variables. Figure 4.3 shows the effect of the SHI on the prices taking both approaches. Both approaches return very similar results. Because the approach including one pre-intervention outcome variable data point offers a better pre-intervention fit, I focus on the approach including the pre-intervention outcome of 2007.

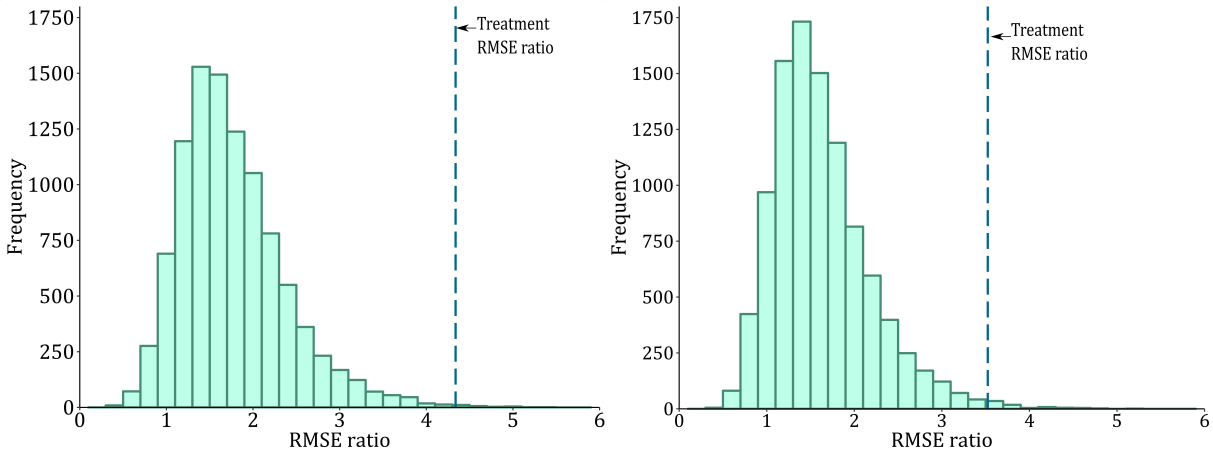
Figure 4.3a indicates that the SHI has a strong negative effect on housing prices, as the treatment and synthetic control prices diverge clearly in 2014 and later. The pre-intervention RMSE is very small, with about 23,000 Swiss Francs (CHF) or 3.56 percent of the average pre-intervention treatment price, which indicates that the synthetic control is an accurate counterfactual. Housing prices are in 2014 (-19 percent), 2015 (-14 percent) and 2016 (-18 percent) clearly lower than the corresponding synthetic control prices. For instance, in 2016, the average treatment housing price was CHF 136,000 lower than the synthetic control price. However, no effect in the post-intervention years 2012 and 2013 is visible. There seems to be a rebound in prices in 2017, where the gap is only -7 percent of the treatment group price. However, the gap widens again in 2018 (-14 percent) and is still significant on the 95 percent level for both years, 2017 and 2018. Hence, prices in treatment regions did not yet recover from the drop in 2014. The RMSE of the post-intervention period is with CHF 101,000 or 11 percent of the average post-intervention



(a) Outcome of 2007 included

(b) No outcome variable included

Figure 4.3: Price development of treatment group and synthetic control and corresponding gaps for both approaches including one and no pre-intervention outcome variable data point to compute synthetic control.
Notes: CI in both approaches are based on 10,000 placebo runs.



(a) Outcome of 2007 included

(b) No outcome variable included

Figure 4.4: Distribution of 10,000 placebo RMSE-ratios.

price clearly higher than in the pre-intervention period. This results in a RMSE-ratio of 4.33. In 2002, the synthetic control is not able to reproduce the treatment price as closely as in other pre-treatment years. However, the gap in 2015 (the smallest gap of the 2014 to 2016 period) is still almost twice as that in 2002.

In Figure 4.3b, we observe a very similar effect for the approach without using the pre-intervention outcome of 2007. In fact, most numbers are very close to those discussed in Figure 4.3a. However, the rebound in prices in 2017 is stronger in this approach. The gap in 2017 is with -4 percent considerably smaller than in Figure 4.3a and not significant anymore. However, the gap increases again in 2018 (-14 percent) and becomes significant on a 95 percent level.

What is the probability of obtaining results of this magnitude by chance? To evaluate the significance of the results obtained above, I run placebo tests as described in Section 4.4.2. I run 10,000 permutation tests and correspondingly obtain 10,000 placebo RMSE-ratios. These 10,000 placebo iterations can be used to construct CI as described in Section 4.4.2. In Figures 4.3a and 4.3b these CI are reflected in the shaded areas. The synthetic control prices in 2014 to 2016 are outside the 99 percent CI and thus, the price decrease in affected municipalities compared to the synthetic control is highly significant on a 99 percent significance level in those years. As mentioned before, there is no effect on prices in the first two years after the approval of the SHI.

As explained in Section 4.4.2, we are going to look as well at the RMSE-ratio distribution of the placebo permutations. This approach is considering the significance of the total post-intervention period instead of individual years as in the CI approach. Because there was no effect in the first two years after the approval of the SHI, the significance of the effect might be lower according to the RMSE-ratio approach. The distribution of these placebo RMSE-ratios is presented in Figures 4.4a and 4.4b. When including one outcome data point in the synthetic control computation only 23 placebo RMSE-ratios are greater than the treatment RMSE-ratio of 4.33 (see Figure 4.4a), which corresponds to a p-value of 0.002 (see Equation 4.6 for more information on the calculation). Hence, the the overall post-intervention effect in Figure 4.3a is

significant with a 99 percent significance level, although there was no effect found in two of five post-intervention years. When not including the outcome variable, only 70 placebo RMSE-ratios are bigger than the treatment ratio of 3.52 (see Figure 4.4b). Hence, in this case the overall post-intervention effect is as well significant on a 99 percent level.

4.6.3 Robustness Checks

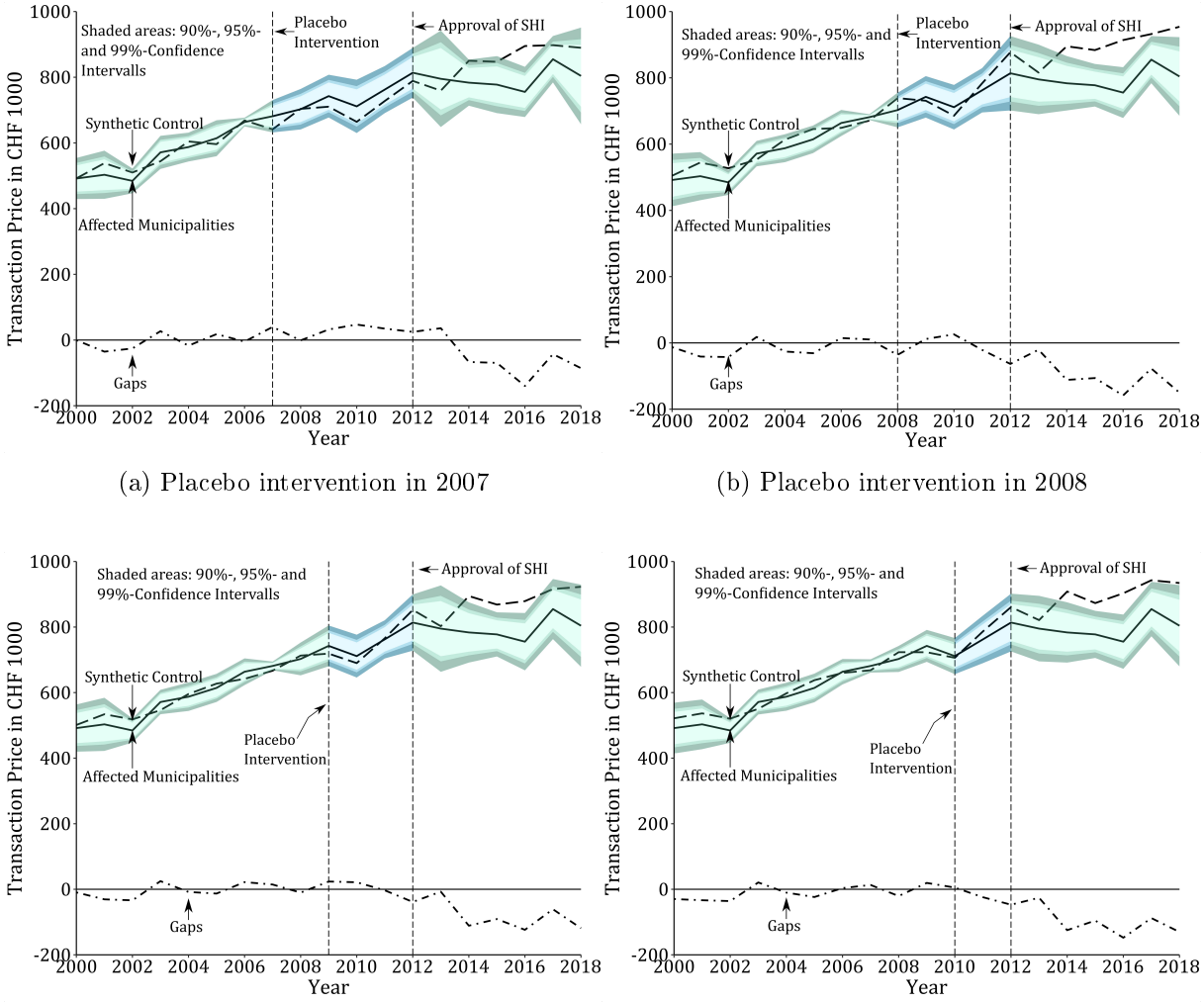
Placebo In-Time

The idea behind the placebo in-time approach is that I introduce a placebo intervention before the actual intervention took place.[16] Therefore, there are three periods in this approach: The pre-placebo period, the between placebo intervention and actual intervention period and the post-intervention period. With this approach I can answer two questions: First, I can check whether the SHI was approved surprisingly or whether it was foreseen by market agents (see 4.2 for a short discussion). If the SHI approval was no surprise, there should be an effect before the actual approval in 2012. Furthermore, I can check whether the gaps of the magnitude found in the baseline estimation above occur as well after a placebo intervention.

Figure 4.5 shows the results of four different placebo interventions from 2007 to 2010. We are mainly interested in the period between placebo intervention and actual intervention (blue-shaded part in Figure 4.5). Because there was no actual treatment in these placebo years, there should be no effect. Therefore, we expect the counterfactual price to be within the inner CI band before the the SHI was accepted in 2012. Using the example of the placebo intervention in 2008 (see Figure 4.5b), we see that the pre-placebo RMSE (CHF 27,000) is larger than the RMSE of the period between placebo intervention and actual intervention (CHF 25,000). This corresponds with a p-value of 0.92 under the null hypothesis that prices between treatment and synthetic control do not differ in the period between placebo and actual intervention. Moreover, synthetic control price trend never leaves the inner CI band in the between-period. Hence, there is no significant placebo effect in the period between the placebo and the actual intervention. The same applies to all four placebo interventions in Figure 4.5. This supports the claim that the results obtained do not occur simply because I stop minimizing the distance between synthetic control and actual treatments in 2012. Furthermore, it confirms that agents did not foresee the outcome of the vote in 2012 and did not adapt their behavior accordingly.

As in the benchmark approach, there is no significant effect in the first two years after the actual intervention in 2012 and a significant negative effect three to five years after the intervention in all placebo approaches in Figure 4.5. Thus, post-treatment effects are very similar to the benchmark results.

[16] The idea behind this kind of placebo tests is the same as behind the DD placebo tests in Appendix A.2.



(a) Placebo intervention in 2007

(b) Placebo intervention in 2008

(c) Placebo intervention in 2009

(d) Placebo intervention in 2010

Figure 4.5: Placebo in-time approach with placebo interventions in 2007 to 2010.
Notes: CI in all approaches are based on 10,000 placebo runs and the outcome of 2004 is included in the construction of the synthetic control.

Only Alpine Municipalities

As shown in Figure 4.1, almost all affected municipalities are located in the Alps. This is the reason, why I do not use the DD approach but SCM to make sure our treatment units and synthetic controls are comparable. However, one might argue that there are events (e.g. adverse effects on tourism industry) which hit Alpine regions harder than any other municipalities.[17] Therefore, I keep only those municipalities in the dataset, which are officially declared as Alpine municipalities according to the FSO.[18]

[17] Note, that the synthetic control of the baseline estimation consists mainly of municipalities located in the Alps (weight of about 65 percent) and/or municipalities for which tourism is an important sector. Thus, it is unlikely that there are shocks which affect only the treatment municipalities but not the synthetic control.

[18] The FSO follows the European mountain areas delineation. I include in this robustness check only municipalities which are classified as "Alpine". See FSO "Räumliche Typologien", URL: `www.agvchapp.bfs.admin.ch/de/typologies/query`

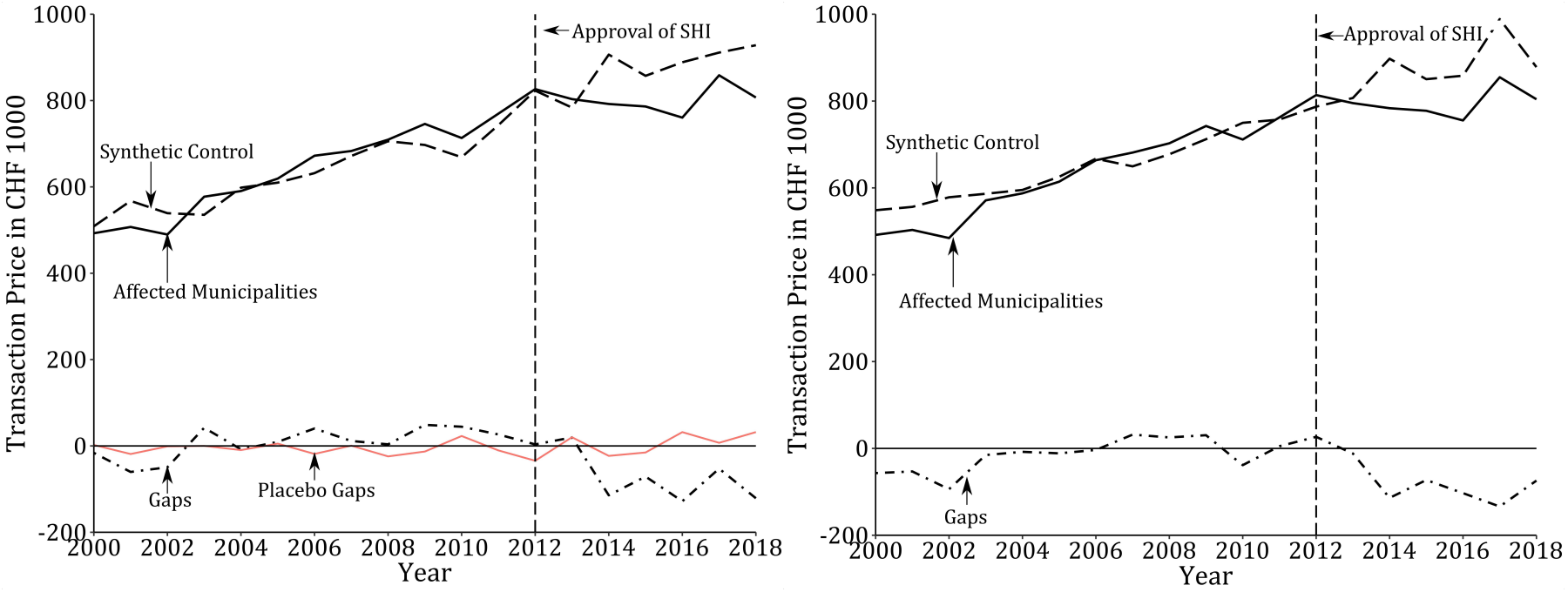
85 of 89 municipalities of the treatment group remain in the dataset when I include only Alpine regions. This shows again, that the SHI affected almost only this one region. On the other hand only 82 of formerly 524 control municipalities remain in the donor pool. Since the donor pool is smaller, we expect a less accurate fit in the pre-intervention period. However, the 82 Alpine control municipalities that remain in the donor pool are likely to be similar to the treatment municipalities, since they are located in the same region. The result is shown in Figure 4.6a. The pre-intervention RMSE is with CHF 35,000 higher than in approaches with bigger donor pools. However, the pattern is very similar to the baseline approaches including all municipalities. There is no effect in the first two years after the intervention, but a strong negative effect three to five years after the intervention (2014 to 2016). The magnitude of the effects in these years is almost the same as in the benchmark approach with -14 percent in 2014, -9 percent in 2015 and -17 percent in 2016. As well as in the benchmark estimation there is a rebound in 2017 (-6 percent), but in 2018 the gap widens again (-15 percent). Hence, the effect of the SHI remains mostly the same, if we include only Alpine regions and the effect seems not to be caused by an adverse effect other than the SHI that only affects the Alpine region. There are no CI computed, because there are less units in the donor pool (82) than in the treatment pool (85). Nevertheless, I used the 82 control municipalities as placebo units and computed a single synthetic control for them. The gaps between this placebo group and its synthetic control are shown in red ("Placebo Gaps") in Figure 4.6a. These post-intervention placebo gaps are clearly smaller than the original gaps.

Neighbor Municipalities

In general, municipalities tend to be more similar to their neighboring municipalities than to municipalities located in different regions. Therefore, control municipalities next to treatment municipalities are likely to receive high weights when constructing the synthetic control. Since no more second homes can be built in treatment municipalities, it is possible that potential buyers are interested in purchasing a second home in the closest municipality not affected by the SHI. Thus, the SHI could also affect the neighbors of treatment municipalities (see Figure 4.1 for the location of neighbors of treated municipalities). This would bias the former results and violate the stable unit treatment value assumption (SUTVA).

Only 10 neighboring municipalities (of 56 in the final dataset) receive a weight of more than 1 percent in the SCM estimation in the main analysis. Nevertheless, in total, almost 64 percent of the synthetic control in the main analysis consists of neighbors of the treatment units. Hence, if the SHI affects these neighboring municipalities, the results obtained earlier could be biased.

In a further robustness check I exclude all neighbor municipalities of the treatment group from the sample to make sure that spillovers on neighbor municipalities do not harm my estimation. Because I reduce the number of municipalities in the donor pool and these neighbor municipalities are likely to be very similar to the treatment group, the pre-intervention fit is expected to be less accurate than in the baseline estimations with bigger donor pools. The results of this approach can be found in Figure 4.6b. The pre-intervention RMSE is with CHF 38,500 bigger than in the



(a) Only Alpine Municipalities

(b) Excluding neighbor municipalities

Figure 4.6: Robustness checks: Approach using only Alpine regions as defined by FSO and estimation excluding all municipalities next to affected municipalities from the donor pool.

Notes: Outcome of 2007 is included in both estimations.

baseline estimations. But again, the pattern remains very similar: There is no effect found in the first two years after intervention, but a strong and negative effect in all later years. While the effect is very similar to the benchmark effect in 2014, the magnitude of the effect in the years 2015 to 2016 is negligibly lower than in the benchmark estimation (between -16 percent and -12 percent in 2015 and 2016 compared to -18 percent to -14 percent in the baseline estimation). However, in contrast to the benchmark estimations, there is no rebound effect in 2017. Because the effects without neighbors are very similar to those in the baseline estimation, except for the effect in 2017, I conclude that my estimation is not significantly biased by spillovers.[19]

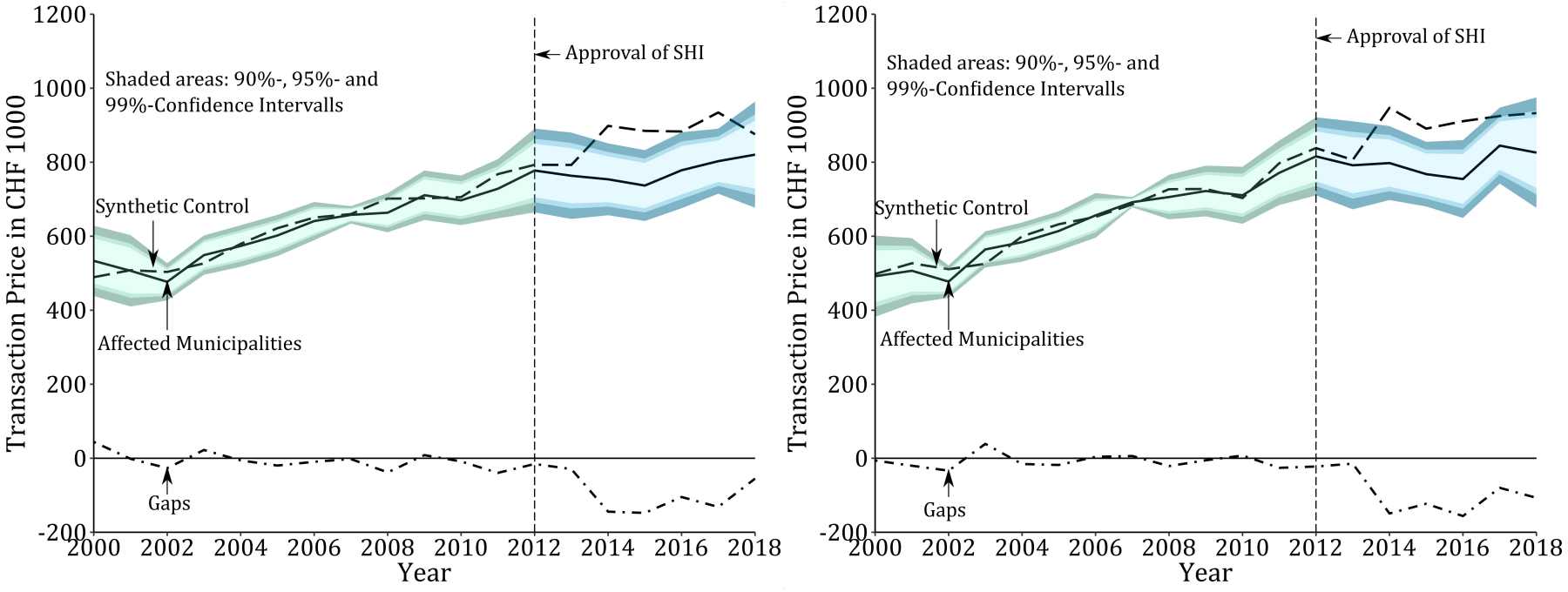
Heterogeneity of the Effect: First Homes vs. Second Homes

As mentioned in Section 4.3, the effect on second homes and pre-law first homes is supposed to be different, if second homes and pre-law first homes are poor substitutes: We would expect an increase in prices of second homes and a decrease in prices for pre-law first homes. I assume that second homes and pre-law first homes are rather good substitutes (see Section 4.3). This assumption could indirectly be tested by estimating the effect separately for second and first homes.

