Universität Bern

Using Field Experiments and Machine Learning to Bridge the Global Learning Gap

Martina Saskia Jakob

Inauguraldissertation zur Erlangung der Würde eines
DOCTOR RERUM SOCIALIUM

der Wirtschafts- und Sozialwissenschaftlichen Fakultät

Promotionsdatum: 14. Dezember 2023

The faculty accepted this thesis on 14/12/2023 at the request of the reviewers Prof. Dr. Mauricio Romero and Prof. Dr. Ben Jann as dissertation, without wishing to comment on the views expressed therein.

Die Fakultät hat diese Arbeit am 14/12/2023 auf Antrag der Gutachter Prof. Dr. Mauricio Romero und Prof. Dr. Ben Jann als Dissertation angenommen, ohne damit zu den darin ausgesprochenen Auffassungen Stellung nehmen zu wollen.

Abstract

This dissertation combines field experiments and machine learning to study education in low- and middle-income countries. The first chapter develops a novel approach to measuring educational attainment from social media data. The second and third chapters evaluate teacher training interventions in El Salvador and Tanzania. The fourth chapter compares bottom-up and top-down approaches to public goods provision in rural El Salvador using deep learning for outcome measurement.

Dissertation

Using Field Experiments and Machine Learning to Bridge the Global Learning Gap

A thesis submitted to attain the degree of

DOCTOR OF SOCIOLOGY
University of Bern

presented by

MARTINA JAKOB
12-124-426

M.A. in Sociology, University of Bern
B.A. in Social Sciences, University of Bern

Supervised by

Prof. Dr. Ben Jann, examiner

Prof. Dr. Mauricio Romero, co-examiner

2023

Acknowledgments

I have the pleasure to look back on a PhD that was only partly defined by a solitary struggle with my computer, but mostly by fruitful collaborations in great teams. In this regard, I am indebted to numerous people whose invaluable contributions made the projects presented in this dissertation possible. First of all, I would like to thank the superb project teams in El Salvador and Tanzania. Compared to your tireless work on the ground, Stephanie, Aracely, Arely, Alonzo, Jonathan, Isabel, Carlos, Chendo, Donatian, Judith, Anna, Staphord, Peter, Diana, and Steven, my role in analyzing the data and writing the papers was a minor contribution.

Throughout my dissertation, I enjoyed the unconditional support of my supervisors and mentors. I want to thank Ben, Mauricio, Adina and Ted for their outstanding guidance on this journey. Likewise, I extend my gratitude to my co-authors – Carla, Konstantin, Daniel and Aymo – for the productive collaborations on the ambitious projects we have shouldered together. In this vein, I am also grateful for the excellent research assistance provided by Malin, Cedric, Elena, Miriam, and Nadja. As a team, we accomplished what would have been unattainable alone.

Over the past few years, I had the pleasure of being part of the stimulating research environments at the University of Bern and the University of California at Berkeley. I want to thank my colleagues in Bern – Benita, Christoph, Barbara, Joël, Lukas, Rudi, Sebastian, and Krzys, among others – for all the insightful and motivating discussions over lunch and around the coffee machine. I am also grateful to my friends from California – David, Ingrid, August, Fanny, Dario, and Felipe – for sharing our PhD experiences during the numerous hikes we embarked on together.

Finally, I would like to thank Sebastian. You accompanied me through all the ups and downs every PhD entails, and provided more support than anyone else. Thank you.

Tina, October 2023

Contents

Introduction

About 85 percent of the global population live in low- or middle-income countries. Their lives are characterized by challenges that differ substantially from those in high-income countries. In a world where research and media attention is overwhelmingly focused on the most developed nations (Plancikova et al., 2021; Harris et al., 2017), it is often overlooked that these challenges are not only more pressing, but also those of the majority. The fields of development economics and sociology advance our understanding of what works to ensure that prosperity, security, and democratic participation are not the privilege of the few, but can be enjoyed by all.

One promising pathway to sustainable development, hailed by scientists and practitioners alike, is through education. Education equips people with essential skills, empowering them to improve their lives and contribute to their societies. At the individual level, it has been consistently documented to yield large private returns, including higher incomes (Angrist and Krueger, 1991; Peet et al., 2015), better health (Cutler and Lleras-Muney, 2006), lower crime propensities (Lochner and Moretti, 2004; Vogl et al., 2012), and greater life satisfaction (Oreopoulos, 2007; Belfield et al., 2006). Education has also been linked to a wide range of desirable aggregate outcomes such as higher rates of economic growth (Hanushek and Woessmann, 2012; Wantchekon et al., 2015), stronger institutions (Milligan et al., 2004; Treisman, 2000), and lower income inequality (Abdullah et al., 2015), highlighting its benefits for society as a whole.

Accordingly, governments and international organization have invested heavily to promote schooling in the developing world. However, while primary gross enrollment rates in low-income countries have risen steadily from 46% in 1970 to 100% in 2008, learning outcomes have failed to keep pace. Only 4 percent of students in low-income countries reach minimum literacy skills towards the end of primary school, compared to 95 percent in high income countries (World Bank, 2018, p. 8). Similarly, results from my own research in El Salvador show that the average fifth grader can answer less than half of the questions pertaining to the math curriculum of grades one and two (Büchel et al., 2022). To address this “learning crisis” in the developing world, we need to (1) find reliable ways to measure and track education, (2) find out what works to improve its quality, and (3) explore its potential to empower people for the advancement of their societies. This dissertation is a collection of articles in development economics and sociology that contribute to these three endeavors.