However, the information on the second home status of transactions in control municipalities is very unreliable in the SRED data: Only 0.37 percent of all pre-intervention transactions in control municipalities are declared as second homes, i.e. not even 47 of about 12,000 transactions per year in *all* control municipalities in Switzerland. Given the average administrative second home rate of more than 10 percent in control municipalities this information does not appear to be reliable. This does not change much after the vote: In the first three years after the vote (2012 to 2014) only 0.3 percent of all transactions in control municipalities involved second homes (i.e. 30 observations per year in *all* 524 control municipalities in Switzerland). There is no unaffected municipality with at least one second home transaction in every year. Hence, there are two major problems when separating first and second homes in control municipalities. First, the information on second home status does not appear to be correct nor reliable for control municipalities. Second, because there are only very few transactions in control municipalities, the aggregate on municipality level contains only a single observation for most of the few municipalities left. Because aggregates consist usually only on one observation, there is a huge volatility in prices and due to these erratic price changes unaffected municipalities do not serve as sensible control group anymore. Therefore, focusing only on second home transactions does not make sense and should be neglected.

In order to shed some more light on this matter, I run the estimation procedure only for first homes. In this case, only 48 affected municipalities are left with at least one first home transaction in every year. I repeat the estimation for first *and* second homes with the same 48 affected municipalities. If second homes and first homes are poor substitutes, we expect a negative effect

[19] Hilber and Schöni (2020) conduct a similar robustness check. This check does not change their results and indicates that including municipalities close to treated municipalities does not bias the estimated effects in the baseline model.



(a) Effect on first homes only

(b) Effect on first and second homes

Figure 4.7: Heterogeneity of the effect: Effect on first homes only vs. effect on first and second homes.
Notes: Only municipalities with at least one first home transaction in every year included (48); CI are based on 10,000 placebo runs and the outcome of 2007 is included in both estimations.

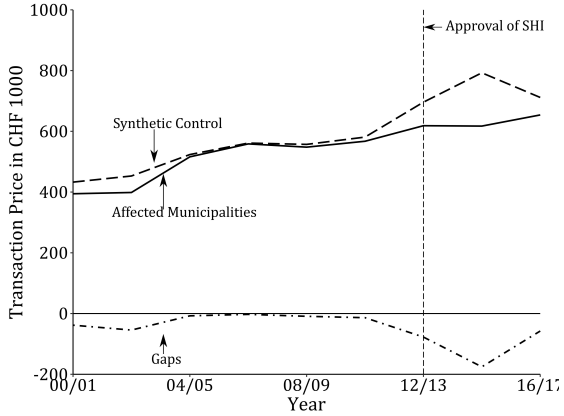
on first homes and a positive effect on second homes. 54 percent of all transactions in treatment municipalities are second homes. Hence, if first and second homes are poor substitutes and the effect on second homes is positive, the effect including only first homes must be substantially more negative than the effect including first and second homes.

The results of both estimations can be found in Figure 4.7. The effect is negative and of a similar magnitude in both approaches. A closer look to the data reveals that the effect including first and second homes is even slightly more negative than the approach including only first homes: While there is no significant effect in the first two years after the vote, the average effect in three to five years after the vote is with -15.6 percent for first and second homes lower than for first homes only with -15.2 percent. These results indicate that the effect on second homes is negative and similar to the effect on first homes. This finding supports the assumption that first and second homes are rather good substitutes. This result contradicts the results of Hilber and Schöni (2020), who find a positive effect for second homes. Hilber and Schöni (2020) estimate the effect separately for second homes.

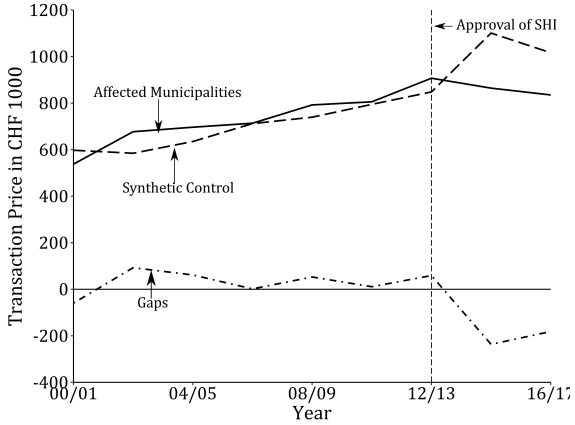
Heterogeneity of Effect: Touristic vs. Less Touristic Municipalities

In this Section, the heterogeneity of the effect is examined in more detail. The best known touristic municipalities with secondary housing rates above 50 percent are considered separately from the other less touristic municipalities. There are 56 affected municipalities with a share of second homes of more than 50 percent and 33 affected municipalities with a lower share. Figure 4.9a shows the effect of the SHI on housing prices in the best-known touristic municipalities. The effect is very similar to the effect in the benchmark estimations: No effect in the first two years after the intervention and then a strong negative effect of -15 percent to -20 percent three to five years after the vote. In the sixth (-5 percent) and seventh (-10 percent) year after the vote, the prices in affected municipalities are still below the synthetic control's prices, however, the effect is not significant anymore. Hence, the effect seems to be slightly stronger in these high-amenity places in years 2014 to 2016, but prices seem to recover more sustainable than in the benchmark estimations. However, differences to the benchmark estimation are very small and should not be over-interpreted.

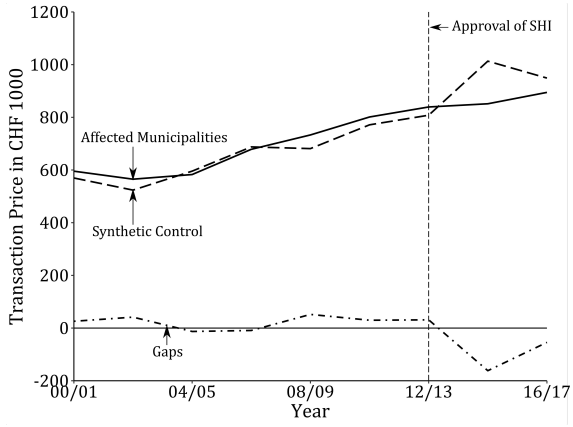
Figure 4.9b shows the effect on less touristic municipalities. The pre-intervention fit is less accurate for this approach (RMSE of CHF 40,000). Post-intervention RMSE is with CHF 103,000 still clearly higher. Post-intervention prices in treatment regions are between 13 percent and 20 percent lower in years 2014 to 2018 and significant on the 95 percent level or more. Hence, less touristic places seem not at all to recover from the price drop in 2014. However, because the pre-intervention fit is less accurate (often outside of the 90 percent CI bounds) this result should be interpreted with caution.



(a) Effect in the canton of Valais



(b) Effect in the canton of Graubünden

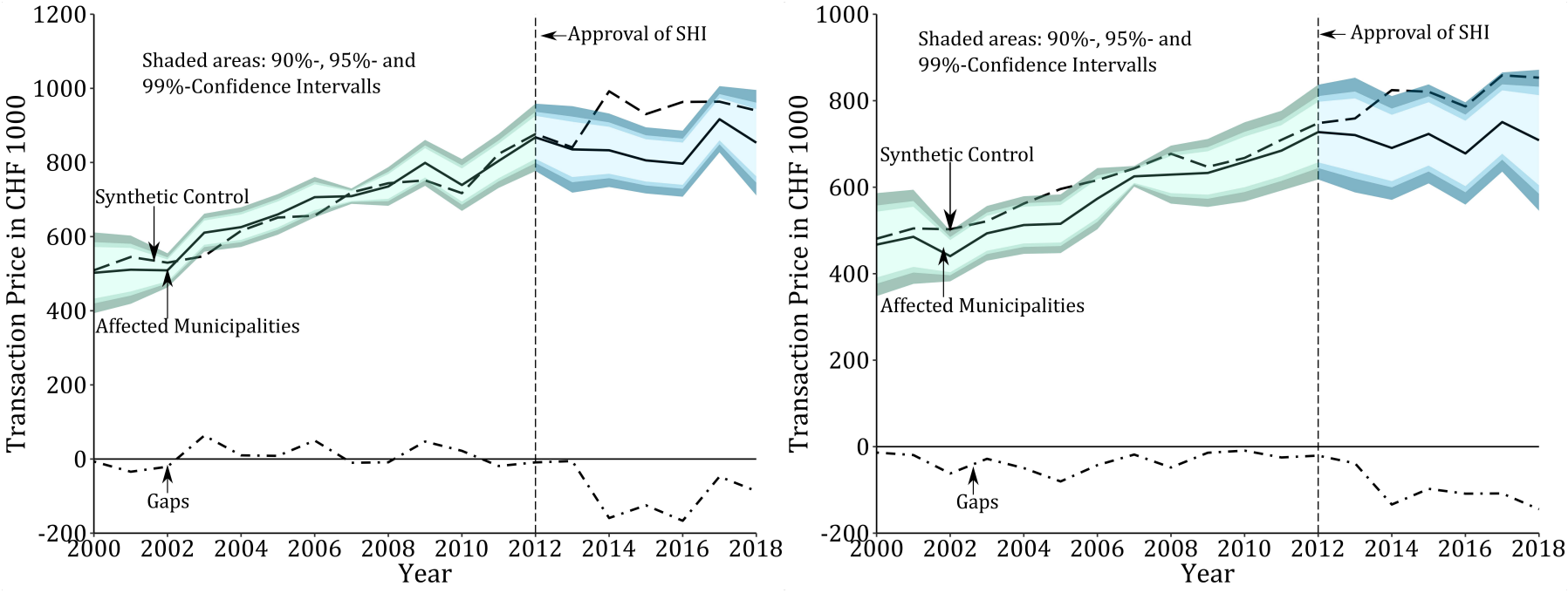


(c) Effect in the canton of Ticino

Figure 4.8: Case studies for the three most affected cantons Valais, Graubünden and Ticino.
Notes: Transactions of two years were pooled and the outcome of 2006/07 is included.

4.6.4 Case Studies of Most Affected Cantons

In this Section three case studies of three cantons are discussed, the cantons of Valais, Graubünden and Ticino. These three cantons are located in the Alps and are clearly the most affected



(a) Effect on most touristic places

(b) Effect on less touristic places

Figure 4.9: Heterogeneity of the effect: Effect on more touristic places (second home share above 50 percent) and less touristic places (second home share between 20 percent and 50 percent).
Notes: CI in both approaches are based on 10,000 placebo runs and the outcome of 2007 is included in both estimations.

regions by the SHI. More than 70 percent of all municipalities are affected in the cantons Valais and Graubünden and 60 percent in the canton of Ticino.[20] The absolute number of municipalities affected is as well clearly the highest in these cantons. Therefore, I estimate the effect of the SHI for these cantons separately. Because this approach is considering regions separately, the number of affected municipalities with at least one observation in each year is too small. Therefore, the transactions of two years are pooled in order to increase the number of municipalities per region.[21]

In opposite to the benchmark estimations or the other case studies, there is a negative effect in the first two years after the vote in the Canton of Valais. The price of affected municipalities is 12.5 percent below the counterfactual price in the years 2012/13. The strongest negative effect is found in 2014/15 with -28.4 percent compared to the control municipalities. Prices seem to recover in 2016/17, where the actual prices are only 8.8 percent below the synthetic control.

There is no effect visible in the first two years after the vote in Graubünden. However, similarly to the Valais, a tremendous negative effect of -27.4 percent can be found in 2014/15. Unlike the Valais, prices do not recover in 2016/17 with a negative effect of -21.7 percent.
Real estate prices in Ticino are as well negatively affected by the SHI. However, the magnitude of the effect is smaller than in Valais and Graubünden. Again, there is no effect in the first two years after vote, but a strong negative of -19 percent in 2014/15. Prices recover in 2016/17 with a negative effect of -6 percent.

In summary, the two most affected cantons, Valais and Graubünden, experience a tremendous negative price effect of about -28 percent in 2014/15. The magnitude of this effect is greater than the magnitude found in benchmark results including all regions. The Ticino is as well affected negatively, however, the effect in Ticino is comparable to the benchmark effect including all regions.

4.7 Evidence for Impact Channels

The results show that the SHI caused a drastic decrease in prices. This decrease in prices can either be explained by an extension in housing supply *or* a fall in demand in affected municipalities. In Section 4.3, different mechanisms through which the SHI might cause such an extension in supply or a fall in demand are discussed. While a boost in supply can only be explained by a one-time glut of last-minute construction permits, there are two different channels that might cause a fall in demand – adverse effects on local economies and a lock-in effect. In this section, I examine whether there is evidence for these three impact channels. This exercise represents an attempt to narrow down whether the different impact channels exist, but it cannot draw definitive conclusions.

[20] The group of affected municipalities in these cantons is consequently larger than the control group. This prevents to compute statistical significance and CI as discussed in Section 4.4.2.
[21] Due to this pooling, all municipalities with at least one observation in every second year is included.

4.7.1 Impact Channel: Increase in Housing Supply

As discussed in Section 4.3, it is possible that a one-time glut of newly build homes might have caused the drop in real estate prices in affected municipalities. This would as well explain the two year delay of the effect after the intervention, because it took time until the homes permitted right after the vote in 2012 were built and came on the market. According to this argument, there must be an increase in housing supply in 2014 and afterward in affected municipalities relative to the synthetic control municipalities. I test this argument by running the same synthetic control procedure as above with housing stock instead of prices as outcome variable. The result can be found in Figure A.2 in Appendix A.5. It is clearly visible that the housing stock development of the counterfactual municipalities is not significantly different from the development in the affected municipalities. Compared to the control housing stock, the SHI rather caused a decrease of the housing stock in affected municipalities. Hence, there is no evidence for the argument that a one-time increase in housing supply caused the drop in prices. Consequently, the SHI seems to affect housing prices via a drop in demand.

4.7.2 Impact Channel: Adverse Effects on Local Economies

We discussed indirect channels that might have caused a drop in housing demand in affected municipalities. One reason that can explain this drop are adverse effects on local economies. The SHI was a setback for the locally very vital construction industry and the SHI prevents the use of second homes or free construction land to cross-subsidize (see Section 4.3 for more details). Such adverse effects on local economies should be reflected in the unemployment rates of affected municipalities. Therefore, I test whether the SHI had an effect on local unemployment rates using the synthetic control method. In order to do that I construct an unemployment rate on municipality level by dividing the number of unemployed by the population of each municipality.[22] The data on number of unemployed by municipality is collected by the SECO and available from 2004 onwards. The SECO strongly recommends to drop municipalities, if the standard deviation of unemployed is greater than a quarter of the average number of unemployed.[23] Following this procedure, 58 percent of all municipalities in Switzerland need to be dropped. In our sample, only 24 affected municipalities and 445 control municipalities remain in the data.

So far I used predictors of real estate prices in order to construct the synthetic control. However, these real estate characteristics are poor predictors for unemployment rates. Further, many important predictors of unemployment are not available on municipality level. Therefore, I compute the synthetic control of the unemployment rate slightly different than the synthetic control of real estate related outcomes. Following Doudchenko and Imbens (2016) I run constrained regressions in order to compute the synthetic control. Constrained regressions are a special

[22] Since there is no data on the working population on municipality level, I divide the number of unemployed by the municipality's population. Population data on municipality level is not available from 2001 to 2006. Therefore, I impute population by municipality for the years 2004 to 2006 by linear interpolation.

[23] If the standard deviation is too big relative to the mean, there is too much noise in order to be able to interpret the change of unemployed.

case of the synthetic control method, where the predictor is the full vector of pre-intervention outcomes.[24]

Results are shown in Figure A.3 in Appendix A.5. While the pre-intervention fit is very good (RMSE=0.12), the unemployment rate is consistently higher in affected municipalities after the intervention: In the first four years after the intervention (2012 to 2015) the unemployment rate was 9 to 11 percent (0.3 percentage points) higher in affected municipalities compared to the synthetic control. However, these differences in the unemployment rate are not significant. In 2017 and 2018, the unemployment rate is significantly higher in affected municipalities (13 to 14 percent). The post-intervention RMSE is with 0.32 2.8 times greater than in the pre-intervention period, which results in an overall p-value of 0.11 based on the RMSE-ratio distribution of 10,000 placebo iterations. Although gaps in unemployment rates are just above of the conventional 0.1-threshold for statistical significance, the results still indicate that unemployment is likely to have increased as a result of the SHI and that local economic activity has cooled down. I interpret this as suggestive evidence that the decrease in real estate prices was partially driven by adverse effects on local economies. As a robustness test, I run the same routine for the full sample (not shown), i.e. I do not exclude municipalities with a standard deviation greater or equal to a quarter of the average of the unemployment rate. The pattern remains similar to the result based on the restricted sample: Unemployment rate is between 7 and 13 percent higher in affected municipalities in the first four years after the intervention. However, these differences are only significant in year 2012. Unemployment rates in 2016 and later are very similar for both groups.

4.7.3 Impact Channel: Legal Uncertainty and the Lock-In Effect

Another impact channel through which the SHI might cause a fall in prices is the legal uncertainty and a general lock-in effect. As discussed in Section 4.3, the final law was agreed upon in 2015 and implemented in 2016. While it is difficult to quantify this legal uncertainty directly, it might have caused a lock-in effect, in which the quality and state of properties sold decreased. Further, the SHI might as well have caused a general lock-in effect irrespective of the legal uncertainty with similar consequences (see Section 4.3). I examine the effect of the SHI on the quality and condition of the houses sold to show whether there is evidence for the existence of the lock-in impact channel. If the quality and condition of the houses decreases due to the SHI, this is an indication that the channel exists.[25]

In this section, I run the SCM as in the benchmark estimations using quality and state of homes as outcome variables. Figure A.4a shows the results for the quality of homes. Although the pre-intervention fit seems to be good, CI bands are very narrow and the quality of property

[24] In this special case of the synthetic control method, the diagonal matrix $\mathbf{V}$ is a N x N identity matrix and consequently $\underset{W \in \omega}{min}(\mathbf{X}_1 - \mathbf{X}_0\mathbf{W})'(\mathbf{X}_1 - \mathbf{X}_0\mathbf{W})$ must be solved, where $\mathbf{X}_1$ and $\mathbf{X}_0$ contain only the pre-intervention outcome of the treatment municipality in $\mathbf{X}_1$ and of the control municipalities in $\mathbf{X}_0$. See Section 4.4 for more information.

[25] Note that the SCM minimizes the pre-intervention distance metric of real estate characteristics. Hence, these characteristics used in the SCM are pre-determined and do not respond to the SHI. This is important since covariates that respond to the treatment are likely to be *bad controls*.

sold is significantly lower in affected municipalities as of 2006. Therefore, the synthetic control might not be a reliable counterfactual in this case. The quality of houses remains lower in affected municipalities than in control municipalities in the post-intervention period and the gaps are larger in the post-intervention period compared to the 2006–2011 period. Further, the post-intervention RMSE is with 0.22 about 2.4 larger than in the pre-intervention period with 0.09 resulting in a p-value of 0.03 based on 10,000 placebo iterations. Although these findings suggest that there might be a decrease in quality of homes, these results should be interpreted wit caution since the synthetic control is not able to replicate affected municipalities in the 2006–2011 period.

Results for the state of real estate sold are presented in Figure A.4b. The synthetic control fits the original housing state well in the pre-intervention with the exception of the years 2006 and 2011. The state of real estate sold in the post-intervention period is significantly lower in affected municipalities compared to control municipalities – especially in the years 2013 to 2015. The gaps in 2013 are for example two to three times larger than in 2011. In 2016, the difference is not significant anymore, however, the fact that the gap widens again in 2017 and 2018 indicates that the lock-in effect might prevail as well for later years. A similar pattern is observed in Figure A.4a where the quality of homes sold in affected municipalities remains significantly lower in affected municipalities even after the final law was implemented. This may be because second home owners are aware that it will be almost impossible to re-buy a second home in affected municipalities in the future due to the supply freeze and are therefore more reluctant to sell their second homes. Hence, only those in need tend so sell their home and therefore, quality might be lower (see Section 4.3).

Summing up, there is no evidence for a one-time boost in supply that might explain the drop in prices in the aftermath of the SHI. Thus, the falling prices must take place through a fall in demand. The evidence for the adverse effect on local economies and the lock-in effect are not clear cut and must be interpreted with caution. However, results point toward the existence of both channels. Hence, it is likely that the decrease in prices caused by the SHI realized through a cooling down of local economic activity and a lock-in effect.

4.8 Conclusion

I estimate the causal effect of a drastic second home construction ban in Switzerland by exploiting a quasi-experimental setup. First, I show that simple DD estimations are not suitable for the context and, therefore, apply the synthetic control method in an innovative way. Results show that there was no short-run effect on real estate prices in the first two years after the intervention (2012 to 2013). However, I found a sharp decrease in prices of -19 percent in the third year, -10 percent in the fourth year and -18 percent in the fifth year after the vote. These results are robust as several checks show. While it is not possible to conclusively pin down the impact channel, data points toward an adverse effect on local economies and a lock-in effect that caused the drop in prices.

The goals of the regulation were to stop splinter development and keep housing affordable for the local population. Estimations show that the SHI had no significant effect on housing supply. However, housing supply was (insignificantly) lower in affected regions compared to the synthetic control, which indicates that the intervention might have contributed to a decrease of urban sprawl. Moreover, estimates clearly show that housing prices decreased in affected municipalities. However, this came at a price: An economic drawback in affected regions as well as a lock-in effect are the probable reasons for the drop in prices. This adverse effect on local economies is likely to have increased local unemployment. Furthermore, local land owners suffer a strong devaluation of their land, because it cannot be used for the construction of second homes anymore. Hence, housing might have become more affordable for locals, however, the devaluation of the land price and economic drawback on these local economies are unintended consequences which harm the local population. Thus, in the light of the global tendency to regulate the second home market this paper shows that regulations should be made more flexible to take account of regional peculiarities and contexts. Drastic regulations may often exhibit unwanted effects as it is the case with the SHI.