Measuring education with alternative data

Accurate and timely data on key development outcomes allows policy-makers to take informed decisions and track progress. However, particularly in developing countries, such data is often lacking. This has given rise to a new research field leveraging alternative data sources to bridge gaps in data availability (Burke et al., 2021). Prominent examples include the prediction of wealth using satellite imagery (Jean et al., 2016; Yeh et al., 2020) or phone records (Blumenstock et al., 2015). The first contribution of my PhD thesis ties into this strand of literature and shows that education can be accurately measured using geocoded Twitter (now X) data.

For this study, we collected over 25 million tweets from Mexico and the United States through the Twitter streaming API. We then constructed a series of interpretable measures, including Twitter penetration and usage statistics, text-based indicators on spelling mistakes, topics and sentiments as well as network indicators. Based on these features, we trained a stacking regressor combining five popular machine learning algorithms to predict educational attainment for Mexican municipalities (N = 2,457) and US counties (N = 3,141). Our results suggest that Twitter features are highly informative about education. Cross-validated predictions account for 70 percent of the variation in years of schooling in Mexico and 65 percent in the United States. This is a stark improvement over previous attempts to predict human capital using satellite data (Head et al., 2017), Google Street View images (Gebru et al., 2017), or Wikipedia articles (Sheehan et al., 2019).

Between 2010 and 2020, the number of internet users in low and middle income countries increased from roughly 1 billion to more than 3 billion. As people in the developing world continue to go online, the use of social media as an alternative data source is poised to become increasingly important. While the main model in our study is based on a large number of tweets collected over two months, we also demonstrate that relatively good performance can be achieved with just three days of tweets, reinforcing the practical applicability of our approach. The results of our study underscore the potential of using social media data and machine learning to understand and track global development, but they also point to challenges and open questions future applications should address. First, our paper discusses how the use of predicted measures in downstream tasks can introduce different types of biases, and shows how these biases can be corrected. With the notable exception of Ratledge et al. (2021), this has been largely neglected in the literature promoting the use of unconventional data sources in the social sciences.

A second challenge that is more specific to the applicability of our methodology is related to the reliance on Twitter data. While the use of the Twitter API was free of charge when we collected the data for this study, fees and restrictions have since been introduced, making our approach more costly. Exploring the potential of other social media networks as a substitute or complement to Twitter data may thus be a promising avenue for future research.

Finally, our paper constitutes a proof of concept in a context where education data is readily available, allowing us to train a predictor and evaluate its performance on ground truth data. The critical next step is to transfer this approach to contexts where no such data is available. This transfer has already taken place for the canonical example of satellite imagery, where a successful proof of concept was followed by applications bridging real data gaps (e.g., Aiken et al., 2022). As the performance of our approach is comparable to that of the satellite data in wealth prediction, using social media data to track education may hold a similar promise.

This project is joint work with Sebastian Heinrich, a PhD student in economics at ETH Zurich, and currently available as a working paper(see Jakob and Heinrich, 2023). I designed the project, collected the data through the Twitter API, implemented the machine learning algorithms, created the main tables and figures, and wrote the paper. I also contributed to the data processing, where Sebastian took the lead.

Strengthening teachers to close the global learning gap

To attain “quality education for all” (United Nations, 2015), we must not only measure and track learning but also find out what works to improve it. A substantial part of my academic work is thus dedicated to understanding why education systems in developing countries are failing and what can be done to make them more effective. My scientific interest in education in disadvantaged areas was sparked by my bachelor's thesis, which focused on inequalities in the accessibility of education. Using a survey with a random sample of 450 high school students in the department of Morazán in El Salvador, I analyzed the mechanisms of educational decision-making in a context of severe economic deprivation. This research project contributed to a better understanding of the processes shaping educational inequality in low-income settings. The results highlight that even as access to education has steadily improved in recent decades, substantial socio-economic disparities and barriers remain. During my PhD, my co-author Benita Combet and I wrote a scientific article based on my thesis, which is now published in *Research in Social Stratification and Mobility* (see Jakob and Combet, 2020).

For my master's thesis, I shifted my focus from the accessibility to the quality of education and submitted a project proposal to the "Impact Award" competition organized by Swiss Agency for Development and Cooperation (SDC) and the Swiss Federal Institute of Technology Zurich (ETH). After my project won the CHF 50,000 grant, I was joined by a team of researchers from the University of Bern and we conducted a large-scale randomized controlled trial (RCT) with 200 primary school classes in El Salvador. The field experiment featured three different treatments, allowing us to explore the potential of Computer-Assisted Learning (CAL) to improve math learning outcomes in public schools: *(i)* additional CAL lessons with a teacher, *(ii)* additional CAL lessons with a technical supervisor and *(iii)* additional traditional classes with a teacher. While the two CAL treatments lead to substantial and roughly equal learning gains, the intervention relying on traditional teaching produced only modest improvements in test scores. Our study was published in the *Journal of Labor Economics* in 2022 (see Büchel et al., 2022) and highlights the potential of technology in bridging educational disparities and promoting inclusive development. At the same time, the low productivity of the teacher-centered intervention raises important questions about the constraints the developing world's teachers face. Understanding these constraints and what works to address them is the focus of the second and third chapter of my dissertation.

In the quest to close the global learning gap, the effectiveness of a wide range of measures, including child health interventions, the provision of school materials, or teacher incentives, has been extensively discussed.[1] However, one crucial pillar of a successful education system (e.g., Barber and Mourshed, 2007; Hanushek, 2011) has received little attention so far: the skills of its teachers. In a recent review on teacher performance in developing countries, Bold et al. (2017, p. 202) conclude that "unfortunately, there are few, if any, well-identified studies on how to effectively improve teacher knowledge and skills and the impact thereof."[2] Two key constraints stand out: (1) gaps in content mastery and (2) outdated pedagogical knowledge.

The second contribution of my dissertation focuses on *teachers' content knowledge* and combines descriptive evidence on the respective shortfalls with an experimental evaluation of a technology-based teacher training program aiming to address them. In the first part of the study, we conducted a math assessment with a random sample of 224 primary school math teachers in the department of Morazán in El Salvador. Even though 97% of the teachers possess a university degree (13-17 years of formal schooling), the average teacher answered only 47% of grade two to grade six questions correctly. For example, only a third of the teachers could add two fractions (36%), and merely one in four (25%) could retrieve information from a descriptive chart. This is in line with recent findings from Subsaharan Africa and India, suggesting that many primary school teachers lack a basic understanding of the concepts they are supposed to teach (Bold et al., 2017; Sinha et al., 2016). Our descriptive results, presented in more detail in Brunetti et al. (2020), underscore the importance of finding effective ways to improve teachers' content knowledge.

In the second part of the study, we thus evaluated a content-centered training for math teachers in El Salvador. Inspired by the positive results of CAL-based instruction for students, we decided to see if this success could be transferred to teachers. The intervention we studied consisted of a five-month in-service program based on *(i)* computer-assisted content training at home and *(ii)* monthly revision workshops. To assess the causal impact of this program, 175 primary school teachers were randomly assigned to either the training program or to a control group. Our results show that treated teachers improved their math skills by 0.29σ immediately after the intervention, but this effect depreciated by about 70 percent one year afterwards. We also intended to measure potential learning gains for students, but our experiment was disrupted by the COVID-19 pandemic, and the respective

[1] For reviews, see Kremer et al. (2013), McEwan (2015) or Ganimian and Murnane (2016), Glewwe and Muralidharan (2016).

[2] See Snilstveit et al. (2015) or Popova et al. (2018) for further reviews on the topic.

assessments had to be canceled due to countrywide school closures. Instead, we conducted a series of simulations to compare the cost-effectiveness of CAL for students with that of CAL for teachers. Despite the lower cost and higher cascading potential of the teacher-centered approach, our results suggest that providing computers directly to pupils is likely to be more effective due to the high depreciation of effects at the teacher level. This study highlights the need for more research on how to make the effects of educational interventions more enduring. It is now published in the *Journal for Development Effectiveness* (see Brunetti et al., 2023), and was a joint project with my supervisor Ben Jann and Konstantin Büchel, Daniel Steffen, and Aymo Brunetti from the Department of Economics at the University of Bern. I was the main responsible for the coordination of the fieldwork in El Salvador and for writing the paper, and I was also involved in the project conceptualization, the design of the measurement instruments, and the data preparation and analysis.

Effective teaching requires not only knowledge of the subject matter, but also strategies for delivering that content to students. Accordingly, the third contribution of my dissertation focuses on *pedagogy*. As modern pedagogical theory stresses the importance of active student engagement, teachers in high-income countries have increasingly adopted student-centered models in the classroom. Yet, more teacher-centered approaches such as lecturing and rote learning are still the norm in many low- and middle-income countries. A study covering seven African countries concludes that only 11% of the teachers engage in pedagogical practices that are generally regarded as good teaching (Bold et al., 2017). In our study, we evaluated if switching to more participatory teaching strategies helps to close the global learning gap. We conducted a randomized controlled trial involving 440 math teachers and over 25,000 students from 220 schools in Tanzania. The intervention comprised a five-day in-service program focusing on participatory and practice-based teaching techniques. Additionally, half of the teachers in the treatment group received laptops with computer-assisted learning (CAL) software for content knowledge refreshment. The impact on students and teachers was assessed using data from standardized student assessments (scraped from the website of Tanzania's national examination council) and additional surveys, interviews, and classroom observations.

Training teachers in participatory pedagogy improved their students' test scores by 0.15σ after two years, with the proportion of top-performing students increasing by 6 percentage points (or 38 percent). This shows that promoting participatory pedagogy can be effective in improving student learning, even in a context with large classrooms and limited teaching aids that make the use of such methods more demanding. The additional provision of CAL software had no discernible effect on student performance, though teachers demonstrated slightly improved content knowledge in the subdomain of number sense and arithmetic. Our data suggest that many teachers in Tanzania already possessed sufficient mastery of their subject for effective teaching. While the average math teacher in our Salvadoran sample scored less than 50% on a test covering the math curriculum from grades two to six, the surveyed teachers in Tanzania achieved 78% correct answers on an almost identical assessment.

To address the persistent learning shortfalls in developing countries, it is essential to equip teachers with effective strategies to handle the challenging situation that schools in these nations present. By providing evidence on the merits of participatory teaching, our study contributes to this endeavor. The paper is currently published as a working paper and was conducted together with my co-authors Konstantin Büchel, Daniel Steffen, and Aymo Brunetti (see, Jakob et al., 2023). I wrote the web scraper to collect the student data, cleaned all the data (i.e., from student and teacher assessments), conducted all quantitative analyses, and wrote the paper, while my co-authors analyzed the qualitative data and took the lead in the coordination of the field work. I also supported the data collection process during two visits to Tanzania.