A. Appendices

A.1 Appendix: Price Trends in Affected and Unaffected Municipalities

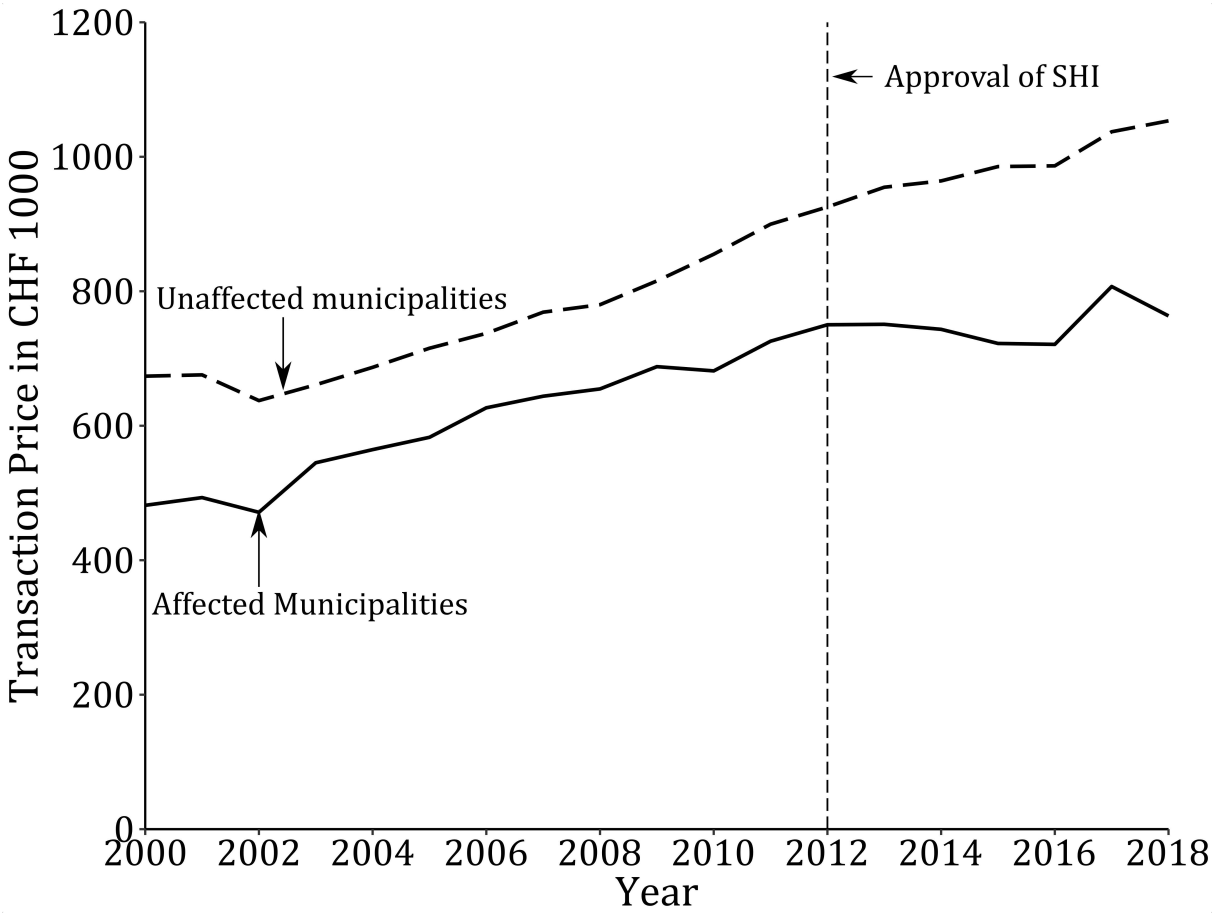


Figure A.1: Trends in transaction prices: Transaction prices in affected vs. transaction prices in unaffected municipalities

A.2 Appendix: Placebo Difference-in-Differences

In Tables A.1 and A.2, the results of placebo interventions are presented. Table A.1 is based on transaction-level data, i.e. every single real estate transaction is an observation. The results of Table A.1 are estimated with the following regression equation:

$$Y_{imcgt} = \alpha + \gamma Treat_g + \lambda_t + \delta(Treat_g * Post_t) + X'_{imcgt}\beta + \mu_c + \epsilon_{imcgt}, \tag{A.1}$$

where Y is the transaction price of the property sold in year t and transaction i, taking place in municipality m and canton c. The municipality either belongs to the control ($g = 0$) or the treatment group ($g = 1$). $Treat_g$ is a dummy taking the value 1 if transaction takes place in an affected municipality and zero otherwise, λ_t is a year dummy, $Post_t$ is a dummy which switches on for post-treatment years, X_{imcgt} is a set of time-variant control variables (see Table A.1) and μ_c are canton fixed effects. The coefficient of interest is δ indicating the effect of the (placebo) treatment. In all estimations in Tables A.1 and A.2 only years 2000 to 2011 are considered, i.e. all years before the actual SHI took place. E.g. for the placebo intervention in 2006 (i.e. column 1 in Table A.1), $Post_t$ switches on for years 2006 to 2011. Because there was no actual intervention until 2012, there should be no effect caused by this placebo intervention. However, Table A.1 reports a significant effect for every placebo intervention.

In Table A.2, data is aggregated by year and municipality. Thus, instead of observations on transaction-level the observations are on municipality level. As in the SCM applied in the main analysis of the paper, only municipalities with at least one transaction in every year is considered. Hence, the results in Table A.2 are based on the same data as in the SCM approach. The estimation equation remains really similar:

$$Y_{mcgt} = \alpha + \gamma Treat_g + \lambda_t + \delta(Treat_g * Post_t) + X'_{mcgt}\beta + \mu_c + \epsilon_{mcgt}, \tag{A.2}$$

The only changes are that the price level Y_{mcgt} and the time-variant controls X_{mcgt} are aggregated on municipality level. Again, effects of placebo interventions in Table A.2 are significant.

Table A.1: Difference-in-differences placebo tests: transaction-level data

	Dependent variable: Log of price				
	Placebo 06[a]	Placebo 07[a]	Placebo 08[a]	Placebo 09[a]	Placebo 10[a]
	(1)	(2)	(3)	(4)	(5)
Treatment	−0.161***	−0.154***	−0.151***	−0.150***	−0.148***
	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)
Treatment x Post	0.032***	0.022***	0.019***	0.016**	0.014*
	(0.006)	(0.006)	(0.007)	(0.007)	(0.008)
Constant	11.098***	11.097***	11.096***	11.096***	11.096***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Year and Canton fixed effects	Yes	Yes	Yes	Yes	Yes
Time-variant contr.[b]	Yes	Yes	Yes	Yes	Yes
Observations	173,280	173,280	173,280	173,280	173,280
Adjusted R^2	0.64	0.64	0.64	0.64	0.64

Notes: Only pre-intervention years 2000 to 2011 are included.
a Each model indicates the year of the placebo intervention. I.e. the dummy *Post* takes value 1 after the placebo intervention and zero otherwise.
b Yearly transactions, number of rooms, plumbing units and garages, quality, state and micro-location of the property as well as second home rate.
*p<0.1; **p<0.05; ***p<0.01; standard errors clustered at municipality-year level in parentheses.

Table A.2: Difference-in-differences placebo tests: municipality-level data

	Dependent variable: Log of price				
	Placebo 06[a]	Placebo 07[a]	Placebo 08[a]	Placebo 09[a]	Placebo 10[a]
	(1)	(2)	(3)	(4)	(5)
Treatment	−0.188***	−0.177***	−0.170***	−0.160***	−0.153***
	(0.022)	(0.021)	(0.021)	(0.021)	(0.020)
Treatment x Post	0.096***	0.087***	0.090***	0.080***	0.076***
	(0.019)	(0.020)	(0.021)	(0.022)	(0.026)
Constant	10.807***	10.806***	10.800***	10.796***	10.793***
	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)
Year and Canton fixed effects	Yes	Yes	Yes	Yes	Yes
Time-variant contr.[b]	Yes	Yes	Yes	Yes	Yes
Observations	7,356	7,356	7,356	7,356	7,356
Adjusted R^2	0.70	0.70	0.70	0.70	0.70

Notes: Only pre-intervention years 2000 to 2011 are included. Transactions are aggregated by year and municipality and only municipalities with at least one transaction in every year are considered.
a Each model indicates the year of the placebo intervention. I.e. the dummy *Post* takes value 1 after the placebo intervention and zero otherwise.
b Yearly transactions, number of rooms, plumbing units and garages, quality, state and micro-location of the property as well as second home rate.
*p<0.1; **p<0.05; ***p<0.01; standard errors clustered at canton-year level in parentheses.

A.3 Appendix: Group-Specific Time Trends

Angrist and Pischke (2008) propose to allow for group-specific time trends in order to check the validity of the common trend assumption. It is discouraging, if these group-specific time trends are significantly different from each other or if the estimated effects of interests are changed by the inclusion of group-specific time trends. Therefore, a group-specific time trend is included in the regression Equations (A.1) and (A.2), while all other variables remain unchanged:

$$Y_{imcgt} = \alpha + \gamma Treat_g + \lambda_t + \rho(Treat_g * t * Pre_t) + \delta(Treat_g * Post_t) + X'_{imcgt}\beta + \mu_c + \epsilon_{imcgt}. \quad \text{(A.3)}$$

Note that t is the time trend and Pre_t is an indicator variable for the pre-intervention period, i.e. Pre_t takes the value 1 for per-intervention years and the value 0 for post-intervention years. The idea of the interaction $Treat_g * t * Pre_t$ is to include only pre-intervention group-specific time trends,[1] which should be the same according to the parallel trends assumption. Again, the estimations are conducted with transaction-level (Table A.3) and municipality-level data (Table A.4). Data of all years (i.e. 2000 to 2018) is included in these estimations. Results show that the group-specific trends are significant and that allowing for group-specific time trends changes the estimated effect of the SHI tremendously for the estimation on transaction-level as well as on municipality-level.[2]

[1] Note that the time trend t starts from 1 for the year 2000 to 19 for year 2018.

[2] Note, that time-variant covariates included in the regression might be *bad controls*, if they respond as well to the SHI. Since this is likely, transaction-level regressions are run with and without these time-variant covariates, while municipality-level regressions are run with time-variant covariates or pre-determined covariates. Pre-determined covariates are the pre-intervention averages of time-variant covariates.

Table A.3: Difference-in-differences with group-specific time trends: transaction-level data

	Dependent variable: Log of price			
	(1)	(2)	(3)	(4)
Treatment	−0.140***	−0.178***	−0.245***	−0.290***
	(0.005)	(0.008)	(0.008)	(0.012)
Treatment x Year x Pre		0.006***		0.007***
		(0.001)		(0.001)
Treatment x Post	−0.019***	0.019**	−0.038***	0.007
	(0.006)	(0.008)	(0.009)	(0.012)
Constant	11.126***	11.130***	13.385***	13.388***
	(0.006)	(0.006)	(0.006)	(0.006)
Year and Canton fixed effects	Yes	Yes	Yes	Yes
Time-variant contr.[c]	Yes	Yes	No	No
Observations	243,824	243,824	243,824	243,824
Adjusted R^2	0.66	0.66	0.21	0.21

Notes: Data of all years (2000 to 2018) is included.
a Yearly transactions, number of rooms, plumbing units and garages, quality, state and micro-location of the property as well as second home rate.
* p<0.10, ** p<0.05, *** p<0.01; standard errors clustered at municipality-year level in parentheses.

Table A.4: Difference-in-differences with group-specific time trends: municipality-level data

	Dependent variable: Log of price			
	(1)	(2)	(3)	(4)
Treatment	−0.163***	−0.272***	−0.059***	−0.148***
	(0.018)	(0.025)	(0.020)	(0.027)
Treatment x Year x Pre		0.017***		0.014***
		(0.003)		(0.003)
Treatment x Post	0.034**	0.145***	−0.003	0.086***
	(0.016)	(0.024)	(0.017)	(0.025)
Constant	10.910***	10.931***	9.876***	9.886***
	(0.035)	(0.035)	(0.076)	(0.076)
Year and Canton fixed effects	Yes	Yes	Yes	Yes
Time-variant contr.[a]	Yes	Yes	No	No
Pre-determ. contr.[b]	No	No	Yes	Yes
Observations	11,647	11,647	11,647	11,647
Adjusted R^2	0.72	0.72	0.68	0.68

Notes: Data of all years (2000 to 2018) is included.
a Yearly transactions, number of rooms, plumbing units and garages, quality, state and micro-location of the property as well as second home rate.
b Pre-determined controls are pre-intervention averages of the time-variant controls.
* p<0.10, ** p<0.05, *** p<0.01; standard errors clustered at canton-year level in parentheses.

A.4 Appendix: Granger Causality Testing

A closely related possibility to check the identification assumption of the DD approach is to test for causality in the spirit of Granger (Angrist and Pischke, 2008). Granger causality tests check whether a past intervention predicts an outcome, while a future intervention does not. I.e. lags and leads of the intervention are included in order to estimate post-treatment effects resp. anticipatory effects. Effects of future policy changes should not matter for the contemporaneous outcome. Leads and lags are included in the regression equation in order to check for these anticipatory and post-treatment effects (Angrist and Pischke, 2008):

$$Y_{imcgt} = \alpha + \gamma Treat_g + \lambda_t + \sum_{\tau=0}^{m} \delta_{-\tau}(Treat_g * Post_{t-\tau}) + \sum_{\tau=1}^{q} \delta_{+\tau}(Treat_g * Post_{t+\tau}) + X'_{imcgt}\beta + \mu_c + \epsilon_{imcgt}. \quad \text{(A.4)}$$

The sums allow for m lags and q leads and again, the equation is estimated with transaction-level and municipality level data.[3] The estimates in Table A.5 show significant effects in several of the five years before the SHI vote took place. This pattern does not appear to be consistent with a causal interpretation of the SHI effect on prices.

[3]Note, that time-variant covariates included in the regression might be *bad controls*, if they respond as well to the SHI. Since this is likely, transaction-level regressions are run with and without these time-variant covariates, while municipality-level regressions are run with time-variant covariates or pre-determined covariates. Pre-determined covariates are the pre-intervention averages of time-variant covariates.

Table A.5: Difference-in-differences with Granger causality testing

	Dependent variable: Log of price			
	Transaction-level[a]		Municipality-level[b]	
	(1)	(2)	(3)	(4)
Treatment	−0.116***	−0.150***	−0.096***	−0.204***
	(0.006)	(0.006)	(0.021)	(0.019)
Treat. x Dummy 2007	0.056***	0.028**	0.081**	0.062*
	(0.017)	(0.012)	(0.036)	(0.037)
Treat. x Dummy 2008	0.031	0.028**	0.107***	0.106**
	(0.019)	(0.013)	(0.038)	(0.042)
Treat. x Dummy 2009	0.032*	0.026**	0.086**	0.101***
	(0.018)	(0.011)	(0.036)	(0.033)
Treat. x Dummy 2010	0.006	0.019*	0.083**	0.116***
	(0.018)	(0.011)	(0.037)	(0.038)
Treat. x Dummy 2011	0.027	0.031***	0.087**	0.109***
	(0.018)	(0.012)	(0.035)	(0.035)
Treat. x Dummy 2012	0.026	0.017	0.098**	0.113***
	(0.018)	(0.012)	(0.041)	(0.036)
Treat. x Dummy 2013	0.002	0.016	0.110***	0.114***
	(0.019)	(0.012)	(0.033)	(0.033)
Treat. x Dummy 2014	−0.010	0.019	0.063	0.121***
	(0.019)	(0.012)	(0.040)	(0.037)
Treat. x Dummy 2015	−0.070***	−0.026*	0.003	0.092**
	(0.021)	(0.014)	(0.039)	(0.038)
Treat. x Dummy 2016	−0.069***	−0.015	−0.009	0.082**
	(0.021)	(0.015)	(0.040)	(0.037)
Treat. x Dummy 2017	−0.042**	−0.028**	0.023	0.041
	(0.021)	(0.014)	(0.038)	(0.037)
Treat. x Dummy 2018	−0.080***	−0.078***	−0.051	−0.029
	(0.022)	(0.015)	(0.040)	(0.040)
Year and Canton fixed effects	Yes	Yes	Yes	Yes
Time-variant contr.[c]	No	Yes	No	Yes
Pre-determ. contr.[d]	No	No	Yes	No
Observations	243,824	243,824	11,647	11,647
Adjusted R^2	0.21	0.66	0.68	0.72

Notes: These estimations include leads and lags of the intervention. Data of all years (2000 to 2018) is included.
a Observations are on transaction level.
b Observations are on municipality level. I.e. transactions are aggregated by year and municipality
c Yearly transactions, number of rooms, plumbing units and garages, quality, state and micro-location of the property as well as second home rate.
d Pre-determined controls are pre-intervention averages of the time-variant controls.
*p<0.1; **p<0.05; ***p<0.01; standard errors clustered at municipality-year resp. canton-year level in parentheses.

A.5 Appendix: Impact Channels

A.5.1 Impact Channel: Increase in Housing Supply

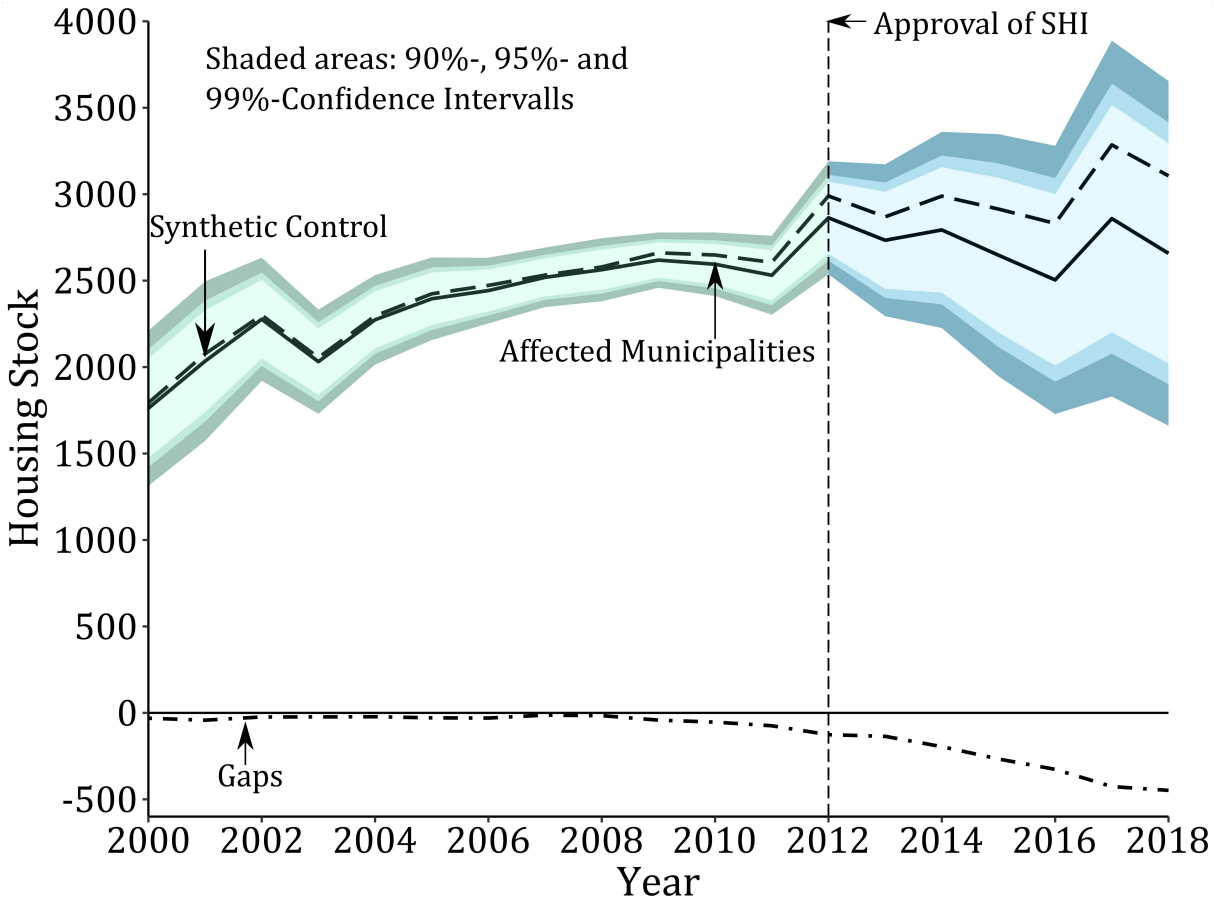


Figure A.2: Effect on housing supply (in dwelling units).
Notes: CI are based on 10,000 placebo runs and the outcome of 2007 is included.