Taken together, the two teacher studies of my dissertation (chapters 2 and 3) provide important insights into the challenges of effective teaching in developing countries. First, our results show that a simple five-day training in participatory can induce teachers to transform their teaching and achieve better results for their students. Leveraging the potential of effective pedagogy may thus be an important pathway towards quality education in the developing world. Second, our results point to important shortfalls in teacher content knowledge. Although both studies suggest that such shortfalls can be partially addressed using CAL-based self-studying, ensuring that knowledge gains are substantial, persistent and passed on to students is not straightforward. Prior research indicates that achieving a 0.1σ gain in student learning would require a 1σ enhancement in teachers' content knowledge (Bau and Das, 2020; Metzler and Woessmann, 2012), a magnitude that surpasses realistic expectations for one-shot educational interventions. If it takes students 6 years to grasp the primary school math curriculum, we cannot realistically expect teachers to achieve the same in a few sessions of in-service training. A potentially more cost-effective solution might be to improve the instruction and selection mechanisms at teacher colleges. Future research could thus evaluate what works to ensure that future generations of educators graduate with fewer subject-related deficiencies. Finally, teacher-centered interventions are often hailed for their sustainability and cost-effectiveness. As typical teachers instruct many generations of students over the course of their professional careers, improving their skills and performance has the potential to produce vast and enduring effects. Our results suggest that there is no guarantee that this potential is realized, as teachers – like everyone else – may forget what they have learned. Finding out which interventions have lasting impacts and how effects can be made more enduring may thus represent an important focus for future research in the field.

Education as empowerment

Education is often hailed as an empowerment tool, enabling individuals and communities to take matters into their own hands and become the primary agents in a participatory development process. This notion is strongly mirrored in the concept of community-driven development (CDD), which has gained popularity as a bottom-up alternative to the conventional top-down approach in international cooperation. Typically, CDD initiatives rely on extensive facilitation processes to empower communities to collectively provide and protect local public goods. While this approach has attracted considerable scientific interest in the last two decades, its effectiveness has not yet been compared with the more traditional top-down strategy it aims to replace. In the last chapter of my dissertation, I offer such a comparison in the context of the waste management problem.

The study is based on a field experiment with 120 communities in rural El Salvador, and compares the effectiveness of two four-month treatments: (i) a traditional top-down intervention where streets are cleaned by an external team of cleaners, and (ii) a community-driven intervention where a facilitator raises awareness for the problem and mobilizes for collective action. We derive an objective measure of waste pollution by taking pictures along all streets and evaluating them using a deep learning model. These contamination assessments are complemented with data from a survey with 2,421 villagers and detailed records of all 883 activities conducted in the context of the two interventions.

In the short term, the traditional intervention reduced solid waste pollution by 0.7–0.8σ or 36 percent. Effects are significantly smaller, but still substantial for the community-driven intervention, with a reduction by 0.5–0.6σ or 29 percent. Long-term estimates show that four months after the end of the treatments, these effects depleted by 80 percent for the traditional intervention and by 60 percent for the community-driven intervention. Our complementary data from surveys and activity records also allows us to explore potential mechanisms behind these effects. We find limited evidence for information effects through increased awareness of the problem or knowledge of others' concern for it. Our results are most in line with a theoretical model where many individuals are willing to contribute to public goods as long as others do so too, but struggle to overcome organizational constraints in the absence of a dedicated leader.

While our study highlights the potential of education, it also points to its limitations. On the one hand, our findings suggest that educational processes can indeed empower communities to act collectively and promote local development. On the other hand, they also show that education will not always translate into action. Hence, the assumption that once people are equipped with the necessary skills, sustainable transformations will automatically follow, is often unrealistic. Building human capital is thus an important component of sustainable global development, but it may need to be complemented with other strategies to guarantee that the acquired knowledge is put to productive use.

This study is joint work with Carla Coccia, a PhD student at the Department of Economics at the University of Bern, and presented as a first draft. Both authors were equally involved in all major aspects of the project (i.e., project design, development of measurement instruments, fine-tuning of the deep learning models, data preparation and data analysis, preparation of figures and tables). In addition, I secured the funding for this project (approximately 100,000 USD) and wrote the paper draft included in this dissertation.

Field experiments and machine learning for development

My dissertation is characterized by the use of rigorous and innovative state-of-the-art methods to understand and promote global development. In particular, my contributions leverage the power of *(i)* randomized controlled trials and *(ii)* machine and deep learning methods.

Randomized controlled trials (RCTs) are considered the gold standard in empirical research. When properly implemented, randomization eliminates all confounding factors and, in the absence of spillover effects, produces an unbiased counterfactual. This offers a unique advantage over alternative methods of causal inference, which tend to rely on much stronger identifying assumptions (Duflo et al., 2008; Gertler et al., 2016). The rapidly growing strand of literature relying on RCTs to study social interventions has produced many important insights into the workings of our societies. Even more importantly, it has helped us understand what does and does not work to improve the livelihoods of the most vulnerable populations. Such insights have often had a real impact on policy-making and, thereby, people's lives. They allow us to learn from experience, continually improve policies, and direct resources towards the most effective initiatives. As a key ingredient to evidence-based policymaking, RCTs are thus an important driver of positive change. My dissertation ties into this strand of empirical research. Out of the four contributions included, along with my master's thesis completed during my PhD years, three involve RCTs.

In recent years, artificial intelligence (AI) has revolutionized many fields and opened up a range of new opportunities. At the same time, the advance of AI may deepen existing disparities if less privileged populations are left further behind. In this context, it is essential to explore ways to harness machine and deep learning methods to understand and address problems specific to low- and middle-income countries. One such problem are persistent gaps in the availability of data to track key social outcomes. The first chapter of my dissertation adds to the literature that leverages alternative data sources through machine learning to bridge such gaps. We provide a proof of concept that education can be accurately predicted based on geocoded Twitter data. The last chapter of my dissertation goes one step further and directly uses deep learning in the context of a field experiment to measure an outcome that would otherwise have been very costly to obtain. Using a fine-tuned state-of-the-art object detection model, we are able to construct an objective measure for contamination by identifying trash in about 200,000 images. To my knowledge, very few studies make use of deep learning in this way.

By combining rigorous methods of causal inference with machine and deep learning, my dissertation provides important insights for informed decision-making in vulnerable contexts. It contributes to closing both the global learning gap and the global data gap.

From research to practice

My decision to pursue a PhD was inspired by the conviction that scientific research can and should serve to understand and address real-world problems. Accordingly, all my projects have a use-oriented focus and aim to provide insights that can be directly translated into actionable solutions. To bridge the gap between research and practice, I dedicated a significant part of my time and effort to engaging with policy makers. In addition to the strong ties with the partner NGOs of my projects – Consciente in El Salvador and Helvetas in Tanzania – I also worked closely with educational authorities (i.e., the Salvadoran and Tanzanian Ministries of Education and the Tanzanian Teachers' Union). In numerous meetings and presentations in El Salvador and Tanzania, the implications of our findings were discussed with local stakeholders. They often found our experimental results, along with descriptive evidence such as teacher and student performance outcomes, very valuable and came to appreciate the benefits of evidence-based policy-making. In this spirit, I also wrote (or co-wrote) three policy briefs, two evaluation reports and two memos for local authorities during my PhD years. As the contributions in the first and last chapter of my dissertation were completed only very recently, the respective public outreach is still ongoing:

- Büchel et al. (2019). "Expanding School Time and the Value of Computer-Assisted Learning: Evidence from a Randomized Controlled Trial in El Salvador". Evaluation report, available at: `www.consciente.ch/calimpact_evaluation_report_april19`

- Büchel et al. (2020). "Self-paced and interactive learning with computers: Does it effectively boost children's math skills?" Policy brief, available at: `www.consciente.ch/policy_brief_cal`

- Brunetti et al. (2020). "Insights from the SITT-Baseline Assessment". Memo.

- Jann and Jakob (2021). "Consulta Magisterial: Conocimientos y Perspectivas de las y los Docentes en Morazán". Memo for the Ministry of Education in El Salvador, available at: `www.consciente.ch/report_catt_mined`

- Jakob et al. (2022). "A Mixed Methods Deluxe Evaluation of the School-Based In-Service Teacher Training (SITT) Program in Tanzania". Evaluation report, available at: www.aymobrunetti.ch/wp-content/uploads/2022/04/SITT_Evaluation_Uni_Bern_April2022.pdf

- Jakob et al. (2023). "Participative and Collaborative Teaching Approaches Make a Difference (Swahili: Mbinu Shirikishi za Ufundishaji Huleta Mabadiliko)". Policy brief.

- Brunetti et al. (2023). "Inadequate Teacher Content Knowledge and What to Do About It". Policy brief, available at: https://www.consciente.ch/catt-policy-brief/

Thanks to our dissemination efforts and the close cooperation with key local actors, my dissertation projects have already had a lasting impact on local policies and practices. As president and founder of Consciente, an NGO working on different education projects in El Salvador, I have been strongly involved in the educational field for over 10 years. This close link has enabled me to feed back the scientific evidence into Consciente's projects to improve initiatives and deploy the limited resources more effectively. The positive evaluation results for the CAL-based math lessons for students (see Büchel et al., 2022) enabled Consciente to scale up the initiative in coordination with the Ministry of Education and secure funding for the long-term continuation of the project. Every year, over 2,000 primary school pupils are now benefiting from the interactive CAL-based math lessons. As the findings on CAL-based teacher training were more mixed, a large-scale follow-up project was launched. In coordination with experts from pedagogical colleges in Switzerland, we designed three different teacher training programs, which we are currently evaluating in an RCT with 340 teachers and 7,000 students. The results of this study will provide further insights on how to improve teacher skills and student learning.[3] Finally, the promising effects reported in the evaluation of the two waste management interventions induced Consciente to incorporate an adapted version of the initiative, combing elements from both treatments, into its programs. The NGO is currently using the insights from the scientific study to improve the sustainability of the project and raise funds for its continuation.

[3] See https://www.socialscienceregistry.org/trials/10035 for the respective RCT registry.

The experimental evaluation of the participatory teaching initiative in Tanzania was originally intended to be a final report on a program that would then be discontinued. However, in response to the positive results, Helvetas decided to continue the project with the strong support of the Ministry of Education and the Tanzanian Teachers' Union, and is currently exploring how to improve its cascading elements. Approximately 300 teachers and their students benefit from the project each year.

Chapter 1

Measuring Human Capital with Social Media Data and Machine Learning

Measuring Human Capital with Social Media Data and Machine Learning

Martina Jakob
University of Bern
martina.jakob@unibe.ch

Sebastian Heinrich
ETH Zurich
heinrich@kof.ethz.ch

October 31, 2023

In response to persistent gaps in the availability of survey data, a new strand of research leverages alternative data sources through machine learning to track global development. While previous applications have been successful at predicting outcomes such as wealth, poverty or population density, we show that educational outcomes can be accurately estimated using geo-coded Twitter data and machine learning. Based on various input features, including user and tweet characteristics, topics, spelling mistakes, and network indicators, we can account for ∼70 percent of the variation in educational attainment in Mexican municipalities and US counties.

Keywords: machine learning, social media data, education, human capital, indicators, natural language processing

JEL Codes: C53, C80, O11, O15, I21, I25

We are grateful to Ben Jann, Carla Coccia, Mauricio Romero, and Joel Ferguson for their helpful comments and suggestions.

1 Introduction

Reliable data on key socio-economic outcomes enables policy-makers to take informed decisions and promote societal development. However, many countries are plagued by a pervasive lack of such data, limiting their ability to track progress and evaluate policies. To address the problem, a growing strand of literature uses alternative data sources such as satellite imagery or phone records to bridge the existing gaps in data availability (Burke et al., 2021). While previous studies have successfully predicted outcomes such as wealth, income or population density, this paper proposes an innovative approach to measuring human capital using geolocated Twitter data.