A.5.2 Impact Channel: Adverse Effects on Local Economies

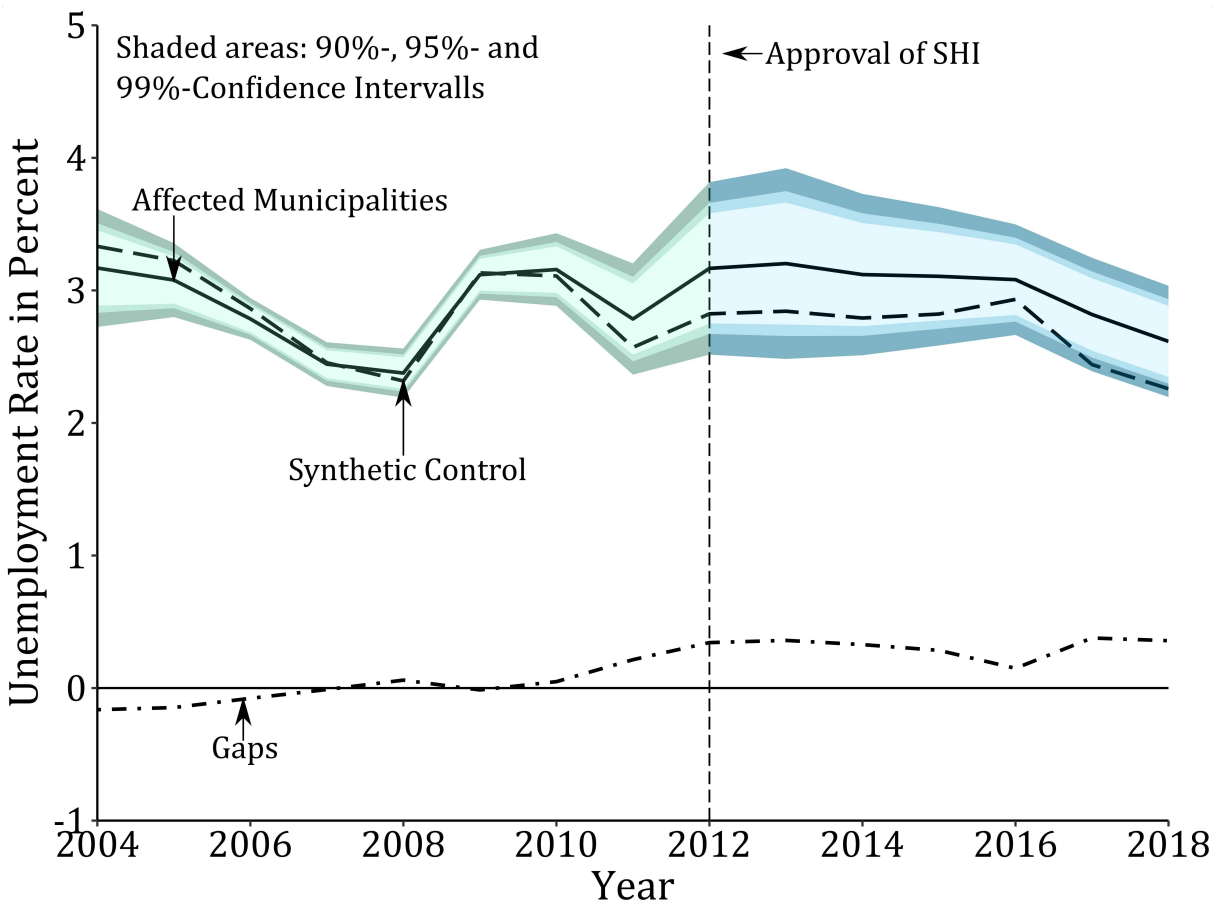
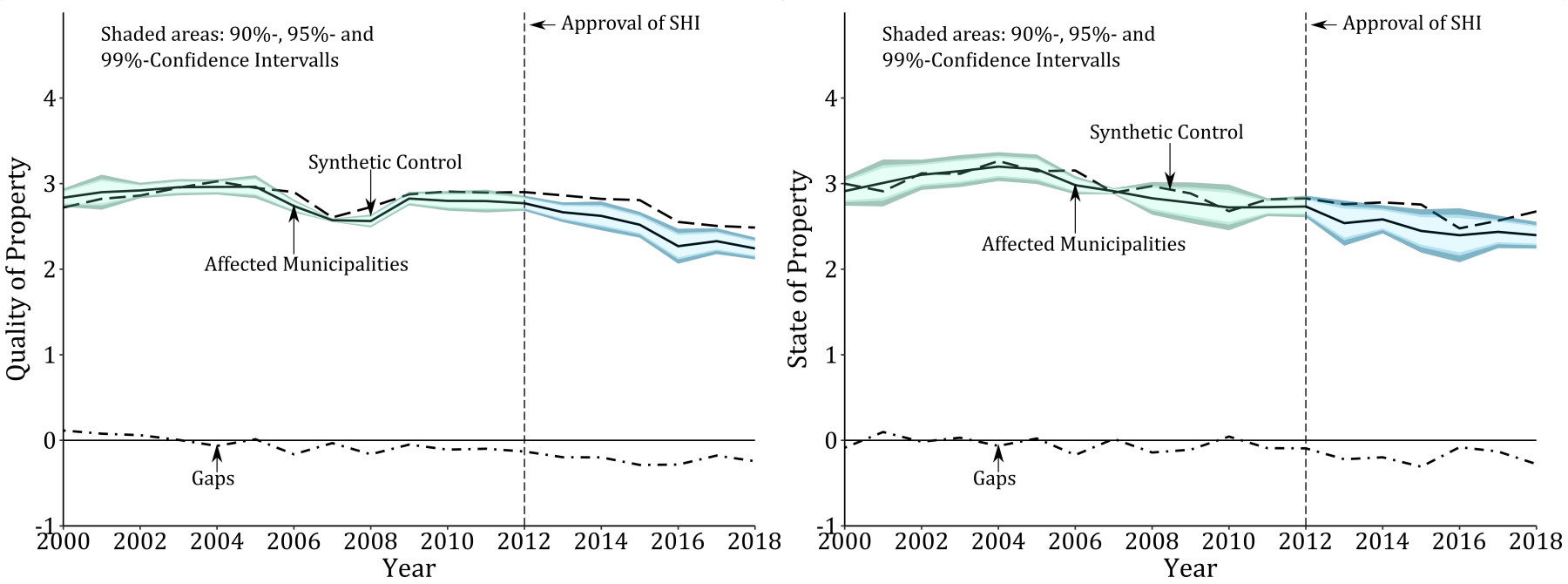


Figure A.3: Effect on unemployment.
Notes: CI are based on 10,000 placebo runs and only the pre-intervention outcome variable is used as predictor (constrained regressions method). Only municipalities with a standard deviation of unemployed smaller than a quarter of the average of unemployed (24 treated municipalities and 445 control municipalities) included.

A.5.3 Impact Channel: Legal Uncertainty and Lock-In Effect



(a) Effect on quality of properties.

(b) Effect on state of properties.

Figure A.4: Effect on property characteristics.
Notes: CI are based on 10,000 placebo runs and outcome of 2007 is included. Quality resp. state of the property are used as outcome variables, while predictors are the same as in benchmark estimations including transaction prices.

5 The Effect of Outward Foreign Direct Investments on Home Employment: Evidence Using Swiss Firm-Level Data*

5.1 Introduction

At present, globalization and international economic interdependence are experiencing a political setback. In many countries that were previously known for their economic openness protectionist forces have gained increasing influence. The most striking example of this turning away from globalization is the election of Donald Trump in the USA. The reason for this departure from economic openness is, among others, the fear that domestic jobs will be relocated.

Accordingly, the discussion about the effects of outward FDI on the domestic job market is reviving and the benefits of investment treaties are doubted. On the one hand, proponents of protectionism often regard outward FDI as a classic zero-sum game: The total number of jobs is fixed and every job that is built up abroad, is a job that is lost at home. This static idea of firms comes closest to the so-called *displacement effect* known in economic theory. Policy makers that support this idea strongly oppose large scale outward FDI in order to ensure home employment and production. However, this static idea of the economy does not correspond to what we observe in reality: FDI are supposed to be crucial to give firms the possibility to remain competitive and guarantee or even create additional domestic jobs in a longer term. Part of this argumentation is related to the *output effect* discussed in economic theory.

Hence, economic theory suggests two important channels through which outward FDI affect home employment: a negative and direct displacement effect channel and a positive and indirect productivity or output effect channel. Opponents of outward FDI focus rather on the displacement effect, while supporters emphasize the output effect and the importance of FDI for firms in order to remain competitive and to survive. Economic theory is not able to predict, whether the gain of domestic jobs due to output effects outweighs the initial loss due to displacement. Therefore, the currently much discussed political question about the effect of outward FDI on home employment boils down to an empirical issue.

The goal of this paper is to empirically examine this effect of outward FDI[1] on *within firm*

[1] Note that we are interested in operational activities of MNEs and not necessarily in financial flows. Therefore,

*This chapter is joint work with Preetha Kalambaden.

domestic employment in the context of a small and open economy with a high relative outward FDI stock, i.e. Switzerland. In particular we aim to answer the question whether firms that engage in outward FDI increase or decrease home employment due to foreign activities. This means that we will not consider horizontal and vertical spillover effects on other firms. There is only limited evidence of these spillover effects. However, Tang and Altshuler (2015) find positive spillover effects of outward FDI on domestic suppliers, showing that at least backward linkages appear to affect home employment positively. Furthermore, we do not take into consideration that firms might have to close down their business in the longer run, if they do not have the possibility to conduct outward FDI. Hence, the overall effect on home employment is likely to be more positive than our within firm estimates suggest.

We can draw on a broad literature, which addresses the same empirical question. The literature can be roughly categorized into three approaches: *i)* papers that pursue an instrumental variable (IV) strategy, *ii)* papers that use matching estimators to form a counterfactual, and *iii)* a paper that exploits a natural experiment. The first strand of literature using an IV strategy is most closely related to our method. In particular, Wright (2013) is quite closely related to this paper. We adopt the empirical strategy of this paper by estimating the displacement and the output effect in two separate steps. Wright (2013) uses sector-level data on US manufacturers and finds a positive overall effect of FDI on total employment. However, he is as well examining the effect on high- and low-skilled labor separately and finds that the total effect on low-skilled labor employment is negative. Among others, the approaches of Desai, Foley and Hines (2009) and Harrison and McMillan (2011) are as well related to our strategy. Desai, Foley and Hines (2009) show that greater foreign employment of US manufacturers is associated with greater domestic employment using firm-level data and applying as well an instrumental variable approach. Harrison and McMillan (2011) find mixed effects. They emphasize that it depends on the type and destination of FDI, whether the overall effect is positive or negative: For firms most likely to perform similar tasks in domestic and foreign affiliate, foreign and domestic employees are substitutes. However, for firms that engage in significantly different tasks at home and abroad, foreign and domestic employments are complements.[2]

The second strand of literature tries to establish a counterfactual for firms that invest abroad for the first time. This strand of literature compares national firms with firms that switch from national to multinational status using matching estimators. Debaere, Lee and Lee (2010) for example, find that South Korean multinational enterprises (MNE), which invest to countries with a lower income than South Korea face lower employment growth than comparable national firms. On the other hand, the authors find no significant difference in employment growth between

we use foreign employment of MNEs instead of actual financial flows or stocks as measure for foreign activities. It is important to know that FDI as we use it in this paper are operational activities (foreign employment) and not financial flows.

[2]This list is obviously incomplete and there are other papers that are related to our approach. Many of them focus on the effect on wage or skill intensity. For instance, Hummels et al. (2014) are estimating the effect of offshoring on wages in Denmark using an instrumental variable approach. Another prominent example are Ottaviano, Peri and Wright (2013) who apply as well an instrumental variable approach and find no negative effect of offshoring on employment level for the US. For a broader overview of offshoring and its effects on wage, skill intensity, investment as well as job loss and creation see Hummels, Munch and Xiang (2018).

firms investing in countries with higher income and comparable national firms. Barba Navaretti, Castellani and Disdier (2010) find, however, positive effects on home employment for Italian and French MNE, irrespective whether they invest in low- or high-income countries.

In a third approach, Sethupathy (2013) uses firm-level data and two events in Mexico as a natural experiment to identify the effects of a fall in the marginal cost of offshoring to Mexico. He finds no evidence of greater domestic job loss in the US due to offshoring. Finally, Crinò (2009) provides an excellent overview of the empirical literature of labor market effects of outward FDI. He finds that FDI mostly have a weak effect on home employment and concludes that results tend to be very mixed depending on countries and offshoring strategy.

Hence, empirical evidence does not clearly show whether overall effects of outward FDI on home employment are positive or negative. The outcomes vary by context and might even indicate contrary effects dependent on the destination market and types of FDI. Or as Lipsey (2004, p.340) puts it in his review of home-effects of FDI: "The effect may depend on whether the foreign operations' relation to home operations is "horizontal" or "vertical," [...] the extent to which the foreign operations are in goods production or in service activities, are in developed or developing countries, or are in industries with plant-level or firm-level economies of scale." The existing literature focuses so far on big economies, while the effect on small economies is clearly less explored.

This paper contributes to the existing literature by analyzing the domestic employment effect of outward FDI for a small economy, Switzerland. Switzerland has one of the highest outward FDI stocks and flows relative to GDP in the world and is thus heavily exposed to the effects of outward FDI. Not surprisingly, some of the largest MNEs such as Nestlé, Roche or Novartis are located in Switzerland. Hence, Switzerland is not only a small economy, but, relative to its size, it engages more strongly in FDI than all other countries examined in the literature so far. Moreover, we construct a novel instrument for foreign employment and our data contains firms from the service and manufacturing sector, while the other studies often focus on the manufacturing sector.[3]

We were able to acquire unique administrative firm-level data from the surveys on cross-border capital linkages from the Swiss National Bank (SNB). We construct a novel instrument for the number of employees abroad by using different exogenous predictors of FDI to estimate the potential employment for each firm in each country in a zero-stage. The idea to estimate a firm's location choices using exogenous predictors is related to the idea of di Giovanni and Levchenko (2009), who compute potential trade flows per sector adapting the classic gravity model approach. Using this novel instrument we estimate the displacement and the output effect in two different steps.

We find no evidence for the existence of a negative displacement channel and clear evidence for a positive output channel, when we are considering only FDI to high-income countries. However,

[3]For instance, Hijzen, Jean and Mayer (2011) focus as well on service and manufacturing sector of France and find a positive effect for horizontal FDI and no effect for vertical FDI. Crinò (2010) considers only the service sector in the US and finds positive employment effects for skilled workers, while less skilled workers might be displaced.

the positive effect of FDI to high-income countries on home employment is rather small. In the case of FDI to lower middle-income countries, the negative displacement effect outweighs the positive output effect and, thus, the cumulative effect on home employment is negative but as well rather moderate. This negative effect is driven by China and disappears as soon as China is excluded from the estimation. Further, we find no evidence for the displacement effect and only partially significant evidence for the output effect of FDI to low-income countries. Again, this potentially positive effect is driven by a single destination country, India. Finally, we are not able to estimate the effect of FDI to upper middle-income countries reliably. Considering all types of FDI, we find that outward FDI have no clear effect on domestic employment for Switzerland, if at all, outward FDI tend to create more domestic jobs within the firm than it relocates.

5.2 Conceptual Framework

Economic theory distinguishes two types of FDI, vertical and horizontal FDI (see e.g. Markusen and Maskus, 2003), which affect home employment through different channels. These types are based on different motivations for an MNE to open affiliates abroad. While it was debated in the early literature which type of FDI is predominant in the world, there is now a consensus that both types of FDI coexist and are important for MNEs (Davies, 2008). Because these types of FDI affect home employment potentially differently (Lipsey, 2004), it is important to understand these diverse types and the mechanisms behind them.[4]

5.2.1 Vertical Foreign Direct Investments

Helpman (1984) and Helpman and Krugman (1985) describe vertical FDI (VFDI) in early models. The motivation behind VFDI is to exploit differences in factor prices between countries. This means that a company produces intermediate goods abroad at lower costs and thus geographically relocates part of the production chain. As a result, there are intra-firm imports of low-wage goods, which were formerly produced domestically. Therefore, in this first relocation step, foreign and home production are substitutes and VFDI are supposed to reduce the number of domestic jobs. However, after this immediate potential displacement of domestic labor, the intermediate goods produced in these foreign affiliates are complementary to the production that remained in the home country. Because factor prices are lower abroad, intermediate goods produced in the foreign affiliates are cheaper. Due to these cost savings the MNE will be able to gain market shares and, therefore, to increase production and employment at home *and* abroad. Furthermore, an expansion in production abroad leads as well to higher demand for headquarter services. Accordingly, VFDI have a negative and immediate displacement effect, as well as an opposing positive effect via competitiveness and output on domestic employment. Because differences in factor prices are the crucial motivation for VFDI, these kind of investments flow typically from high-wage countries to low-wage countries.

[4]See Barba Navaretti, Venables and Barry (2006) for a very broad overview of MNE activities and different types of FDI as well as their effects.

5.2.2 Horizontal Foreign Direct Investments

In the case of horizontal FDI (HFDI), not only one stage but the entire production process is replicated in an affiliate abroad. I.e. the same products are manufactured in different locations. The motivation behind HFDI is the reduction of transport costs, market seeking, technology sourcing and exploitation of firm scale economies (see e.g. Markusen, 1984, for an early version of HFDI models or Markusen and Venables, 2000). Since in the case of HFDI the same products are manufactured at home and abroad, home and foreign production are substitutes. Outward HFDI thus tend to reduce domestic exports, which reduces the demand for domestic employment (Helpman, Melitz and Yeaple, 2004; Lipsey, 2004). However, HFDI allow more efficient sales in the foreign market leading to a stronger penetration of that market and accordingly, an increase in production abroad. This more complex organization and a further expansion of production due to gains in market shares lead to a higher demand for headquarter services and other complementary products, which are typically provided by the parent company (Helpman and Krugman, 1985). This gain in market share and the following rise in output increases the demand for domestic jobs. Furthermore, technology sourcing might increase home productivity, which leads as well to higher output and employment. So, there is again a negative displacement effect and a positive effect via output. The effect of HFDI on home employment strongly depends on whether a firm has exported much to a country before the foreign investment takes place. If there were only few or no exports in the forefront of the investment, there is little or no home production to be substituted and the displacement effect is negligible or even nonexistent. The more the firm exported to that country before it opened an affiliate there, the stronger is the displacement effect and it depends on the size of the output effect whether number of jobs increase or decrease at home. Because HFDI are motivated by technology sourcing or market seeking, this kind of FDI typically happens between high-income countries. Thus, according to classic theories both VFDI and HFDI have ambiguous effects on home employment: negative displacement effects as well as positive effects due to an expansion in production.

5.2.3 Trading Tasks

More recently Grossman and Rossi-Hansberg (2008) presented an alternative approach which examines the wage effects of offshoring. Instead of goods, tasks are traded in this new model. Wright (2013) reformulates this model in order to be able to estimate the effect on labor demand instead of wage. He, then, decomposes the effect of offshoring on labor demand in three channels: a direct *displacement effect*, an *output effect* and a *substitution effect*.[5] Wright (2013) differentiates between low- and high-skilled labor and assumes that only low-skilled labor can be outsourced. The displacement effect negatively impacts domestic employment, because if firms move more tasks overseas, it takes less domestic tasks to produce a unit of a good. Thus,

[5]We do not discuss the substitution effect, because our data is not detailed enough to estimate this effect. However, the substitution effect is not included in the empirical literature discussed in the introduction (except in Wright, 2013, of course). Wright (2013) does not find a significant impact of this substitution effect on employment and does not focus on this channel.

domestic labor demand falls. The output effect increases domestic labor demand by generating productivity gains via cost-savings. The substitution effect first reflects the substitution between the high-skill factor and the low-skill factor (factor substitution) and second within the low-skill factor the substitution between domestic tasks and foreign tasks (task substitution). While the factor substitution has a positive effect on domestic low-skill employment (and a negative on high skill home employment), task substitution has a negative effect on low-income employment at home. The displacement and the output effect identified by Wright (2013) are closely related to the effects described in earlier literature discussed above.

Summing up, theory comes up with different channels and opposing effects of FDI on home employment. Two channels seem to be important in all models and for both types of FDI: the direct displacement effect and the indirect output or productivity effect. These two channels have opposing effects on home employment and theory is not able to predict which will be the dominating channel. Hence, the theory does not come to a clear conclusion as to whether outward FDI lead to a loss or gain of jobs in the home country. It is important to keep in mind, that the motivation for FDI is decisive in order to investigate the effects on employment, because HFDI and VFDI do affect employment via different mechanisms. Therefore, these different investment types might have distinct effects on domestic labor demand.

5.3 Empirical Strategy

5.3.1 Baseline Specification

As outlined in the previous section two opposing channels explain the relation between domestic and foreign employment. Following Wright (2013), we estimate these two channels in two separate steps reflected in these estimation equations:

$$\ln Empl_{it}^{D} = \alpha^{D} + \beta_1^{D} \ln Empl_{it-1}^{F} + \beta_2^{D} \ln Y_{it-1} + X_{it}'\beta^{D} + \delta_t^{D} + \gamma_i^{D} + \varepsilon_{it}^{D} \tag{5.1}$$

$$\ln Y_{it} = \alpha^{O} + \beta_1^{O} \ln Empl_{it-1}^{F} + X_{it}'\beta^{O} + \delta_t^{O} + \gamma_i^{O} + \varepsilon_{it}^{O}, \tag{5.2}$$

where Equation (5.1) estimates the displacement effect and Equation (5.2) the output effect. $Empl_{it}^{D}$ is domestic labor of the MNE i and $Empl_{it}^{F}$ is the number of employees working abroad. Y_{it} is output measured in net revenue. We control for a set of additional firm specific variables X_{it} (exports, imports and capital). We have no access to export and import data on firm level, they are constructed on industry level.[6] Further, we include year (δ_t) and firm fixed effects (γ_i).

In the first step, we estimate the displacement effect, which quantifies the direct effect of offshoring, where domestic workers are replaced by foreign workers. Therefore, home employment

[6] Firm-fixed effects ensure that time-invariant level differences are absorbed. This means that only the change of exports and imports over time is relevant. Including exports and imports on industry-level is therefore based on the assumption that imports and exports of a MNE develop in the same way as the average of its industry.

$Empl_{it}^{D}$ is regressed on lagged foreign workers $Empl_{it-1}^{F}$ (see Equation 5.1). As discussed in Section 5.2, home employment is as well affected by foreign employment via an opposing indirect effect. This indirect channel affects home employment via firm output. In order to isolate the displacement effect in Equation (5.1), we have to cancel this output channel out. As Wright (2013), we are doing this by holding output fix, i.e. by controlling for output Y_{it-1}. When output is fixed, foreign and home employment are substitutes and more foreign employment means less employment at home. Consequently, we expect the displacement effect to be negative. Because the labor market is not fully flexible, it takes time until a dismissal or hiring of staff realizes. Therefore, both $Empl_{it}^{F}$ and Y_{it} are lagged by one year.

In a second step, we estimate the output effect. As discussed in Section 5.2, the output effect is an indirect effect which works via cost savings and an increase in production. Therefore, we need to estimate the output effect in two steps. We estimate the effect of outward FDI on total output (see Equation 5.2). Again, we lag $Empl_{it}^{F}$, because it takes time until the opening of a new foreign affiliate affects output. However, a significant effect of foreign employment on output in Equation (5.2) is not sufficient to show that the output channel exists. The output effect channel consists of two parts: The effect of foreign employment on output on the one hand and the effect of output on home employment on the other. With Equation (5.2), only the first part of this channel is established. Hence, in order to fully capture the output effect channel, we need to show as well that the effect of output on home employment in Equation (5.1) is significant (and positive). Accordingly, the indirect effect of outward FDI on home employment is only given if both – the effect of outward FDI on output and the effect of output on home employment – can be substantiated. The overall effect of FDI on domestic employment is finally identified by adding up the coefficients from estimating the displacement and the output effect.

By applying a fixed effects model, we control for time-invariant and firm-specific variation, however, there might exist time-variant firm-specific variables that are not observed. One example for time-variant unobserved variables are technology shocks which are absorbed in the error term but affect domestic and foreign employment. This could cause endogeneity issues, which we face by adopting an instrumental variable strategy.