Specifically, we construct a series of interpretable measures of human capital at low administrative units (municipality in Mexico and county in the United States) based on over 25 million tweets. Our feature matrix includes simple Twitter penetration (e.g., user densities) and usage statistics (e.g., tweet length), text-based indicators on spelling mistakes (e.g., frequency of grammar mistakes), topics (e.g., share of tweets about science), and sentiments (e.g., share of negative tweets) as well as network indicators (e.g., closeness centrality). For each input, we compute cluster-level estimates based on geographical neighbors, and use them both as additional features and to impute missing values. We then train a stacking regressor combining five machine learning algorithms — elastic net regression, gradient boosting, support vector regression, nearest neighbor regression, and a feed-forward neural network — to predict educational attainment for Mexican municipalities (N = 2,457) and US counties (N = 3,141). We apply grid search to tune the relevant hyperparameters of each model, and evaluate the performance of the final models using five-fold cross-validation.

Our predictions account for 70 percent of the variation in years of schooling in Mexican municipalities and 65 percent in US counties. Where, how and what people tweet is thus highly informative about human capital. Within both countries, Twitter data appears to be particularly well-suited for distinguishing higher levels of education. For example, we achieve an r^2 of 0.70 when predicting county-level shares of US adults holding a bachelor's degree, while the corresponding r^2 for the percentage that completed high school is only 0.50. We observe a similar, though less pronounced relationship, for Mexico with an r^2 of 0.69 for the share with post-basic education and 0.61 for the percentage completing primary education.

Our focus on a limited number of meaningful variables also allows us to study which (groups of) features are most predictive of educational outcomes. In most models, user density emerges as the single most important predictor of educational outcomes. Twitter penetration features are particularly informative in Mexico, where (on their own) they account for 57 percent of the variation in educational outcomes, compared to 37 percent in the

US. Similarly, error and network features appear to be strongly related to human capital in Mexico ($r^2 = 0.55$ and 0.51, respectively), but less so in the US ($r^2 = 0.42$ and 0.34, respectively). General tweet statistics and topics have consistently high predictive power in both countries (r^2 between 0.5 and 0.6). In Mexico and the United States including cluster-level features is critical, improving model performance by almost 10 percentage points.

The main challenge to model performance arises in sparsely populated areas with low Twitter penetration. Accordingly, the population-weighted r^2 for years of schooling is 0.85 for Mexico and 0.70 for the US (compared to 0.70 and 0.65 in our unweighted base model). Similarly, restricting the evaluation sample to areas with at least ten users would increase performance to 0.74 in Mexico and 0.68 in the US. We also explore how model performance evolves depending on the data collection period, finding that we can achieve relatively high predictive power with just three days of tweet data, namely an r^2 of 0.66 for Mexico and 0.58 for the United States.

Using wealth data for Mexico and income data for the US, we further explore how our human capital measure performs in downstream tasks by comparing regression results based on predicted vs. ground truth education measures. We find that slope coefficients tend to be biased not only when using the predicted indicator as an independent variable, but also when it acts as the dependent variable. The latter bias results from the typical model tendency to overpredict for low and underpredict for high values and is likely to affect most applications. When using a loss function that penalizes quintile-specific biases (see Ratledge et al., 2022), the bias effectively disappears, and regression coefficients based on our predicted indicator become very similar to their ground truth counterparts. Our simulations show that when appropriately modeled, predicted indicators can produce correct estimates in downstream regression tasks as long as they serve as the outcome and not the treatment variable.

This paper contributes to the recent literature exploring the combined potential of non-conventional data sources and machine learning to measure and understand socio-economic development. While a range of outcomes including wealth (Jean et al., 2016; Blumenstock, Cadamuro, and On, 2015; Yeh et al., 2020; Aiken et al., 2022), population density (Stevens et al., 2015; Wardrop et al., 2018), crop yield (Lobell, 2013; Burke and Lobell, 2017; Sun et al., 2019), informal settlements (Kuffer, Pfeffer, and Sliuzas, 2016; Mboga et al., 2017), electricity access (Ratledge et al., 2022), and disease spread (Wesolowski et al., 2012; Chang et al., 2021) have been accurately predicted using satellite or phone data, previous attempts to infer human capital have been less successful. Head et al. (2017) use satellite data to predict educational attainment in Rwanda, Nigeria, Haiti and Nepal, achieving an average r^2 of $\sim$0.55. The predictive power of other data sources, such as Google Street View images (Gebru et al., 2017) or Wikipedia articles (Sheehan et al., 2019), appears to be even lower, accounting for less than 40 percent of the variation in educational outcomes. We show that by using geolocated Twitter data and natural language processing, we cannot only derive a more accurate indicator of human capital than previous studies but also achieve similar performance to the renowned wealth prediction with satellite data.

We also add to the literature leveraging social media data for social science research. Almost five billion people worldwide used at least one social media platform in 2023, and another billion is projected to join until 2027, as emerging and developing economies are catching up (Poushter, Bishop, and Chwe, 2018; Statista, 2022). Thus, using social media data to understand and track development is likely to become increasingly relevant in low- and middle-income countries where the scarcity of reliable traditional data sources tends to be most pronounced. Social media data has been used to predict or study diverse outcomes such as migration (Huang et al., 2020; Yin, Gao, and Chi, 2022), social capital (**chetty2022social**), censorship (King, Pan, and Roberts, 2013), alcohol consumption (Curtis et al., 2018) or stock market prices (Bollen, Mao, and Zeng, 2011). Moreover, micro-evidence suggests that social media posts are informative about individual users' educational characteristics (Smirnov, 2020; Gómez et al., 2021). This paper goes one step further and shows that despite the high endogenous selection in social media usage (Mellon and Prosser, 2017), the respective data can be used to derive accurate education estimates at low administrative units within countries.