5.3.2 Instrumental Variable Strategy

We are proposing a novel instrument for firm-level outward FDI. We construct the potential foreign employment for each firm using exogenous FDI predictors. Our approach is similar to the strategy used in Desai, Foley and Hines (2009) and the gravity based technique often used in the trade literature (see e.g. Santos Silva and Tenreyro, 2006). Desai, Foley and Hines (2009) construct firm-specific weighted averages of foreign GDP growth as predictor for foreign activity of that firm. The predicted growth rates of foreign activity are then employed to explain changes in domestic activity. The idea behind the instrument is that FDI locations differ significantly between firms and these locations are exposed to different exogenous developments (in Desai, Foley and Hines, 2009, different GDP growth rates) which affect FDI positions. Part of our argument is very similar: We know that FDI destination countries of Swiss MNEs differ significantly across

firms. Given these locations, we can observe exogenous and country-specific shocks, which affect FDI choices of Swiss MNEs. Let us for example assume that one firm's investments are concentrated in Germany, while the other firm's investments are concentrated in the United Kingdom (UK). The firm which is operating in Germany is more exposed to shocks in Germany than the other firm. Hence, a positive shock in Germany is – at least in the short term – supposed to have a bigger positive impact on foreign activities of the firm with mostly German operations. As predictors of these shocks we take inward FDI stock of country c minus Swiss outward FDI stock into the same country c, the exchange rate between the US-Dollar (USD) and country $c's$ currency, as well as other variables described below which have been shown to be important exogenous predictors of FDI flows in the gravity literature (e.g. in Carr, Markusen and Maskus, 2001; Head, Mayer and Ries, 2009; Egger and Pfaffermayr, 2004). We take inward FDI stocks as predictor since Swiss MNEs are highly likely to invest in those countries, where MNEs of other countries invest. The idea behind the exchange rate is the following: If there are two firms, one with affiliates mostly in Germany and the other in the UK and the Euro depreciates, German employees become relatively cheaper and firms which have already affiliates in Germany will expand foreign activities relatively more than firms with affiliates concentrated in the UK.[7]

A challenge is that we observe all the predictors of outward FDI on country level. However, we need firm-specific predictions of foreign employment based on the exogenous predictors named above. We apply the approach of di Giovanni and Levchenko (2009) to overcome this problem. The goal is to predict firm-specific foreign employment in a zero-stage in these three steps:

$$Log\ Empl^{f}_{ict} = \alpha + \beta_{1i} Log\ FDI^{*}_{ct} + X'_{ct}\beta_i + \epsilon_{ict} \tag{5.3}$$

$$\widehat{Log\ Empl}^{f}_{ict} = \hat{\alpha} + \hat{\beta}_{1i} Log\ FDI^{*}_{ct} + X'_{ct}\hat{\beta}_i \tag{5.4}$$

$$\widehat{Empl}^{f}_{it} = \sum_{c=1}^{C} e^{\ \widehat{Log\ Empl}^{f}_{ict}} \tag{5.5}$$

$Log\ Empl^{f}_{ict}$ is the log of foreign employment of firm i in country c and year t. FDI^{*}_{ct} is total inward FDI stock of country c subtracted by the outward FDI stock of Switzerland (CH) in country c ($FDI_{ct} - FDI^{CH}_{ct}$). X_{ct} is a set of exogenous predictors of FDI: exchange rate between the USD and foreign country c's currency, population (log), capital-labor ratio, investment and trade costs (log), distance from Switzerland to c (log), dummy variable for existing investment treaties between Switzerland and c and a dummy variable for common language.[8]

The key of the approach is the first step, i.e. estimation Equation (5.3). Following di Giovanni

[7] We take the USD dollar as base currency, because fluctuations in the exchange rate of the Swiss franc are likely to be caused by events that affect the performance of Swiss firms.

[8] Note that the exogenous predictors should not affect home employment of a Swiss firm via any other channel than foreign employment. Therefore, it is important to include time fixed effects: These time fixed effects absorb global shocks (e.g. a downturn in global economy) that may affect predictors (e.g. FDI^{*}_{ct}) as well as Swiss firms directly. Further, it is important to keep in mind that we control for import and export such that predictors (e.g. distance between countries or population) only affect home employment of a Swiss-based firm via FDI.

and Levchenko (2009), we regress firm-level foreign employment on country-specific predictors to get firm-specific coefficients β_i, i.e. we run regression Equation (5.3) for each firm i. We get different firm-specific coefficients, because firms might follow different foreign investment strategies. Firm-specific investment strategies might address different host-countries, be more or less sensitive to different predictors and change over time. For example, capital-labor ratio might be more important for some firms than for others, depending on the investment strategy and the production function of the firms. In a second step, we predict potential foreign employment per country, $\widehat{Empl}^f_{ict}$, based on exogenous predictors of Equation (5.3). Hence, we keep only the exogenous variation of foreign employment, while the endogenous part in the error term is left out. In a final step, we compute the total potential foreign employment per firm by summing up the exponential of the predicted log of country-specific foreign employment over all countries for each firm (see Equation 5.5).[9] Having predicted potential foreign employment $\widehat{Empl}^f_{it}$, we apply a 2SLS strategy to estimate Equations (5.1) and (5.2) using $\widehat{Empl}^f_{it}$ as instrument.

Since firms have affiliates and therefore positive numbers of foreign employment only in a few countries, they report a lot of zeros for most other countries. This implies that we have to deal with many zero values which would get lost when taking logs. These zero values contain important information in order to consistently estimate the coefficients in Equation (5.3) and allow us to consider cases were firms open up new plants in a foreign country. We face this issue by following Santos Silva and Tenreyro (2006) and use the Poisson pseudo-maximum likelihood (PPML) estimator in order to estimate Equation (5.3).

5.4 Data

The main data source on multinational activities of Swiss firms are the surveys on cross-border capital linkages from the SNB which covers basically firms with a FDI balance sheet lager than 10 million Swiss Francs (CHF). Our data include domestic and foreign employment on a country level of Swiss multinational enterprises over the period 1994 to 2016. It also covers data on firm characteristics such as industry classification and ownership as well as extensive information on domestic and foreign capital links of the firms. To get access to firm-level data we were obliged to obtain the consent of the respective firms due to confidentiality issues and the data protection rule of the SNB. We got access to data of 139 firms. Our sample covers around 56 percent of all domestic employees working for Swiss MNEs. Further, data on firm characteristics such as net revenues and property, plants, equipment (called capital) were extracted from Worldscope, Thomson Reuter's Datastream. The remaining missing data on firm characteristics were finally gathered through access to historical annual reports by the Swiss Economics Archive (Schweizerisches Wirtschaftsarchiv). However, we were not able to fill all the missing values for variables on firm characteristics. Exports and imports are obtained from UN Comtrade Database for trade in goods and from the SNB[10] for trade in service.

[9] In order to get the absolute total of potential foreign employment, we need to sum the exponential because the PPML estimator returns logarithmized results of potential foreign employment per firm and country.

[10] Database can be accessed via `https://data.snb.ch/en`.

Table 5.1: Descriptive Statistics

	Characteristics of Swiss multinationals			
	Minimum	Median	Mean	Maximum
Home Employment	109	1434	4361	53,201
Foreign Employment	9	2390	14,021	275,947
Revenue (in Mio. CHF)	43	2005	7698	95,902
Capital (in Mio. CHF)	17	409	7733	404,094
Countries per firm	1	36	45	124
Employment per Affiliate	1	94	592	56,288

We further assemble data on FDI predictors in order to construct our instrumental variable (see Section 5.3.2). We obtain data on distance between Switzerland and a certain destination, population size and information on common language from the CEPII gravity database.[11] Data on investment costs (Global Competitiveness Index, GCI), exchange rates, capital-labor ratio and bilateral trade costs are retrieved from the World Bank database. Data on FDI stocks and information on bilateral investment treaties are gathered from UNCTAD and information on preferential trade arrangements between countries including WTO investment areas are provided by Word Bank.

We need to drop a number of firms from our sample due to several reasons. First, we drop firms which never have a non-zero observation in foreign employment (34 firms never have a non-missing or non-zero value). Second, we drop Swiss-based subsidiary companies of foreign corporations and consider only corporations headquartered in and directed from Switzerland. Thirdly, we drop firms for which the PPML-estimation does not converge, because we are not able to predict potential foreign employment reliably (8 firms). Finally, we have to drop one or two firms in each estimation, because the instrument (prediction of foreign employment) of these firms is a clear and highly influential outlier.[12]

Our panel is highly unbalanced with hardly any values for the years before 2002. Due to modifications of the methodical concepts in 2014[13] and limited availability of other data used, we are finally left with a sample covering the period 2002 to 2013. The data include firms in manufacturing as well as service (including banks and insurance companies). Table 5.1 shows the summary statistics of the firm characteristics. The size of the MNEs vary considerably in our sample. A few firms operate mainly globally and report high values of foreign employment while others operate mainly domestically. I.e. the number of workers employed in Switzerland varies between 109 and 53,201 per firm and the number of workers abroad between 9 and 275,947. The

[11] Database can be accessed via `cepii.fr`.

[12] These one or two outliers cause a drop of the first-stage Kleibergen-Paap F-statistic of our 2SLS prediction to below 1 from a convenient value of clearly more than the critical 10. We indicate for each estimation, how many firms had to be dropped because of outliers. Usually, a small bank with very few employees at home and abroad is dropped, as well as a firm in the energy sector. Fixed effects estimations show that estimates are otherwise not sensible to the inclusion or exclusion of these firms.

[13] Until 2013 staff numbers included both minority and majority participations and were stated in relation to the capital participation of the investor. As of 2014 – in line with international methodology – staff numbers only include majority participations. Further, no longer proportional, but absolute numbers of staff abroad are stated.

average firm in the dataset has about 4,361 domestic employees and 14,021 employees working abroad. Furthermore, firms in our dataset have an affiliate in at least one country and on average in 45 countries. The affiliate size ranges from 1 employee to a maximum of about 56,000 employees with an average of 592 employees. The median firm has affiliates in 36 different countries with a median size of 94 employees.

Table 5.2: Employment by Destination

	2002	2013
Domestic employment	223,482	429,080
Total foreign employment	907,752	1,283,284
Employment in high-income countries	627,047	672,839
Share of foreign empl. in high-income	69%	52%
Employment in upper middle-income countries	71,358	130,305
Share of foreign empl. in upper middle-income	8%	10%
Employment in lower middle-income countries	166,575	368,167
Share of foreign empl. in lower middle-income	18%	29%
Employment in low-income countries	42,773	111,972
Share of foreign empl. in low-income	5%	9%

Note: Grouping into income categories according to World Bank classification in 2002.

Table 5.2 reports total number of domestic employment and employment in foreign countries aggregated over all firms and classified by income-level aggregates. We split employment in foreign affiliates according to the country's income level in high-, upper middle-, lower middle- and low-income. The classification is done according to the World Bank income classification of 2002. While Swiss MNEs located their affiliates mostly in high-income countries in 2002 (69 percent of all foreign employees were engaged in high-income countries), lower middle-income and low-income countries have gained importance as destination market for foreign investments in the last decade. During the wave of globalization especially countries like China and India, which are classified as lower middle-income or low-income countries, became increasingly more attractive for foreign investment. These two countries are the main driver of the rise in employment in the lower middle-income and low-income country aggregates.[14] In the period observed, foreign employment increased in high-income countries by almost 12 percent, while it more than doubled in lower middle- and low-income countries. These findings are well aligned with global FDI patterns: Almost all FDI in the 1990s took place between high-income countries and tended consequently to be HFDI, but VFDI have become increasingly important in recent decades (Barba Navaretti, Venables and Barry, 2006, p.32). Figure A.1 in Appendix A.1 displays the ten most important destination countries in terms of foreign employment for Switzerland. In 2002 the US and Germany were clearly the most important destination countries followed by other

[14] India was classified as low-income country until 2009 and is, therefore, included in our low-income aggregate.

major economies in Europe as the United Kingdom and France. Further behind, large emerging countries like Brazil, Russia and China were following as well as other major European economies as Italy and Spain. Over the last decade major emerging economies such as China, Brazil and India became more important countries for outward FDI. In 2013, China was as important as Germany for Swiss multinationals while the US remained the main destination of FDI in terms of employment in foreign affiliates.

5.5 Results

5.5.1 Benchmark Estimations

Table 5.3 reports the results of estimating Equation (5.1) quantifying the displacement effect. Models in columns (1) to (3) are estimated with the fixed effects estimation approach, while models in columns (4) to (6) show the results of the IV estimation within a 2SLS framework. The results in column (1) show that there is no significant correlation between domestic and foreign employment, when we do only control for firm and year fixed effects. Since we expect a negative displacement and a positive output effect to be at work, which might outweigh each other, this result is not meaningful. Therefore, we control for output to isolate the displacement channel. By doing so, the effect of foreign employment becomes, as expected negative and significant (see columns 2 and 3). An increase of foreign employment of 10 percent would lead to a decrease of domestic within firm employment of 1 percent. This is in line with the theory and the underlying concept of the estimation strategy. We control for capital, exports and imports in order to capture changes in capital intensity and trade-related movements. These additional controls do not alter the coefficient of interest.

To cope with potential endogeneity issues, we instrument foreign employment as described in Section 5.3. The Kleibergen-Paap F-statistics of the first stage show that our instrument is relevant. Using instrumented foreign employment, the displacement effect disappears completely (columns 5 and 6): The coefficient of interest becomes insignificant and is close to zero. Only output seems to be relevant and is positively associated with domestic employment. However, the coefficient of output loses significance, as soon as we control for capital, exports and imports.

In a second step we estimate the output effect, which represents the indirect productivity effect from offshoring on domestic employment. We have discussed in Section 5.2 that we need two steps to fully substantiate the output effect: The effect of foreign employment on output and, additionally, the effect of output on home employment is expected to be positive. First, we consider the effect of foreign employment on output. In order to do that we estimate Equation (5.2) by using the same instrument as in Equation (5.1) for foreign employment. Results are reported in Table 5.4. We find a highly significant and positive association between foreign employment and output. The estimate in column (4) indicates that an increase in foreign employment by 10 percent is associated with a rise in output of 3.8 percent. Hence, the first part of the output effect channel is established: Higher foreign employment is significantly and positively associated

with output. Results in Table 5.3 show that the evidence of the second part of the output channel – whether output is positively associated with domestic employment – is less clear. Output is on the one hand positively and significantly correlated with home employment when we are not including capital and trade (column 5). But on the other hand, the IV point estimates of output become marginally smaller if we include all controls and bootstrap standard errors get inflated. Therefore, the IV estimates of the relation between output and domestic employment remain positive but turn insignificant in column (6). Hence, we do not find clear evidence of the second part of the output channel. However, results point toward the existence of the output channel.

Summing up, we find that the negative displacement effect seems to be irrelevant when using the IV approach, while it depends on the model whether we find a positive output effect. Hence, outward FDI do not appear to have an important effect on home employment when considering all types of FDI. As mentioned, depending on the FDI strategy of a firm different mechanisms might be at work and considering all types of FDI in one estimation might not allow to disentangle these different effects. Further, in 2002, about 69 percent of all Swiss outward FDI flow to other high-income countries and are therefore, mainly HFDI. Even though middle- and low-income countries gain importance over time, high-income countries remain the primary destination of outward FDI of Swiss MNEs. As discussed in Section 5.2, HFDI substitute domestic exports. However, if exports to a certain country were low or zero before firms conduct HFDI in that country, there are not much exports to be substituted and thereby the displacement effect is small or nonexistent. Hence, an explanation why there seems not to exist a clear displacement effect in Switzerland, might be the investment strategy of Swiss firms seeking to open (new) markets. The small domestic market and potentially higher transport costs due to the lack of sea access,

Table 5.3: Displacement Effect

Dep. Var.:	Not instrumented			Instrumented		
Log of Home Empl.	(1)	(2)	(3)	(4)	(5)	(6)
Lag For. Empl. (log)	0.01	−0.09***	−0.10***	0.22	0.06	0.04
	(0.05)	(0.03)	(0.03)	(0.21)	(0.27)	(0.29)
Lag Output (log)		0.47***	0.39***		0.35*	0.29
		(0.13)	(0.12)		(0.20)	(0.23)
Capital (log)			0.08			0.08
			(0.07)			(0.09)
Exports (log)			0.09			0.10
			(0.18)			(0.20)
Imports (log)			−0.04			−0.01
			(0.08)			(0.15)
Observations	557	557	557	557	557	557
First stage F-stat.				34.26	19.05	18.21
Firm & Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–3 and bootstrap std. err. (10,000 iterations) for columns 4–6. Two firms have been removed from this estimation for reasons explained in footnote 12.
* p<0.10, ** p<0.05, *** p<0.01.

Table 5.4: Output Effect

Dep. Var.:	Not instrumented		Instrumented	
Log of Output	(1)	(2)	(3)	(4)
Lag For. Empl. (log)	0.24***	0.18***	0.45***	0.38***
	(0.08)	(0.07)	(0.13)	(0.13)
Capital (log)		0.15*		0.12
		(0.08)		(0.08)
Exports (log)		0.34***		0.29**
		(0.12)		(0.11)
Imports (log)		0.01		0.06
		(0.08)		(0.11)
Observations	557	557	557	557
First stage F-stat.			34.26	29.08
Firm & Year FE	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–2 and bootstrap std. err. (10,000 iterations) for columns 3–4. Two firms have been removed from this estimation for reasons explained in footnote 12.
* p<0.10, ** p<0.05, *** p<0.01.

may further explain why Swiss MNEs are opening up affiliates overseas primarily in order to gain market access. In order to further investigate and disentangle the effects of different types of FDI on home employment, we need to distinguish between vertical and horizontal outward FDI. This is done in the following section.

5.5.2 FDI by Destination

The distinction between VFDI and HFDI is crucial in determining the displacement and the output effect. While VFDI are motivated by making use of wage differentials and cost savings, market seeking and technology sourcing are main objectives of HFDI. Therefore, the mechanisms at work are different and effects of the respective type of outward FDI might be different.

A crude measure to differentiate between the types of FDI is by looking at destination countries and classify investments to lower-income countries as VFDI and to high-income countries as HFDI.[15] This distinction by income levels is based on the idea that wages are lower in low-income countries and therefore, they are more attractive for VFDI. Furthermore, purchasing power is relatively low and therefore, are these low-income markets less attractive to sell products. In high-income countries, on the other hand, purchase power is high and the technology closer to the frontier, which is important for market seeking or technology sourcing and therefore HFDI. We incorporate this distinction between VFDI and HFDI by splitting countries to high-, upper resp. lower middle- and low-income countries according to the World Bank classification in the year 2002.

[15] This link between type of FDI and destination country is, for instance, as well done in Harrison and McMillan (2011) and Debaere, Lee and Lee (2010)

FDI to High-Income Countries

In a first step, we focus on FDI to high-income countries. This means, we run the zero stage (Equations 5.3 to 5.5) as well as estimations of the displacement and output effect considering only foreign employment in high-income countries. Figure A.2 in Appendix A.1 shows the share of employment in the 10 most important high-income destinations. In 2002 the USA was clearly the most important destination followed by Germany and further behind other major European

Table 5.5: Displacement Effect for FDI into high-income countries

Dep. Var.:	Not instrumented			Instrumented		
Log of Home Empl.	(1)	(2)	(3)	(4)	(5)	(6)
Lag For. Empl. (log)	0.04	−0.05	−0.05	0.15	0.03	0.01
	(0.05)	(0.03)	(0.03)	(0.14)	(0.14)	(0.14)
Lag Output (log)		0.43***	0.36***		0.37***	0.31**
		(0.13)	(0.12)		(0.13)	(0.15)
Capital (log)			0.08			0.08
			(0.07)			(0.08)
Exports (log)			0.10			0.09
			(0.18)			(0.19)
Imports (log)			−0.02			−0.02
			(0.08)			(0.13)
Observations	553	553	553	553	553	553
First stage F-stat.				37.00	25.28	26.00
Firm & Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–3 and bootstrap std. err. (10,000 iterations) for columns 4–6. Two firms have been removed from this estimation for reasons explained in footnote 12.
* p<0.10, ** p<0.05, *** p<0.01.

economies as the United Kingdom, France or Italy. The most salient change in 2013 compared to 2002 is that very distant countries like Canada, Japan and Australia seem to become more important destinations. As mentioned above, we expect most of these FDI to be HFDI.

The regression results for the effect of investing to high-income countries and therefore performing HFDI are shown in Tables 5.5 and 5.6. The Kleibergen-Papp F-statistics of the first stage of our estimations are much higher compared to the values in the baseline regressions including all types of FDI and show that our instrument is relevant. Compared to the baseline results in Table 5.3, the magnitude of the displacement effect in the fixed effects approach is cut in half and is not significant anymore (see columns 2 and 3 in Table 5.5). The results of the IV strategy show a similar pattern as in the baseline regression: Point estimates of foreign employment in Table 5.5 are close to zero and clearly not significant. Hence, we do not find any evidence of the negative displacement channel. On the other hand, foreign employment is positively and significantly associated with output (see Table 5.6) and output is positively and significantly associated with home employment (see Table 5.5). There is significant evidence of the existence of both steps of the output effect. Therefore, outward FDI to high-income countries

Table 5.6: Output Effect for FDI into high-income countries

Dep. Var.:	Not instrumented		Instrumented	
Log of Output	(1)	(2)	(3)	(4)
Lag For. Empl. (log)	0.24***	0.16**	0.34***	0.28***
	(0.08)	(0.07)	(0.11)	(0.08)
Capital (log)		0.15*		0.12
		(0.07)		(0.08)
Exports (log)		0.32**		0.27**
		(0.12)		(0.12)
Imports (log)		−0.02		0.00
		(0.07)		(0.12)
Observations	553	553	553	553
First stage F-stat.			37.00	35.64
Firm & Year FE	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–2 and bootstrap std. err. (10,000 iterations) for columns 3–4. Two firms have been removed from this estimation for reasons explained in footnote 12.
* p<0.10, ** p<0.05, *** p<0.01.

stimulate domestic employment – even though the effect is rather small. A simple combination of both steps of the output effect as in Wright (2013) gives us the following overall effect: An increase of foreign employment by 10 percent is associated with an increase of home employment of about 0.9 percent (including all controls) to 1.3 percent (excluding capital and trade controls) via the output channel. Since results in this sections show that HFDI do not substitute exports (i.e. there is no displacement effect), one might conclude that the main motivation of outward FDI in Switzerland is opening new markets or technology sourcing.

FDI to Upper Middle-Income Countries

Our instrument is not valid for this category of FDI (see Kleibergen-Papp F-statistics of the first stage in Tables A.1 and A.2 in Appendix A.2). Therefore, we refrain from discussing the results. Nevertheless, the results can be found in Appendix A.2.

Reasons why we fail to reliably estimate the effect in this case, might be that the upper-middle income group is a relatively small group of heterogeneous countries and overall only a relatively small number of Swiss MNE employees is active in these countries. In total, Swiss MNEs in our sample are active in only 23 upper middle-income countries, which makes the upper middle-income economies the smallest category in terms of number of destinations. Figure A.3 in Appendix A.1 shows that these are mostly Eastern European or Latin American destinations. Compared to the most important countries of other categories these are relatively small economies. With a share of 8 percent of total foreign employment in 2002 and no country in the top ten destinations for Swiss MNE upper middle-income countries are less important as destination for Swiss MNE than high-income or lower middle-income countries.

FDI to Lower Middle-Income Countries

Lower middle-income countries are with a share of 18 percent in 2002 and 29 percent in 2013 the most important destination for Swiss outward FDI after high-income countries. Figure A.4 in Appendix A.1 shows that most important lower middle-income destination countries are big emerging markets outside of Europe as Brazil, Russia or China. These three large emerging

Table 5.7: Displacement Effect for FDI into lower middle-income countries

Dep. Var.:	Not instrumented			Instrumented		
Log of Home Empl.	(1)	(2)	(3)	(4)	(5)	(6)
Lag For. Empl. (log)	−0.08*	−0.11**	−0.11**	−0.11	−0.17*	−0.16*
	(0.04)	(0.05)	(0.05)	(0.08)	(0.09)	(0.09)
Lag Output (log)		0.31***	0.28**		0.35***	0.32**
		(0.09)	(0.10)		(0.12)	(0.14)
Capital (log)			0.01			0.01
			(0.05)			(0.13)
Exports (log)			0.17			0.13
			(0.25)			(0.27)
Imports (log)			−0.07			−0.07
			(0.14)			(0.35)
Observations	386	386	386	386	386	386
First stage F-stat.				27.42	23.17	26.51
Firm & Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–3 and bootstrap std. err. (10,000 iterations) for columns 4–6. One firm has been removed from this estimation for reasons explained in footnote 12.
* p<0.10, ** p<0.05, *** p<0.01.

markets also belong to the most important destinations for Switzerland when considering all destinations. Due to the high wage level in Switzerland, lower-middle-income countries might mainly be of interest for VFDI for Swiss MNE. However, some of these countries such as China, Brazil and Russia might be as well interesting for HFDI, because of their market size and increasing purchase power (at least during the period observed).