Finally, this paper makes two methodological contributions. First, it ties into the nascent methodological discussion on the validity of predicted indicators for downstream regression tasks (Ratledge et al., 2022). While the main focus of the previous literature has been on achieving high predictive performance, we also discuss how regression estimates are affected by different biases and show how the most detrimental of these biases can be corrected. Second, we propose an innovative solution to deal with sparse or noisy data in areas of low population density. By allowing our models to not only learn from data in the observed units, but also from spatial neighbors, we achieve a substantial improvement in performance. This approach could be beneficially transferred to other applications, as geographical information is usually readily available and many outcomes are spatially correlated.

2 Data and Methods

2.1 Collection and Processing of Twitter Data

We used the Twitter Streaming API to compile a large tweet dataset for Mexico and the United States. Twitter's Streaming API grants real-time access to information on 1% of all tweets, including the text of each tweet as well as a series of tweet and user characteristics.[2] Our final dataset consists of 2,686,779 geo-localized tweets from 123,309 users for Mexico and 22,610,134 tweets from 943,164 users for the United States, gathered between July and August 2021. The tweets included in our final dataset were selected based on three criteria:

1. *Geographical location*: We excluded all tweets that were not posted from within the geographic territory of the respective country. In the case of the United States, we use all tweets from the mainland, Alaska and Hawaii, but not from unincorporated territories such as Puerto Rico or the Virgin Islands. We also exclude tweets without precise location information (i.e., less than municipality/county level precision). Our final sample comprises tweets with exact coordinates (MX: 3%, US: 3%), neighborhood or point of interest (poi) level precision coordinates (MX: 2%, US: 2%), and city-level precision coordinates (MX: 95%, US: 94%).

2. *Language*: For each country, only tweets written in the main native language (i.e., Spanish for Mexico and English for the United States) are included.

3. *Source*: One key concern regarding the reliability of Twitter data is that many tweets are automatically spread through APIs rather than individually created by a human user. We thus restrict our sample to content that is posted through the main four channels for human users: iPhone, Android, iPad, and Instagram.[3] This excludes tweets generated through third-party APIs from platforms such as Foresquare or CareerArc (approximately 1 percent of geo-localized tweets in Mexico and 7 percent in the United States).

To compute municipality or county-level statistics, we follow a three-stage procedure. First, each tweet is assigned to a geographical unit (i.e., municipality or county) based on its coordinate data. While this is straightforward for exact coordinates, we have to apply different types of consistency checks to find the correct unit when coordinate information consists of a city, poi, or neighborhood level bounding box.[4]

[2] The use of the Twitter streaming API was free of charge until the beginning of February 2023, when a fee was introduced.

[3] For tweets posted through Instagram, we exclude all tweets using the default text ("Just posted a photo @...") rather than a message specified by the user. Tweets posted through the Twitter website are not included in our sample as they do not have any associated coordinates.

[4] In most cases, assignment to the geographical unit harboring the centroid of the tweet bounding box yielded correct results. However, particularly in the Mexican case, where location precision for tweets tends to be lower (and city level-precision as defined by Twitter refers to municipalities rather than places within municipalities), we combine spatial joins with name matching to ensure all tweets are assigned to the correct entity.

Next, we approximate the home municipality or county for each user. If users tweet from more than one geographical entity (MX: 33% of users, US: 35% of users), we assign all their tweets to the entity from which they tweeted the most. For users with equal numbers of tweets in two or more entities (MX: 1%, US: 2%), we use the number of tweets posted during non-work hours on weekdays to break ties. This procedure results in the reassignment of 14 percent of tweets in Mexico and 12 percent of tweets in the United States. Tweets that cannot be unambiguously assigned to a municipality through this procedure are dropped (MX: 0.4%, US: 0.2%).

Finally, data is aggregated at the municipality or county level using the unit-level sum, mean, or median depending on the distribution of the underlying variables (for details, see Section 2.3 and Appendix C). To give equal weight to all users irrespective of their degree of activity, all tweet-level variables are first aggregated at the user level.

2.2 Survey Data

While many countries lack timely and spatially disaggregated information on educational outcomes, such data are available for both Mexico and the United States, allowing us to train and test a prediction algorithm in two different settings. Our main outcome variable is years of schooling for both countries, but we also look at the share of adults holding different educational degrees to better understand at which point of the educational distribution our models work best (see Table A8). We use data from the 2020 census for Mexico and from the American Community Survey (2017–2021, 5-year estimates) for the United States.[5] Following Barro and Lee (2013), we approximate county level years of schooling for the US based on the proportions holding different educational degrees and the averages for the years of schooling these degrees correspond to.[6]

Section C in the Appendix presents summary statistics on all outcome variables. In the average Mexican municipality, 28 percent of the population holds a post-basic degree, 54 percent graduated from secondary school, 76 percent finished primary school, and the average person completed 7.8 years of schooling. The corresponding figures in US counties are 23 percent with a bachelor degree, 54 percent with some college, 88 percent with a high

[5] The Mexican census data is publicly available at `https://www.inegi.org.mx/datosabiertos/` while data from the American Community Survey can be accessed at `https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data/`.

[6] Average years of schooling for a given county are calculated by $\sum_j h_j \, Dur_j$, where h_j indicates the fraction of the population having attained education level j and Dur_j indicates the respective duration to attain level j. We use data from the *Current Population Survey*, specifically the *2021 Annual Social and Economic (ASEC) Supplement*, to compute estimates for Dur_j. In Mexico, this approximation is not necessary as average years of schooling are included in the census data.

school degree, and 13.3 years of schooling.[7]

2.3 Features

Our feature matrix comprises municipality-level information on *(i)* Twitter penetration, *(ii)* Twitter usage, *(iii)* spelling mistakes, *(iv)* topics, *(v)* sentiment, and *(vi)* user networks. In addition, we also include population density estimates.[8] To advance our understanding of the aspects of people's online behavior that are most predictive of human capital, we deliberately focus on a limited number of interpretable features rather than using, for example, tweet text embeddings (for a detailed overview, see Section C and D in the Appendix).