Tables 5.7 and 5.8 show the results of FDI to lower middle-income countries. In contrast to our findings of overall FDI and FDI to high-income countries, we find evidence of a negative displacement effect in the IV model (columns 5 and 6 in Table 5.7). IV estimates show that an increase of foreign employment in lower middle-income countries by 10 percent is associated with a significant decrease of employment at home by 1.6 percent. Furthermore, there is evidence for the positive output effect: Foreign employment is positively and significantly associated with output (see Table 5.8), while output is positively associated with home employment (see Table 5.7). When we are combining the effects as it is done in Wright (2013), we find that overall an increase in foreign employment in lower middle-income countries by 10 percent is associated with a decrease of home employment by about 1.1 percent. Hence, the negative displacement effect outweighs the positive output effect in this case.

Table 5.8: Output Effect for FDI into lower middle-income countries

Dep. Var.:	Not instrumented		Instrumented	
Log of Output	(1)	(2)	(3)	(4)
Lag For. Empl. (log)	0.13**	0.11**	0.16**	0.17**
	(0.05)	(0.05)	(0.08)	(0.07)
Capital (log)		0.13		0.12
		(0.11)		(0.15)
Exports (log)		0.41**		0.44*
		(0.17)		(0.23)
Imports (log)		0.01		0.01
		(0.20)		(0.36)
Observations	386	386	386	386
First stage F-stat.			27.42	30.19
Firm & Year FE	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1-2 and bootstrap std. err. (10,000 iterations) for columns 3–4. One firm has been removed from this estimation for reasons explained in footnote 12.
* p<0.10, ** p<0.05, *** p<0.01.

So, while overall FDI and in particular FDI to high-income countries tend to have a positive effect on home employment, FDI to lower middle-income countries seem to decrease home employment in the short-run. Hence, the different mechanics behind HFDI and VFDI actually do affect home employment differently: While the positive output effect dominates for HFDI, the negative output effect is dominating for VFDI to lower middle-income countries.

FDI to Low-Income Countries

Ultimately, we look at FDI to low-income countries. Figure A.5 in Appendix A.1 shows that almost 40 percent of foreign employees in low-income countries are located in India. Other important low-income destination countries are Zambia, Indonesia or Pakistan. Due to the low purchase power, most of these countries do not seem to be interesting for HFDI.[16] Moreover, low-income countries are generally not at the technological frontier and therefore, technology sourcing is as well unlikely to be the motivation behind FDI to these countries. Wage differentials seem to be the main motivation behind FDI in these countries. Furthermore, countries with unstable political institutions and very low purchase power but rich in natural resources as the Democratic Republic of Congo seem to be attractive destinations because of their natural resources and not because of low wages and cost-savings in production. Hence, it is not likely that production stages from Switzerland will be shifted to countries with very low incomes but rich in natural resources. The goal behind FDI to these countries is rather natural resource sourcing than a substitution of Swiss production. Thus, FDI to low-income countries appear to be heterogeneous and it is not clear what the prevailing motivation behind FDI to low-income countries is. Further,

[16] Countries like India or Indonesia could of course be as well interesting for HFDI because of their market size and rapidly growing middle-class. However, HFDI are overall rather the exception, while VFDI are the prevailing FDI type flowing to these type of countries.

it is important to know, that only very few Swiss firms open up relatively large affiliates in these countries. In 2013 only 54 firms of 103 in the data have affiliates in low-income countries, while 100 of 103 firms in the data have affiliates in high-income countries.

Table 5.9: Displacement Effect for FDI into low-income countries

Dep. Var.:	Not instrumented			Instrumented		
Log of Home Empl.	(1)	(2)	(3)	(4)	(5)	(6)
Lag For. Empl. (log)	−0.09	−0.12*	−0.09	0.02	−0.05	0.03
	(0.06)	(0.07)	(0.06)	(0.13)	(0.11)	(0.21)
Lag Output (log)		0.27**	0.21*		0.24**	0.12
		(0.11)	(0.11)		(0.10)	(0.15)
Capital (log)			−0.04			−0.02
			(0.03)			(0.15)
Exports (log)			0.58***			0.73**
			(0.20)			(0.37)
Imports (log)			0.00			0.06
			(0.10)			(0.34)
Observations	250	250	250	250	250	250
First stage F-stat.				88.46	93.86	49.20
Firm & Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–3 and bootstrap std. err. (10,000 iterations) for columns 4–6.
* p<0.10, ** p<0.05, *** p<0.01.

Tables 5.9 and 5.10 present the results of the estimations considering only FDI to low-income countries. The results might be compared to the results found in the benchmark estimations in Table 5.3 and 5.4: The displacement effect disappears completely as soon as we instrument foreign employment. The point estimate is again close to zero and clearly not significant (see Table 5.9). On the other hand, we find a positive association between foreign employment and output for the fixed effects approach and for the IV approach (see Table 5.10). Output tends to be positively associated with domestic employment, although the positive association is not significant when including all control variables (see column 6 in Table 5.9). So, in the case of FDI to low-income countries we do not find evidence of a negative displacement effect and we find no robust evidence of a positive output effect. Similarly as for the estimation with upper middle-income countries, this lack of significance might be attributed to the limited number of observations, but also to the heterogeneity of countries in our low-income sample. On the one hand the data covers foreign employment in countries with large market size such as India and Indonesia which are interesting for VFDI (and HFDI), on the other hand it also includes countries rich in natural resources (e.g. Democratic Republic of Congo, Zambia, etc.), which are rather interesting for resource sourcing instead of production for MNEs.

Table 5.10: Output Effect for FDI into low-income countries

Dep. Var.:	Not instrumented		Instrumented	
Log of Output	(1)	(2)	(3)	(4)
Lag For. Empl. (log)	0.14**	0.16**	0.27*	0.30**
	(0.06)	(0.06)	(0.15)	(0.13)
Capital (log)		0.14		0.16
		(0.13)		(0.23)
Exports (log)		0.76**		0.87*
		(0.37)		(0.51)
Imports (log)		0.10		0.18
		(0.13)		(0.42)
Observations	250	250	250	250
First stage F-stat.			88.46	62.10
Firm & Year FE	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–2 and bootstrap std. err. (10,000 iterations) for columns 3–4.
* p<0.10, ** p<0.05, *** p<0.01.

5.5.3 Robustness Checks

In Section 5.5.2 we have grouped the countries into different income categories according to the World Bank definition from 2002. We have used this classification as a rough measure to classify the FDI into HFDI and VFDI. However, results discussed before could be driven by large destination countries or particularities of the classification itself. Therefore, we run two different robustness checks for each destination category, except for the upper middle-income group, where we were not able to estimate benchmark results reliably with the data at hand.

First, we exclude dominant countries that might drive the results. In the case of FDI to high-income countries we exclude the US with a share of almost 30 percent of the foreign employees in high-income countries. In the case of lower-middle income countries we exclude China with a share of 35 percent in 2013 and in the case of low-income countries we drop India with a share of almost 40 percent in 2013.

Second, we re-run the whole estimation by destination group using the classification of the World Bank in the year 2013 instead of 2002. Many countries changed the income group until 2013 (see Table A.15 in Appendix A.4 for an overview) and therefore, this robustness check allows us to test whether results found in Section 5.5.2 stem from particularities of the classification.

Robustness Check: Drop Dominant Destinations

In a first robustness check we drop the US as dominant destination in the high-income group. Results in Tables A.3 and A.4 in Appendix A.3.1 show that the exclusion of the USA does not importantly change the results: We still do not find any evidence of the displacement effect when instrumenting for foreign employment (columns 5 and 6 in Table A.3), but – as in the benchmark

results – we find evidence for both steps of the positive output effect.

In a second robustness check we drop China as the dominant destination in the lower middle-income group (see Tables A.7 and A.8 in Appendix A.3.3). Dropping China alters the results completely. The negative displacement effect disappears: Points estimates are small and insignificant. While there is still significant evidence for the first step of the output effect (Table A.8), there is no evidence for the second step anymore: Point estimates of the second step are cut in half and not significant anymore (Table A.7 columns 5 and 6). So, it appears that the negative overall effect of FDI to lower-middle income countries found in the benchmark estimations is driven by China and disappears completely as soon as China is excluded.

In a third robustness check India as the dominant destination of the low-income group is dropped. Tables A.11 and A.12 in Appendix A.3.5 show that the results found in the benchmark estimation for low-income countries heavily depend on India: Not only does the instrument lose its relevance (Kleibergen-Papp F-statistics of the first stage are consistently below the value of 5), but we do not find any evidence for the displacement or output effect anymore.

In conclusion, results for lower middle- and low-income destination are driven by single dominant countries and not robust when these countries are dropped. However, the positive effect of FDI to high-income destination is robust when dropping the US as dominant destination.

Robustness Check: Income Classification of 2013

In this section we re-estimate the regressions by income level classification of 2013 instead of 2002. Table A.15 in Appendix A.4 shows to which income group countries belonged in 2002 and 2013. Tables A.5 and A.6 in Appendix A.3.2 show the effect of FDI to high-income countries according to the classification of 2013. Point estimates remain very similar to the benchmark estimation. The only remarkable difference to the benchmark results is that the second part of the output effect loses significance when all controls are included (see Table A.5 column 6).

Considering the robustness check of the lower middle-income category, it is important to know that seven of the ten most important lower middle-income countries in 2002 were classified either as upper middle- or high-income countries in 2012. Most importantly, China – which is driving the results in the benchmark estimation – as well as Brazil and Russia switched from being classified as lower-middle income countries in 2002 to upper-middle income countries in 2013. On the other hand, many important low-income destinations are classified as lower middle-income countries in 2013. In particular, India is classified as a lower-middle income country since 2009. Tables A.9 and A.10 in Appendix A.3.4 present the results using the 2013 classification. We find no evidence for the negative displacement effect. Results show further that foreign employment is positively associated with output (see Table A.10 columns 3 and 4). Hence, we find evidence for the first part of the output effect. Finally, it depends on the specification whether we find evidence for the second part of the output effect (see Table A.9): We only find a positive and significant association between output and home employment in column (5) but not when we include all covariates in column (6). Hence, as soon as we are considering the classification of

2013 the effect of FDI to lower middle-income countries changes from negative to (tendentiously) positive. This is no surprise, since most important lower middle-income destinations of 2002 became upper middle-income countries, while most important low-income destinations moved up to the lower middle-income category. So, the results we are finding in this robustness check is very similar to the results of the benchmark estimation of the low-income category. However, this robustness check shows that results are sensitive to the classification of countries.

In the case of low-income countries, many important countries moved up to the lower middle-income group. Most importantly India, that was driving the results in the benchmark estimations, but in total eight of the ten most important destinations in 2002 were classified as lower middle-income countries in 2013. Consequently, observations drop to 100 and it is not surprising that we do not find any significant results (see Tables A.13 and A.14 in Appendix A.3.6).

In conclusion, robustness tests show that the positive effect found for high-income countries is fairly robust: Results do not change much if the dominant country is dropped nor if the income classification of 2013 is used. However, effects found for all other income levels do not seem to be robust and heavily depend on single dominant countries or the year of the classification. In particular, it becomes apparent that the negative effect found for lower middle-income countries is driven by China (which is comparable to the finding of Debaere, Lee and Lee (2010) for South Korea).

5.6 Conclusion

Economic theory suggests that there are negative as well as positive effects of offshoring on domestic labor demand. However, theory is not able to predict clearly, whether positive effects are able to offset negative effects. Empirical work does not come to a clear conclusion either: Results depend on the context of the country observed and on the type and destination of FDI. We use firm-level data containing firms of the manufacturing and the service sector in order to examine the effect of offshoring on home employment in the case of Switzerland, a small economy with relatively high outward FDI stock.

We find that it is crucial to distinguish between different types of FDI. Using fixed effects and an instrumental variable approach we find no evidence of the negative displacement effect, but a positive and significant output effect of FDI to high-income countries (i.e. mainly HFDI). On the other hand we find a significant and negative displacement effect which outweighs the positive output effect in the case of FDI to lower middle-income countries (i.e. mainly VFDI). However, while the positive effect found for FDI to high-income countries is robust, the effect of FDI to lower-middle income countries is driven by China and disappears as soon as China is excluded. Further, it is important to keep in mind that these positive short-run effects of HFDI and negative effects of VFDI are rather moderate. We find no evidence of a negative displacement effect when we are considering the IV results of FDI to all countries and only to low-income countries. For both – FDI to all countries and only to low-income countries – results point toward to a positive output and, hence, overall effect. However, the effect of FDI to low-

income countries is driven by single dominating destinations and not robust to changes in the income classification definitions.

Summing up, Swiss outward FDI stock and flows are tremendous in relative size, but do barely affect total domestic jobs within firms. If so, there seems to be rather a positive effect than a negative. It is important to keep in mind that the goal of this paper is to estimate the overall effect of outward FDI on home employment and that effects might be very different between low- and high-skilled labor (see e.g. Wright, 2013).

A reason why the displacement effect does not seem to exist in Switzerland might be, that Swiss MNEs follow a HFDI strategy and primarily invest in other high-income countries. HFDI seem to stimulate total domestic jobs, although the magnitude of the effect is rather small. So, we are concluding that outward FDI do not endanger the total number of domestic jobs in the case of Switzerland – on the contrary they seem to create jobs, especially if the MNE is investing into another high-income country. Although there is a trend to more outward FDI into upper middle- but more importantly lower middle- and low-income countries, the majority of Swiss outward FDI still flows into other high-income countries.

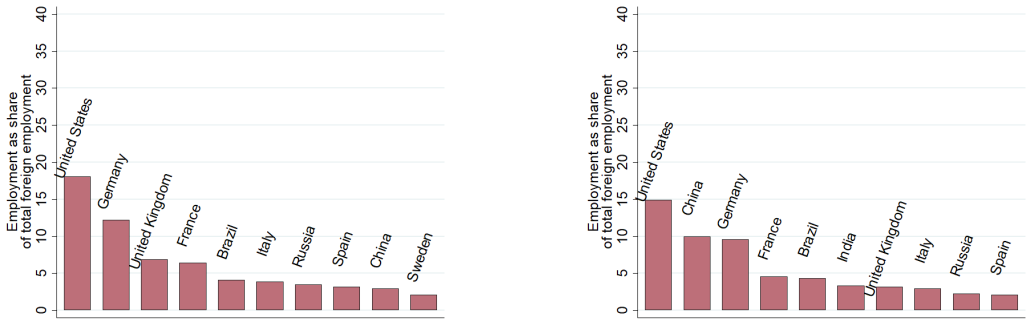
There are limitations in comparing our results with other existing evidence given the different estimation strategy and underlying data. Our approach is most related to Desai, Foley and Hines (2009), who find positive effects of foreign activity of US MNEs on domestic employment. Harrison and McMillan (2011) present evidence of different effects given the destination country and the tasks performed abroad, where investments to low-income countries are associated with lower domestic employment, while they find positive effects of FDI into high-income countries. Wright (2013) finds as well a slightly positive overall effect. However, he examines as well the effect on low skilled labor, where he finds a negative effect. Hijzen, Jean and Mayer (2011) and Debaere, Lee and Lee (2010) pursue a different empirical approach but find similar effects. Hijzen, Jean and Mayer (2011) find positive effects of HFDI and no effects of VFDI in France, while Debaere, Lee and Lee (2010) find negative effects of FDI in lower-income countries and positive effects on employment of FDI to higher-income countries for South Korea.

Finally, it is important to stress that our estimation approach and the underlying data do not take into consideration that firms might have to close down their business in the longer run, if they do not have the possibility to engage in outward FDI. Furthermore, we do not consider backward or forward spillovers on other firms.[17] These points indicate that the long-run positive effect of outward FDI on home employment might even be more pronounced than our findings suggest.

[17] E.g. Tang and Altshuler (2015) find positive spillover effects of outward FDI on domestic suppliers.

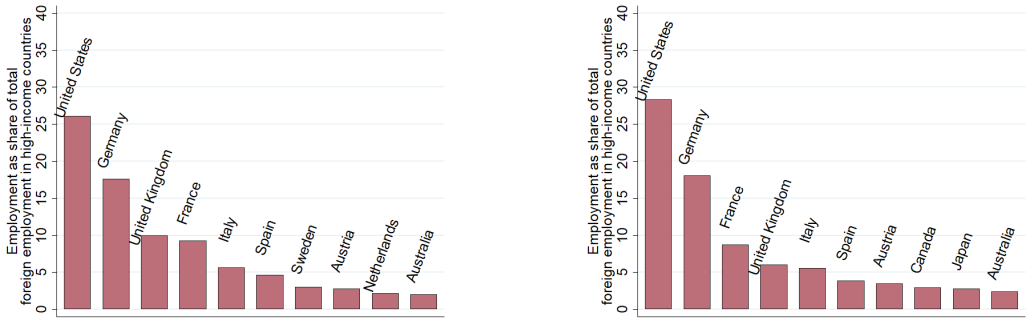
A. Appendices

A.1 Appendix: Foreign Employment by Destination



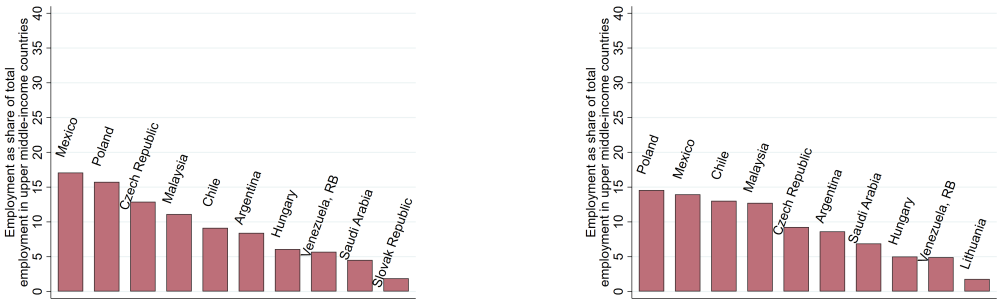
(a) 2002 (N=78) (b) 2013 (N=103)

Figure A.1: Share of total foreign employment by destination
Notes: Number of firms N in parentheses.



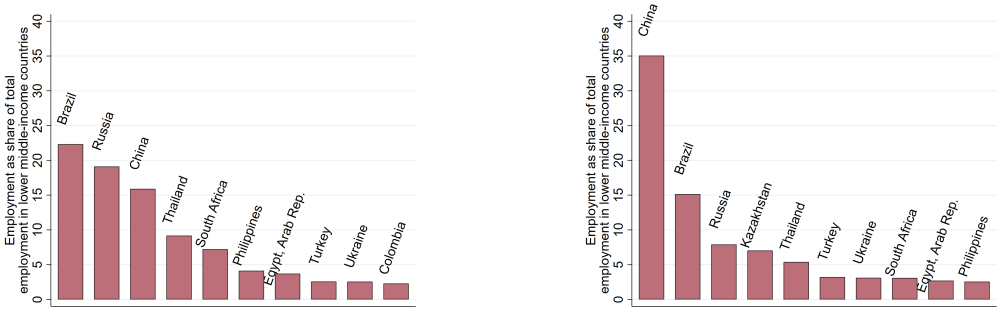
(a) 2002 (N=78) (b) 2013 (N=100)

Figure A.2: Share of total foreign employment in high-income countries
Notes: Number of firms N in parentheses.



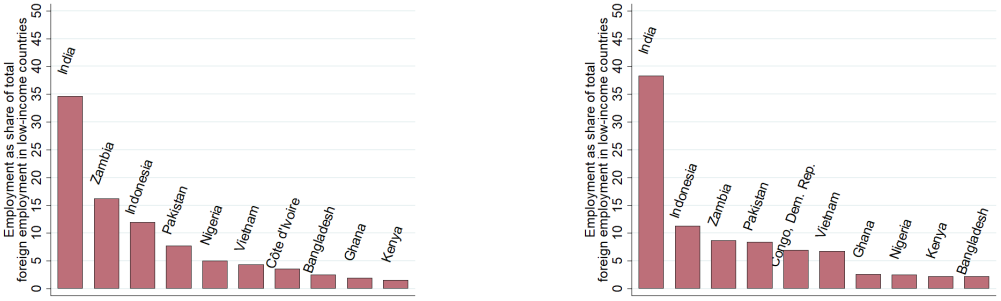
(a) 2002 (N=46) (b) 2013 (N=74)

Figure A.3: Share of total foreign employment in upper middle-income countries
Notes: Number of firms N in parentheses.



(a) 2002 (N=45) (b) 2013 (N=73)

Figure A.4: Share of total foreign employment in lower middle-income countries
Notes: Number of firms N in parentheses.



(a) 2002 (N=32) (b) 2013 (N=54)

Figure A.5: Share of total foreign employment in low-income countries
Notes: Number of firms N in parentheses.

A.2 Appendix: FDI to Upper Middle-Income Countries

Table A.1: Displacement Effect for FDI into upper middle-income countries

Dep. Var.:	Not instrumented			Instrumented		
Log of Home Empl.	(1)	(2)	(3)	(4)	(5)	(6)
Lag For. Empl. (log)	0.01	−0.02	−0.01	0.09	0.07	0.01
	(0.04)	(0.03)	(0.03)	(0.28)	(1.91)	(1.20)
Lag Output (log)		0.36**	0.21*		0.32	0.20
		(0.16)	(0.12)		(1.77)	(0.76)
Capital (log)			0.01			0.01
			(0.06)			(0.21)
Exports (log)			0.66***			0.66
			(0.18)			(1.72)
Imports (log)			0.08			0.08
			(0.09)			(0.46)
Observations	369	369	369	369	369	369
First stage F-stat.				2.80	2.81	2.93
Firm & Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–3 and bootstrap std. err. (10,000 iterations) for columns 4–6. Five firms have been removed from this estimation for reasons explained in footnote 12. * p<0.10, ** p<0.05, *** p<0.01.

Table A.2: Output Effect for FDI into upper middle-income countries

Dep. Var.:	Not instrumented		Instrumented	
Log of Output	(1)	(2)	(3)	(4)
Lag For. Empl. (log)	0.13*	0.07	0.06	−0.02
	(0.07)	(0.05)	(0.48)	(0.45)
Capital (log)		0.17		0.19
		(0.13)		(0.27)
Exports (log)		0.54**		0.54
		(0.22)		(0.41)
Imports (log)		0.11		0.10
		(0.21)		(0.44)
Observations	369	369	369	369
First stage F-stat.			2.80	2.79
Firm & Year FE	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–2 and bootstrap std. err. (10,000 iterations) for columns 3–4. Five firms have been removed from this estimation for reasons explained in footnote 12. * p<0.10, ** p<0.05, *** p<0.01.

A.3 Appendix: Robustness Checks

A.3.1 High-Income Countries: Drop USA

Table A.3: Displacement effect for FDI into high income countries: Drop USA

Dep. Var.:	Not instrumented			Instrumented		
Log of Home Empl.	(1)	(2)	(3)	(4)	(5)	(6)
Lag For. Empl. (log)	0.03	−0.04	−0.04	0.05	−0.04	−0.04
	(0.06)	(0.04)	(0.04)	(0.15)	(0.13)	(0.12)
Lag Output (log)		0.42***	0.34***		0.42***	0.34**
		(0.13)	(0.12)		(0.13)	(0.14)
Capital (log)			0.08			0.08
			(0.07)			(0.09)
Exports (log)			0.11			0.11
			(0.19)			(0.19)
Imports (log)			−0.02			−0.02
			(0.08)			(0.12)
Observations	545	545	545	545	545	545
First stage F-stat.				60.63	33.00	32.34
Firm & Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–3 and bootstrap std. err. for columns 4–6. Two firms have been removed from this estimation for reasons explained in footnote 12. * p<0.10, ** p<0.05, *** p<0.01.

Table A.4: Output effect for FDI into high-income countries: Drop USA

Dep. Var.:	Not instrumented		Instrumented	
Log of Output	(1)	(2)	(3)	(4)
Lag For. Empl. (log)	0.21***	0.14**	0.21*	0.18**
	(0.08)	(0.06)	(0.13)	(0.09)
Capital (log)		0.16*		0.16*
		(0.08)		(0.09)
Exports (log)		0.33**		0.31**
		(0.13)		(0.14)
Imports (log)		−0.02		−0.01
		(0.08)		(0.13)
Observations	545	545	545	545
First stage F-stat.			60.63	51.44
Firm & Year FE	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–2 and bootstrap std. err. for columns 3–4. Two firms have been removed from this estimation for reasons explained in footnote 12. * p<0.10, ** p<0.05, *** p<0.01.

A.3.2 High-Income Countries: Income Classification of 2013

Table A.5: Displacement effect for high-income countries: Income class. 2013

Dep. Var.:	Not instrumented			Instrumented		
Log of Home Empl.	(1)	(2)	(3)	(4)	(5)	(6)
Lag For. Empl. (log)	0.04	−0.05	−0.06*	0.20	0.11	0.10
	(0.05)	(0.03)	(0.03)	(0.15)	(0.17)	(0.17)
Lag Output (log)		0.43***	0.36***		0.31*	0.24
		(0.13)	(0.12)		(0.16)	(0.17)
Capital (log)			0.08			0.07
			(0.07)			(0.08)
Exports (log)			0.10			0.08
			(0.18)			(0.20)
Imports (log)			−0.03			−0.01
			(0.08)			(0.13)
Observations	553	553	553	553	553	553
First stage F-stat.				32.60	23.63	24.03
Firm & Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–3 and bootstrap std. err. for columns 4–6. Two firms have been removed from this estimation for reasons explained in footnote 12. * p<0.10, ** p<0.05, *** p<0.01.

Table A.6: Output effect for high-income countries: Income class. 2013

Dep. Var.:	Not instrumented		Instrumented	
Log of Output	(1)	(2)	(3)	(4)
Lag For. Empl. (log)	0.25***	0.17**	0.31***	0.26***
	(0.08)	(0.07)	(0.10)	(0.08)
Capital (log)		0.15*		0.13
		(0.08)		(0.08)
Exports (log)		0.31**		0.27**
		(0.12)		(0.12)
Imports (log)		−0.02		0.00
		(0.07)		(0.12)
Observations	553	553	553	553
First stage F-stat.			32.60	31.97
Firm & Year FE	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–2 and bootstrap std. err. for columns 3–4. Two firms have been removed from this estimation for reasons explained in footnote 12. * p<0.10, ** p<0.05, *** p<0.01.

A.3.3 Lower-Middle Income Countries: Drop China

Table A.7: Displacement effect for FDI into lower-middle income countries: Drop China

Dep. Var.:	Not instrumented			Instrumented		
Log of Home Empl.	(1)	(2)	(3)	(4)	(5)	(6)
Lag For. Empl. (log)	0.00	−0.03	−0.02	0.08	0.05	0.04
	(0.02)	(0.03)	(0.03)	(0.10)	(0.11)	(0.11)
Lag Output (log)		0.24***	0.21**		0.17	0.14
		(0.09)	(0.10)		(0.12)	(0.15)
Capital (log)			0.01			0.00
			(0.05)			(0.12)
Exports (log)			0.13			0.17
			(0.24)			(0.29)
Imports (log)			−0.03			−0.05
			(0.11)			(0.28)
Observations	370	370	370	370	370	370
First stage F-stat.				18.74	18.68	17.61
Firm & Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–3 and bootstrap std. err. for columns 4–6. Two firms have been removed from this estimation for reasons explained in footnote 12. * p<0.10, ** p<0.05, *** p<0.01.

Table A.8: Output effect for FDI into lower-middle income countries: Drop China

Dep. Var.:	Not instrumented		Instrumented	
Log of Output	(1)	(2)	(3)	(4)
Lag For. Empl. (log)	0.13***	0.10**	0.18**	0.14*
	(0.05)	(0.04)	(0.08)	(0.07)
Capital (log)		0.13		0.12
		(0.10)		(0.15)
Exports (log)		0.37**		0.38*
		(0.17)		(0.23)
Imports (log)		−0.01		−0.02
		(0.20)		(0.38)
Observations	370	370	370	370
First stage F-stat.			18.74	18.72
Firm & Year FE	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–2 and bootstrap std. err. for columns 3–4. Two firms have been removed from this estimation for reasons explained in footnote 12. * p<0.10, ** p<0.05, *** p<0.01.

A.3.4 Lower Middle-Income Countries: Income Classification of 2013

Table A.9: Displacement effect for lower middle-income countries: Income class. 2013

Dep. Var.:	Not instrumented			Instrumented		
Log of Home Empl.	(1)	(2)	(3)	(4)	(5)	(6)
Lag For. Empl. (log)	−0.12**	−0.16**	−0.13**	−0.03	−0.10	−0.03
	(0.06)	(0.06)	(0.06)	(0.09)	(0.09)	(0.12)
Lag Output (log)		0.31***	0.26**		0.28***	0.18
		(0.11)	(0.11)		(0.11)	(0.14)
Capital (log)			−0.05			−0.03
			(0.03)			(0.12)
Exports (log)			0.48***			0.64**
			(0.17)			(0.32)
Imports (log)			0.02			0.04
			(0.08)			(0.27)
Observations	267	267	267	267	267	267
First stage F-stat.				66.65	61.76	47.39
Firm & Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–3 and bootstrap std. err. for columns 4–6. Two firms have been removed from this estimation for reasons explained in footnote 12. * p<0.10, ** p<0.05, *** p<0.01.

Table A.10: Output effect for lower-middle income countries: Income class. 2013

Dep. Var.:	Not instrumented		Instrumented	
Log of Output	(1)	(2)	(3)	(4)
Lag For. Empl. (log)	0.16***	0.18***	0.25***	0.31***
	(0.04)	(0.05)	(0.07)	(0.06)
Capital (log)		0.14		0.14
		(0.12)		(0.21)
Exports (log)		0.77**		0.93**
		(0.34)		(0.45)
Imports (log)		0.04		0.06
		(0.15)		(0.37)
Observations	267	267	267	267
First stage F-stat.			66.65	56.63
Firm & Year FE	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–2 and bootstrap std. err. for columns 3–4. Two firms have been removed from this estimation for reasons explained in footnote 12. * p<0.10, ** p<0.05, *** p<0.01.

A.3.5 Low-Income Countries: Drop India

Table A.11: Displacement effect for FDI into low-income countries: Drop India

Dep. Var.:	Not instrumented			Instrumented		
Log of Home Empl.	(1)	(2)	(3)	(4)	(5)	(6)
Lag For. Empl. (log)	0.14*	0.12	0.07	0.26	0.09	0.08
	(0.07)	(0.08)	(0.09)	(0.70)	(0.52)	(2.63)
Lag Output (log)		0.22*	0.20		0.23	0.20
		(0.11)	(0.11)		(0.19)	(0.35)
Capital (log)			−0.04			−0.04
			(0.03)			(0.68)
Exports (log)			0.33			0.33
			(0.24)			(3.21)
Imports (log)			−0.04			−0.04
			(0.09)			(0.75)
Observations	156	156	156	156	156	156
First stage F-stat.				4.12	3.82	4.25
Firm & Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–3 and bootstrap std. err. for columns 4–6. * p<0.10, ** p<0.05, *** p<0.01.

Table A.12: Output effect for FDI into low-income countries: Drop India

Dep. Var.:	Not instrumented		Instrumented	
Log of Output	(1)	(2)	(3)	(4)
Lag For. Empl. (log)	0.12	0.10	0.78	0.73
	(0.11)	(0.12)	(5.05)	(12.47)
Capital (log)		0.10		0.19
		(0.11)		(1.60)
Exports (log)		0.67		0.13
		(0.62)		(11.63)
Imports (log)		0.07		0.24
		(0.26)		(3.75)
Observations	156	156	156	156
First stage F-stat.			4.12	4.13
Firm & Year FE	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–2 and bootstrap std. err. for columns 3–4. * p<0.10, ** p<0.05, *** p<0.01.

A.3.6 Low-Income Countries: Income Classification of 2013

Table A.13: Displacement effect for low-income countries: Income class. 2013

Dep. Var.:	Not instrumented			Instrumented		
Log of Home Empl.	(1)	(2)	(3)	(4)	(5)	(6)
Lag For. Empl. (log)	0.06	0.04	0.01	−0.01	−0.21	0.00
	(0.07)	(0.06)	(0.06)	(1.25)	(5.20)	(6.32)
Lag Output (log)		0.16	0.59*		0.31	0.59
		(0.28)	(0.32)		(4.98)	(6.91)
Capital (log)			−0.34*			−0.35
			(0.17)			(3.02)
Exports (log)			0.46			0.48
			(0.34)			(6.68)
Imports (log)			−0.34			−0.35
			(0.53)			(4.72)
Observations	100	100	100	100	100	100
First stage F-stat.				7.29	14.55	11.54
Firm & Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–3 and bootstrap std. err. for columns 4–6. * p<0.10, ** p<0.05, *** p<0.01.

Table A.14: Output effect for low-income countries: Income class. 2013

Dep. Var.:	Not instrumented		Instrumented	
Log of Output	(1)	(2)	(3)	(4)
Lag For. Empl. (log)	0.08	0.05	0.63	0.04
	(0.06)	(0.03)	(2.29)	(0.23)
Capital (log)		0.40***		0.40**
		(0.08)		(0.19)
Exports (log)		−0.56***		−0.54
		(0.11)		(0.81)
Imports (log)		1.34***		1.35**
		(0.38)		(0.69)
Observations	100	100	100	100
First stage F-stat.			7.29	9.65
Firm & Year FE	Yes	Yes	Yes	Yes

Notes: Clustered std. err. in parentheses for columns 1–2 and bootstrap std. err. for columns 3–4. * p<0.10, ** p<0.05, *** p<0.01.

A.4 List of Countries and Income Classification

Table A.15: Income Classification

Country Names	2002	2013
Antigua and Barbuda	H	H
Aruba	H	H
Australia	H	H
Austria	H	H
Bahamas	H	H
Bahrain	H	H
Barbados	H	H
Belgium	H	H
Bermuda	H	H
Brunei Darussalam	H	H
Canada	H	H
Cayman Islands	H	H
Cyprus	H	H
Denmark	H	H
Finland	H	H
France	H	H
Germany	H	H
Greece	H	H
Hong Kong, China	H	H
Iceland	H	H
Ireland	H	H
Isle of Man	H	H
Israel	H	H
Italy	H	H
Japan	H	H
Korea, Rep.	H	H
Kuwait	H	H
Luxembourg	H	H
Macao, China	H	H
Malta	H	H
Netherlands	H	H
New Zealand	H	H
Norway	H	H
Portugal	H	H
Qatar	H	H
Singapore	H	H
Slovenia	H	H
Spain	H	H
Sweden	H	H
Taiwan, China	H	H
United Arab Emirates	H	H
United Kingdom	H	H
United States	H	H
Virgin Islands (U.S.)	H	H
Argentina	UM	UM
Belize	UM	UM
Botswana	UM	UM
Chile	UM	H
Costa Rica	UM	UM
Croatia	UM	H
Czech Republic	UM	H
Dominica	UM	UM
Estonia	UM	H
Gabon	UM	UM
Grenada	UM	UM
Hungary	UM	UM
Latvia	UM	H
Lebanon	UM	UM
Libya	UM	UM
Lithuania	UM	H
Malaysia	UM	UM
Mauritius	UM	UM
Mexico	UM	UM
Oman	UM	H
Panama	UM	UM
Poland	UM	H
Saudi Arabia	UM	H
Slovak Republic	UM	H
St. Kitts and Nevis	UM	H
St. Lucia	UM	UM
Trinidad and Tobago	UM	H
Uruguay	UM	H
Venezuela	UM	UM
Albania	LM	UM
Algeria	LM	UM
Armenia	LM	LM
Belarus	LM	UM
Bolivia	LM	LM
Bosnia and Herzegovina	LM	UM
Brazil	LM	UM
Bulgaria	LM	UM
China	LM	UM
Colombia	LM	UM
Cuba	LM	UM
Dominican Republic	LM	UM
Ecuador	LM	UM

Cont. Table A.15: Income Classification

Country Names	2002	2013
Egypt	LM	LM
El Salvador	LM	LM
Guatemala	LM	LM
Honduras	LM	LM
Iran	LM	UM
Iraq	LM	UM
Jamaica	LM	UM
Jordan	LM	UM
Kazakhstan	LM	UM
Macedonia	LM	UM
Morocco	LM	LM
Namibia	LM	UM
Paraguay	LM	LM
Peru	LM	UM
Philippines	LM	LM
Romania	LM	UM
Russia	LM	H
South Africa	LM	UM
Sri Lanka	LM	LM
Syria	LM	LM
Thailand	LM	UM
Tunisia	LM	UM
Turkey	LM	UM
Turkmenistan	LM	UM
Ukraine	LM	LM
West Bank and Gaza	LM	LM
Afghanistan	L	L
Angola	L	UM
Azerbaijan	L	UM
Bangladesh	L	L
Benin	L	L
Burkina Faso	L	L
Burundi	L	L
Cambodia	L	L
Cameroon	L	LM
Central African Republic	L	L
Chad	L	L
Congo, DR	L	L
Congo, Rep.	L	LM
Côte d'Ivoire	L	LM
Ethiopia	L	L
Georgia	L	LM
Ghana	L	LM
Guinea	L	L
Haiti	L	L
India	L	LM
Indonesia	L	LM
Kenya	L	L
Kyrgyz Republic	L	LM
Lao	L	LM
Liberia	L	L
Madagascar	L	L
Malawi	L	L
Mali	L	L
Mauritania	L	LM
Moldova	L	LM
Mongolia	L	LM
Mozambique	L	L
Myanmar	L	L
Nepal	L	L
Nicaragua	L	LM
Niger	L	L
Nigeria	L	LM
Pakistan	L	LM
Papua New Guinea	L	LM
Rwanda	L	L
Senegal	L	LM
Sierra Leone	L	L
Sudan	L	LM
Tajikistan	L	L
Tanzania	L	L
Timor-Leste	L	LM
Togo	L	L
Uganda	L	L
Uzbekistan	L	LM
Vietnam	L	LM
Yemen	L	LM
Zambia	L	LM
Zimbabwe	L	L

Notes:
H: High-income;
UM: Upper-middle income;
LM: Lower-middle income;
L: Low income

Bibliography

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105(490):493–505.

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59(2):495–510.

Abadie, Alberto and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93(1):113–132.

Aber, Lawrence, Catalina Torrente, Leighann Starkey, Brian Johnston, Edward Seidman, Peter Halpin, Anjuli Shivshanker, Nina Weisenhorn, Jeannie Annan and Sharon Wolf. 2017. "Impacts after One Year of "Healing Classroom" on Children's Reading and Math Skills in DRC: Results from a Cluster Randomized Trial." *Journal of Research on Educational Effectiveness* 10(3):507–529.

Acemoglu, Daron, Simon Johnson, James Kwak and Todd Mitton. 2016. "The Value of Connections in Turbulent Times: Evidence from the United States." *Journal of Financial Economics* 121(2):368–391.

Albornoz, Facundo, Maria Anauati, Melina Furman, Mariana Luzuriaga, Maria Podesta and Ines Taylor. 2018. "Training to Teach Science: Experimental Evidence from Argentina." World Bank Policy Research Working Paper No. 8589. URL: `www.worldbank.org/` (last access: 10.03.2020).

Angrist, Joshua and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics. An Empiricist'sCompanion.* Princeton: Princeton University Press.

Angrist, Joshua and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." *Journal of Economic Perspective* 24(2):3–30.

Araujo, Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *Quarterly Journal of Economics* 131(3):1415–1453.

Attanasio, Orazio, Camila Fernández, Emla Fitzsimons, Sally Grantham-McGregor, Costas Meghir and Marta Rubio-Codina. 2014. "Using the Infrastructure of a Conditional Cash Transfer Program to Deliver a Scalable Integrated Early Child Development Program in Colombia: Cluster Randomized Controlled Trial." *BMJ* 349:1–12.

Autor, David. 2003. "Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing." *Journal of Labour Economics* 21(1):1–42.

Badarinza, Cristian and Tarun Ramadorai. 2018. "Home away from Home? Foreign Demand and London House Prices." *Journal of Financial Economics* 130:532–555.

Baird, Sarah, Aislinn Bohren, Craig McIntosh and Berk Özler. 2015. "Designing Experiments to Measure Spillover Effects." PIER Working Paper No. 15-021. URL: `www.ssrn.com/` (last access: 10.03.2020).

Bando, Rosangela and Xia Li. 2014. "The Effect of In-Service Teacher Training on Student Learning of English as a Second Language." IDB Working Paper Series No. 529. URL: `https://www.iadb.org/en` (last access: 20.06.2020).

Banerjee, Abhijit and Esther Duflo. 2011. *Poor Economics*. London: Penguin Books.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland and Michael Walton. 2016. "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India." NBER Working Paper No. 22746. URL: `https://www.nber.org/papers` (last access: 24.06.2020).

Banerjee, Abhijit, Shawn Cole, Esther Duflo and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122(3):1235–1264.

Barba Navaretti, Giorgio, Davide Castellani and Anne-Célia Disdier. 2010. "How Does Investing in Cheap Labour Countries Affect Performance at Home? Firm-Level Evidence from France and Italy." *Oxford Economic Papers* 62(2):234–260.

Barba Navaretti, Giorgo, Anthony Venables and Frank Barry. 2006. *Mulitnational Firms in the World Economy*. Princeton: Princeton University Press.

Barber, Michael and Mona Mourshed. 2007. "How the World's Best-Performing Schools Systems Come out on Top." Mc Kinsey Report. URL: `https://www.mckinsey.com/industries/social-sector/our-insights/how-the-worlds-best-performing-school-systems-come-out-on-top` (last access: 10.03.2020).

Bassi, Marina, Costas Meghir and Ana Reynoso. 2016. "Education Quality and Teaching Practices." NBER Working Paper, No. 22719. URL: `https://www.nber.org/papers/` (last access: 23.06.2020).

Bau, Natalie and Jishnu Das. 2017. "The Misallocation of Pay and Productivity in the Public Sector: Evidence from the Labor Market for Teachers." World Bank Policy Research Working Paper No. 8050. URL: `www.worldbank.org/` (last access: 10.03.2020).

Bau, Natalie and Jishnu Das. 2020. "Teacher Value-Added in a Low-Income Country." *American Economic Journal: Economic Policy* 12(1):62–96.

Baumert, Jürgen and Mareike Kunter. 2013. The COACTIVE Model of Teachers' Professional Competence. In *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers*, ed. Mareike Kunter et al. New York: Springer pp. 25–48.

Berlinski, Samuel and Matias Busso. 2013. "Challenges in Educational Reform: An Experiment on Active Learning in Mathematics." *Economic Letters* 156:172–175.

Besley, Timothy and Robin Burgess. 2004. "Can Labor Regulation Hinder Economic Performance? Evidence from India." *Quarterly Journal of Economics* 119(1):91–134.

Beuermann, Diether, Emma Naslund-Hadley, Inder Ruprah and Jennelle Thompson. 2013. "The Pedagogy of Science and Environment: Experimental Evidence from Peru." *Journal of Development Studies* 49(5):719–736.

Bietenbeck, Jan, Marc Piopiunik and Simon Wiederhold. 2018. "Africa's Skill Tragedy: Does Teachers' Lack of Knowledge Lead to Low Student Performance?" *Journal of Human Resources* 53(33):553–578.

Billmeier, Andreas and Tommaso Nannicini. 2013. "Assesing Economic Liberalization Episodes: A Synthetic Control Approach." *Review of Economics and Statistics* 95(3):983–1001.

Blimpo, Moussa, Pedro Carneiro, Pamela Jervis and Todd Pugatch. 2019. "Improving Access and Quality in Early Childhood Development Programs: Experimental Evidence from the Gambia." World Bank Policy Research Working Paper No. 8737. URL: `www.worldbank.org/` (last access: 20.06.2020).

Bold, Tessa, Deon Filmer, Ezequiel Molina and Jakob Svensson. 2019. "The Lost Human Capital: Teacher Knowledge and Student Achievement in Africa." World Bank Policy Research Working Paper No. 8849. URL: `www.worldbank.org/` (last access: 10.03.2020).

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson and Waly Wane. 2017a. "Enrollment without Learning: Teacher Effort, Knowledge and Skill in Primary Schools in Africa." *Journal of Economic Perspectives* 31(4):185–204.

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Christophe Rockmore, Brian Stacy, Jakob Svensson and Waly Wane. 2017b. "What Do Teachers Know and Do? Does It Matter? Evidence from Primary Schools in Africa." World Bank Policy Research Working Paper No. 7956. URL: `www.worldbank.org/` (last access: 10.03.2020).

Brunetti, Aymo, Konstantin Büchel, Martina Jakob, Ben Jann, Christoph Kühnhanss and Daniel Steffen. 2020. "Teacher Content Knowledge in Developing Countries: Evidence from a Math Assessment in El Salvador." Working Paper No. 2005, Department of Economics, University of Bern. URL: `www.vwi.unibe.ch` (last access: 10.03.2020).

Bruns, Barbara and Javier Luque. 2014. *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean.* Washington D.C.: The World Bank.

Carr, David, James Markusen and Keith Maskus. 2001. "Estimating the Knowledge-Capital Model of the Multinational Enterprise." *American Economic Review* 91(3):693–708.

Carrillo, Paul, Mercedes Onofa and Juan Ponce. 2011. "Information Technology and Student Achievement: Evidence from a Randomized Experiment in Ecuador." IDB Working Paper Series No. 223. URL: `www.econstor.eu` (last access: 10.03.2020).

Cavallo, Eduardo, Sebastian Galiani, Ilan Noy and Juan Pantano. 2013. "Catastrophic Natural Disasters and Economic Growth." *Review of Economics and Statistics* 95(5):1549–1561.

Chao, Chi-Chur and Eden Yu. 2014. "Housing Markets with Foreign Buyers." *Journal of Real Estate Finance and Economics* 50:207–218.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan and Halsey Rogers. 2006. "Missing in Action: Teacher and Health Worker Absence in Developing Countries." *Journal of Economic Perspectives* 20(1):91–116.

Chen, Christopher, Nitish Jain and Alex Yang. 2019. "The Impact of Trade Credit Provision on Retail Inventory: An Empirical Investigation Using Synthetic Controls." Working Paper, URL: `www.ssrn.com/index.cfm/en/`.

Chetty, Raj, John Friedman and Jonah Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104(9):2633–2679.

Chinco, Alex and Christopher Mayer. 2015. "Misinformed Speculators and Mispricing in the Housing Market." *Review of Financial Studies* 29(2):486–522.

Cilliers, Jacobus, Brahm Fleisch, Cas Prinsloo and Stephen Taylor. 2019. "How to Improve Teaching Practice? An Experimental Comparison of Centralized Training and In-Classroom Coaching." *Journal of Human Resources* 66:203–213.

Cilliers, Jacobus, Brahm Fleisch, Janeli Kotze, Mpumi Mohohlwane and Stephen Taylor. 2020. "The Challenge of Sustaining Effective Teaching: Spillovers, Fade-Out, and the Cost-Effectiveness of Teacher Development Programs." *Unpublished manuscript.*

Codoni, Davide and Ueli Grob. 2013. "Auswirkungen der Zweitwohnungsinitiative auf den Tourismus im Schweizer Alpenraum." *Die Volkswirtschaft* 4:17–20.

Conn, Katharine. 2017. "Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations." *Review of Educational Research* 87(5):863–898.

Crinò, Rosario. 2009. "Offshoring, Multinationals and Labour Market: A Review of the Empirical Literature." *Journal of Economic Surveys* 23(2):197–249.

Crinò, Rosario. 2010. "Service Offshoring and White-Collar Employment." *Review of Economic Studies* 77(2):595–632.

Cvijanovic, Dragana and Christophe Spaenjers. 2015. "Non-Resident Demand and Property Prices in Paris." HEC Paris Research Paper No. 2015.

Davies, Ronald. 2008. "Hunting High and Low for Vertical FDI." *Review of International Economics* 16(2):250–267.

de Ayala, R.J. 2009. *The Theory and Practice of Item Response Theory.* New York: Guilford Press.

Debaere, Peter, Hongshik Lee and Joonhyung Lee. 2010. "It Matters Where You Go: Outward Foreign Direct Investment and Multinational Employment Growth at Home." *Journal of Development Economics* 91(2):301–309.

Desai, Mihir, Fritz Foley and James Hines. 2009. "Domestic Effects of the Foreign Activities of US Multinationals." *American Economic Journal: Economic Policy* 1(1):181–203.

Dhaliwal, Iqbal, Esther Duflo, Rachel Glennerster and Caitlin Tulloch. 2014. Comparative Cost-effectiveness Analysis to Inform Policy in Developing Countries: a General Framework with Applications for Education. In *Education Policy in Developing Countries*, ed. Paul Glewwe. Chicago and London: University of Chicago Press pp. 285–338.

di Giovanni, Julian and Andrei Levchenko. 2009. "Trade Openness and Volatility." *Review of Economics and Statistics* 91(3):558–585.

Doudchenko, Nikolay and Guido Imbens. 2016. "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis." NBER Working Paper No. 22791, URL: `www.nber.org/papers/` (last access: 02.07.2020).

Digestyc, Direccion General de Estadistica y Censos El Salvador. 2018. "Encuesta de Hogares de Direccion General de Estadistica y Censos 2017 (EHPM)." Online avaliable, URL: `www.digestyc.gob.sv` (last access: 25.07.2018).

Du, Zaicho and Lin Zhang. 2015. "Home-Purchase Restriction, Property Tax and Housing Price in China: a Counterfactual Analysis." *Journal of Econometrics* 188(2):558–568.

Egger, Peter and Michael Pfaffermayr. 2004. "Distance, Trade and FDI: a Hausman–Taylor SUR Approach." *Journal of Applied Econometrics* 19(2):227–246.

Escueta, Maya, Andre Nickow, Philip Oreopoulos and Vincent Quant. 2020. "Upgrading Education with Technology: Insights from Experimental Research." *Journal of Economic Literature.* forthcoming.

Favilukis, Jack and Stijn Van Nieuwerburgh. 2017. "Out-of-Town Home Buyers and City Welfare." CEPR Discussion Paper No. 12283, URL: `www.cepr.org` (last access: 19.09.2019).

Fleisch, Brahm, Volker Schöer, Gareth Roberts and Amy Thornton. 2016. "System-Wide Improvement of Early-Grade Mathematics: New Evidence from the Gauteng Primary Language and Mathematics Strategy." *International Journal of Educational Development* 49:157–174.

Ganimian, Alejandro and Richard Murnane. 2016. "Improving Education in Developing Countries: Lessons From Rigorous Impact Evaluations." *Review of Educational Research* 86(3):719–755.

Glennerster, Rachel and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide.* Princeton: Princeton University Press.

Glewwe, Paul and Karthik Muralidharan. 2016. Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps and Policy Implications. In *Handbook of the Economics of Education*, ed. Eric Hanushek, Stephen Machin and Ludger Woessmann. Amsterdam: Elsevier pp. 653–743.

Glewwe, Paul, Michael Kremer and Sylvie Moulin. 2009. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics* 1(1):112–135.

Grossman, Gene and Esteban Rossi-Hansberg. 2008. "Trading Tasks: A Simple Theory of Offshoring." *American Economic Review* 98(5):1978–1997.

Hanushek, Eric. 1992. "The Trade-off Between Child Quantity and Quality." *Journal of Political Economy* 100(1):84–117.

Hanushek, Eric. 2011. "The Economic Value of Higher Teacher Quality." *Economics of Education Review* 30(3):466–479.

Harrison, Ann and Margaret McMillan. 2011. "Offshoring Jobs? Multinationals and U.S. Manufacturing Employment." *Review of Economics and Statistics* 93(3):857–875.

He, Fand, Leigh Linden and Margaret MacLeod. 2009. "A Better Way to Teach Children to Read? Evidence from a Randomized Controlled Trial." Working Paper, Columbia Unversity.

Head, Keith, Thierry Mayer and John Ries. 2009. "How Remote Is the Offshoring Threat?" *European Economic Review* 53(4):429–444.

Helpman, Elhanan. 1984. "A Simple Theory of International Trade with Multinational Corporations." *Journal of Political Economy* 92(3):451–471.

Helpman, Elhanan, Marc Melitz and Stephen Yeaple. 2004. "Export Versus FDI with Heterogeneous Firms." *American Economic Review* 94(1):300–316.

Helpman, Elhanan and Paul Krugman. 1985. *Market Structure and Foreign Trade: Increasing Returns, Imperfect Competition, and the International Economy.* Cambridge: MIT Press.

Hess, Simon. 2017. "Randomization Inference with Stata: A Guide and Software." *Stata Journal* 17(3):630–651.

Hijzen, Alexander, Sébastien Jean and Thierry Mayer. 2011. "The Effects at Home of Initiating Production Abroad: Evidence from Matched French Firms." *Review of World Economics* 147(3):457–483.

Hilber, Christian and Olivier Schöni. 2020. "On the Economic Impacts of Constraining Second Home Investments." *Journal of Urban Economics* 118:103266.

Hummels, David, Jakob Munch and Chong Xiang. 2018. "Offshoring and Labor Markets." *Journal of Economic Literature* 56(3):981–1028.

Hummels, David, Rasmus Jørgensen, Jakob Munch and Chong Xiang. 2014. "The Wage Effects of Offshoring: Evidence from Danish Matched Worker-Firm Data." *American Economic Review* 104(6):1597–1629.

Imbens, Guido and Jeffrey Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47(1):5–86.

Johnson, Angela, Catherine Galloway, Elliott Friedlander and Claude Goldenberg. 2019. "Advancing Educational Quality in Rwanda: Improving Teachers' Literacy Pedagogy and Print Environments." *International Journal of Educational Research* 98:134–145.

Jukes, Matthew, Elizabeth Turner, Margaret Dubeck, Katherine Halliday, Hellen Inyega, Sharon Wolf, Stephanie Simmons Zuilkowski and Simon Brooker. 2016. "Improving Literacy Instruction in Kenya Through Teacher Professional Development and Text Messages Support: A Cluster Randomized Trial." *Journal of Research on Educational Effectiveness* 10(3):449–481.

Kanfer, Ruth and Phillipp Ackerman. 2004. "Aging, Adult Development, and Work Motivation." *Academy of Management Review* 29(3):440–458.

Kaufmann, Philippe and Thomas Rieder. 2012. "Zweitwohnungsstopp: Mögliche Auswirkungen auf die Immobilienpreise in den Tourismusregionen." *Die Volkswirtschaft* 6:63–66.

Kaul, Ashok, Stefan Klössner, Gregor Pfeifer and Manuel Schieler. 2017. "Synthetic Control Methods: Never Use All Pre-Intervention Outcomes as Economic Predictors." MPRA Working Paper No. 83790, URL: `https://mpra.ub.uni-muenchen.de/` (last access: 19.09.2019).

Kerwin, Jason and Rebecca Thornton. 2020. "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures." *Review of Economics and Statistics.* forthcoming.

Kraft, Matthew, David Blazar and Dylan Hogan. 2018. "The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence." *Review of Educational Research* 88(4):547–588.

Kreif, Noémi, Richard Grieve, Dominik Hangartner, Alex James Turner, Silviya Nikolova and Matt Sutton. 2015. "Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units." *Health Economics* 25:1514–1528.

Kremer, Michael, Conner Brannen and Rachel Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science* 340(6130):297–300.

Lai, Fang, Renfu Luo, Lixiu Zhang, Xinzhe Huang and Scott Rozelle. 2015. "Does Computer-Assisted Learning Improve Learning Outcomes? Evidence from a Randomized Experiment in Migrant Schools in Beijing." *Ecomomics of Education Review* 47(1):34–48.

Leamer, Edward. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73(1):31–43.

Leme, Maria, Paula Louzanoc, Vladimir Ponczek and André Portela Souzaa. 2012. "The Impact of Structured Teaching Methods on the Quality of Education in Brazil." *Economics of Education Review* 31(5):850–860.

Linden, Leigh. 2008. "Complement or Substitute? The Effect of Technology on Student Achievement in India." infoDev Working Paper No. 17. URL: `www.worldbank.org/` (last access: 10.03.2020).

Lipsey, Robert. 2004. Home- and Host-Country Effects of Foreign Direct Investment. In *Challenges to Globalization: Analyzing the Economics*, ed. Robert Baldwin and Alan Winters. Chicago: University of Chicago Press pp. 333–382.

Lucas, Adrienne, Patrick McEwan, Moses Ngware and Moses Oketch. 2014. "Improving Early-grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda." *Journal of Policy Analysis and Management* 33(4):950–976.

Markusen, James and Anthony Venables. 2000. "The Theory of Endowment, Intra-Industry and Multinational Trade." *Journal of International Economics* 52(2):209–234.

Markusen, James and Keith Maskus. 2003. General-Equilibrium Approaches to the Multinational Enterprise: A Review of Theory and Evidence. In *Handbook of International Trade*, ed. Kwan Choi and James Harrigan. Malden: Blackwell Publishing pp. 320–352.

Markusen, James R. 1984. "Mulitnationals, Mutli-Plant Economies and the Gains from Trade." *Journal of International Economics* 16(3–4):205–226.

Mbiti, Isaac. 2016. "The Need of Accountability in Education in Developing Countries." *Journal of Economic Perspectives* 30(3):109–132.

Mbiti, Isaac, Karthik Muralidharan, Maruicio Romero, Youdi Schipper, Constantine Manda and Rakesh Rajani. 2019. "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania." *Quarterly Journal of Economics* 134(3):1627–1673.

McEwan, Patrick. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." *Review of Educational Research* 85(3):353–394.

Metzler, Johannes and Ludger Woessmann. 2012. "The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation." *Journal of Development Economics* 99(2):486–496.

Miguel, Edward and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72(1):159–217.

MINED, Ministerio de la Educación Ciencia y Tecnología de El Salvador. 2019. "Informe de Resultados: PAES 2019." Online avaliable, URL: `https://www.mined.gob.sv` (last access: 17.06.2020).

MINED, Ministerio de la Educacion de El Salvador. 2013. "Elementos para el Desarrollo del Modelo Pedagogico del Sistema Educativo Nacional – Escuela Inclusiva de Tiempo Pleno." Online avaliable, URL: `https://www.mined.gob.sv/jdownloads/Institucional/modelopedagogico.pdf` (last access: 14.01.2018).

Mo, Di, Linxiu Zhang, Jiafu Wang, Weiming Huang, Yao Shi, Matthew Boswell and Scott Rozelle. 2015. "Persistence of Learning Gains from Computer Assisted Learning: Experimental Evidence from China." *Journal of Computer Assisted Learning* 31:562–581.

Molina, Ezequiel, Adelle Pushparatnam, Sara Rimm-Kaufman and Keri Ka-Yee Wong. 2018. "Evidence-Based Teaching. Effective Teaching Practices in Primary School Classrooms." World Bank Policy Research Working Paper No. 8856. URL: `www.worldbank.org/` (last access: 10.03.2020).

Morgan, Kari and Donald Rubin. 2012. "Rerandomization to Improve Covariate Balance in Experiments." *Annals of Statistics* 40(2):1263–1282.

Mouton, Johann. 1995. "Second Language Teaching for Primary School Students: An Evaluation of a New Teaching Method." *Evaluation and Program Planning* 18(4):391–408.

Muralidharan, Karthik. 2017. Field Experiments in Education in Developing Countries. In *Handbook of Economic Field Experiments*, ed. Abhijit Banerjee and Esther Duflo. Amsterdam: Elsevier pp. 323–385.

Muralidharan, Karthik, Abhijeet Singh and Alejandro Ganimian. 2019. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." *American Economic Review* 109(4):1426–1460.

Ng, Thomas and Daniel Feldman. 2008. "The Relationship of Age to Ten Dimensions of Job Performance." *Journal of Applied Psychology* 93(2):392–423.

Nonoyama-Tarum, Yuko and Kurt Bredenberg. 2009. "Impact of School Readiness Program Interventions on Children's Learning in Cambodia." *International Journal of Educational Development* 29(1):39–45.

Ottaviano, Gianmarco, Giovanni Peri and Greg Wright. 2013. "Immigration, Offshoring, and American Jobs." *American Economic Review* 103(5):1925–1959.

Özler, Berk, Lia Fernald, Patricia Kariger, Christin McConnell, Michelle Neuman and Eduardo Fraga. 2016. "Combining Preschool Teacher Training with Parenting Education." World Bank Policy Research Working Paper No. 7817. URL: `www.worldbank.org/` (last access: 10.03.2020).

Pallante, Daniel and Young-Suk Kim. 2013. "The Effect of a Multicomponent Literacy Instruction Model on Literacy Growth for Kindergartners and First-Grade Students in Chile." *International Journal of Psychology* 48(5):747–761.

Piper, Benjamin and Medina Korda. 2011. "EGRA Plus: Liberia." Program Evaluation Report, RTI International. URL: `https://eric.ed.gov/` (last access: 20.06.2020).

Piper, Benjamin, Stephanie Simmons Zuilkowski, Margaret Dubeck, Evelyn Jepkemei and Simon King. 2018. "Identifying the Essential Ingredients to Literacy and Numeracy Improvement: Teacher Professional Development and Coaching, Student Textbooks, and Structured Teachers' Guides." *World Development* 106:324–336.

Pomeranz, Dina. 2019. "Empirische Revolution in der Verwaltung." *Die Volkswirtschaft* 10:4–7.

Popova, Anna, David Evans, Mary Breeding and Violeta Arancibia. 2018. "Teacher Professional Development around the World: The Gap between Evidence and Practice." World Bank Policy Research Working Paper No. 8572. URL: `www.worldbank.org/` (last access: 10.03.2020).

Pournara, Craig, Jeremy Hodgen, Jill Adler and Vasen Pillay. 2015. "Can Improving Teachers' Knowledge of Mathematics Lead to Gains in Learners' Attainment in Mathematics?" *South African Journal of Education* 35(3):1–10.

Pritchett, Lant and Amanda Beatty. 2014. "Slow Down, You're Going too Fast: Matching Curricula to Student Skill Levels." *International Journal of Educational Development* 40:276–288.

Rockoff, Jonah. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review Papers and Proceedings* 94(2):247–252.

Sá, Filipa. 2016. "The Effect of Foreign Investors on Local Housing Markets: Evidence from the UK." CEPR Discussion Paper No. 11658. URL: `www.ssrn.com/index.cfm/en/` (last access: 19.09.2019).

Sacerdote, Bruce. 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*, ed. Eric Hanushek, Stephen Machin and Ludger Woessmann. Amsterdam: Elsevier pp. 249–277.

Sailors, Misty, James Hoffman, David Pearson, Nicola McClung, Jaran Shin, Liveness Phiri and Tionge Saka. 2014. "Supporting Change in Literacy Instruction in Malawi." *Reading Research Quarterly* 49(2):209–231.

San Antonio, Diosdado, Nelson Morales and Leo Moral. 2011. "Module-Based Professional Development for Teachers: a Cost-Effective Philippine Experiment." *Teacher Development* 15(2):157–169.

Santos Silva, João and Silvana Tenreyro. 2006. "The Log of Gravity." *Review of Economics and Statistics* 88(4):641–658.

Schaie, Warner and Sherry Willis. 2013. "The Seattle Longitudinal Study of Adult Cognitive Development." *ISSBD Bulletin* 57(1):24–29.

Sethupathy, Guru. 2013. "Offshoring, Wages, and Employment: Theory and Evidence." *European Economic Review* 62:73–97.

Sinha, Shabnam, Rukmini Banerji and Wilima Wadhwa. 2016. *Teacher Performance in Bihar, India: Implications for Education*. Washington D.C.: The World Bank.

Snilstveit, Birte, Jennifer Stevenson, Daniel Phillips, Martina Vojtkova, Emma Gallagher, Tanja Schmidt, Hannah Jobse, Maisie Geelen, Maria Pastorello and John Eyers. 2015. "Interventions for Improving Learning Outcomes and Access to Education in Low- and Middle- Income Countries: A Systematic Review." 3ie Systematic Review 24. Online available, URL: `https://www.3ieimpact.org/evidence-hub/publications/systematic-reviews/`.

Somerville, Tsur, Long Wang and Yang Yang. 2020. "Using Purchase Restrictions to Cool Housing Markets: A Within-Market Analysis." *Journal of Urban Economics* 115:103189.

Spratt, Jennifer, Simon King and Jennae Bulat. 2013. "Independent Evaluation of the Effectiveness of Institut pour l'Education Populaire's "Read-Learn-Lead" (RLL) Program in Mali." Program Evaluation Report, RTI International. URL: `https://www.rti.org/` (last access: 20.06.2020).

Tan, Jee-Peng, Julia Lane and Gerard Lassibille. 1999. "Student Outcomes in Philippine Elementary Schools: An Evaluation of Four Experiments." *World Bank Review* 13(3):493–508.

Tang, Jitao and Rosanne Altshuler. 2015. "The Spillover Effects of Outward Foreign Direct Investment on Home Countries: Evidence from the United States." Working Paper 1503, Oxford University Centre for Business Taxation. URL: `www.sbs.ox.ac.uk/` (last access: 15.03.2020).

Unesco, United Nations Educational, Scientific and Cultural Organization. 2019. "UNESCO Institute for Statistics Database." Online available, URL: `http://data.uis.unesco.org/` (last access: 04.12.2019).

The Economist. 2017. "Technology is Transforming What Happens When a Child Goes to School." published in *Briefing* section of the print edition under headline "Machine Learning", July 22nd 2017.

Wing, Coady, Kosali Simon and Ricardo Bello-Gomez. 2018. "Designing Difference in Difference Studies: Best Practices for Public Health Policy Research." *Annual Review of Public Health* 39:453–469.

Wolf, Sharon. 2019. "Year 3 Follow-Up of the 'Quality Preschool for Ghana' Interventions on Child Development." *Developmental Psychology* 55(12):2587–2602.

Wolf, Sharon, Lawrence Aber, Jere Behrman and Edward Tsinigo. 2019. "Experimental Impacts of the "Quality Preschool for Ghana" Interventions on Teacher Professional Well-being, Classroom Quality, and Children's School Readiness." *Journal of Research on Educational Effectiveness* 12(1):10–37.

World Bank. 2018. *World Development Report 2018: Learning to Realize Education's Promise*. Washington D.C.: World Bank.

Wright, Greg. 2013. "Revisiting the Employment Impact of Offshoring." *European Economic Review* 66:63–83.

Yan, Yan and Hongbind Ouyang. 2018. "Effects of House-Sale Restrictions in China: a Difference-in-Difference Approach." *Applied Economics Letters* 25(15):1051–1057.

Yang, Yihua, Linxiu Zhang, Junxia Zeng, Xiaopeng Pang, Fang Lai and Scott Rozelle. 2013. "Computers and the Academic Performance of Elementary School-Aged Girls in China's Poor Communities." *Computers & Education* 60(1):335–346.

Yoshikawa, Hirokazu, Diana Leyva, Catherine Snow, Ernesto Treviño, Clara Barata, Christina Weiland, Celia Gomez, Lorenzo Moreno, Andrea Rolla, Nikhit D'Sa and Mary Arbour. 2015. "Experimental Impacts of Teacher Professional Development Program in Chile on Preschool Classroom Quality and Child Outcomes." *Developmental Psychology* 51(3):309–322.

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *Quarterly Journal of Economics* 134(2):557–598.

Zhang, Linxiu, Fang Lai, Xiaopeng Pang, Hongmei Yi and Scott Rozelle. 2013. "The Impact of Teacher Training on Teacher and Student Outcomes: Evidence from a Randomised Experiment in Beijing Migrant Schools." *Journal of Development Effectiveness* 5(3):339–358.

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Koautorenschaften sowie alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Artikel 36 Absatz 1 Buchstabe o des Gesetzes vom 5. September 1996 über die Universität zum Entzug des aufgrund dieser Arbeit verliehenen Titels berechtigt ist.

Bern, 31. August 2020

Daniel Steffen