Table 1: Summary statistics by education level for selected features

	Mexico			United States		
	Bottom 25%	Top 25%	All	Bottom 25%	Top 25%	All
User density	0.23	0.86	0.47	0.79	2.49	1.45
Tweet density	1.94	16.70	7.00	12.03	45.14	24.40
Tweet length	68.75	72.88	69.94	77.09	82.05	80.86
Account age	5.03	6.34	5.67	6.67	7.51	7.06
Tweets per year	1,306.55	362.71	841.93	648.93	351.58	495.19
Favorites per tweet	5.02	1.34	3.76	1.52	2.14	1.73
Error total	24.60	23.54	25.28	15.23	13.14	13.87
Error grammar	0.17	0.15	0.17	0.65	0.47	0.55
Error typos	12.18	10.66	12.47	7.48	6.92	7.19
Topic science	1.84	1.92	1.87	1.58	1.82	1.69
Topic relationships	6.66	5.72	6.27	5.31	4.42	4.76
Sentiment positive	0.39	0.37	0.38	0.50	0.50	0.50
Offensive language	0.15	0.16	0.15	0.17	0.16	0.16
Network clos. centr.	0.06	0.31	0.16	0.28	0.42	0.34
Number of Areas	430	429	1,714	723	723	2,889

Municipality (MX) or county (US) averages for selected features by educational outcome. The bottom 25% and top 25% refer to the municipalities/counties in the lowest or highest quartile with regard to years of schooling. Only areas with at least one tweet are included. Features are not log-transformed.

Twitter penetration data (4 features) consists of the total number of tweets and users as well as the number of users and tweets relative to the population (referred to as user

[7] MX: Estimates for years of schooling, primary and secondary completion are provided for the population aged 16 or more, while the share with post-basic education is defined for adults (i.e., over 18). US: All education statistics refer to the population aged 25 or older.

[8] Population data is globally available; consequently, its inclusion does not limit the external validity of our approach. Population data is also necessary for the computation of tweet and user densities. A model using only population estimates will serve as our benchmark against which the performance of our approach is compared.

and tweet densities). We further include general information on *Twitter usage* (11 features) such as the average tweet length, number of followers, user mobility, account age, number of emojis per tweet, or the share of tweets posted during working hours or from an iPhone. To obtain estimates for the frequencies of different *spelling mistakes* (MX: 23 features, US: 16 features), we use a Python wrapper for "LanguageTool", an open-source grammar, style, and spell checker. LanguageTool is available in over 25 languages, including English and Spanish, and classifies the detected errors into different categories such as grammar, typos, casing, punctuation, or style.[9] We include the total number of errors per 1,000 characters and the corresponding figures for each category. To determine the *topics* of each tweet (19 features), we use a pre-trained multi-label tweet classification model (Ushio and Camacho-Collados, 2022). This allows us to estimate the probability a given tweet is about a specific topic such as news, celebrity, sports, or science. As no pre-trained tweet classification models are available in Spanish, we translate all Spanish tweets to English using a pre-trained model based on the Marian NMT framework (Junczys-Dowmunt et al., 2018) to determine the topic distributions of our Mexican tweets.[10] A further group of inputs comprises features related to *sentiments* (4 features), such as the share of tweets with negative or positive sentiments, offensive language, or hate speech. They are generated using pre-trained classification models for Spanish and English tweets.[11] Finally, we also add *network indicators* (4 features), such as degree and closeness centrality. We use quotes and mentions to construct a user-to-user network and subsequently aggregate this network to the municipality or county level. We take the log of right-skewed features and standardize all features before training.[12]

To address potential problems related to sparse or noisy data in areas of low population density, we develop a procedure that allows our model to learn from spatial neighbors. For each unit (i.e., municipality or county), we create a cluster consisting of the focal unit and all its spatial neighbors and compute cluster-level estimates for each of our features. We use this information about Twitter usage in the broader area around each unit in three ways: First, we add the cluster-level estimates as additional inputs to our feature matrix (i.e., for each unit and measure, we include both unit and cluster-level values). Second, we use cluster-level features to impute missing values in units without tweets using an elastic net regression model. This provides estimates for features that cannot be observed in the absence of tweets, and is necessary as most machine learning algorithms cannot deal with

[9] See `https://dev.languagetool.org/languages` for information on language availability.

[10] The model is provided via the HuggingFace library: `https://huggingface.co/docs/transformers/model_doc/marian`.

[11] The classification models are provided by the same library used for the topic classification above.

[12] Appendix D documents which variables are log-scaled. Following Stahel (2000), we use $log(x + c)$ to deal with zeros, with x as the values of a particular feature and $c = Q^2_{0.25} \,/\, Q_{0.75}$, where $Q_{0.25}$ and $Q_{0.75}$ are the first and the third quartile based on feature values $x > 0$.

missing values. Third, in units with less than 5 tweets, we replace extreme outliers with imputed values using the same imputation procedure.[13]

Table 1 shows the mean of selected features by educational level for both countries (see Section C in the Appendix for complete summary statistics).This simple inspection already reveals a strong correlation between Twitter features and educational outcomes. In both countries, user and tweet density is markedly higher in places with more educated populations. Similarly, users in more educated areas tend to write longer tweets, make fewer errors and talk about different topics (e.g., science rather than relationships). On the other hand, users in areas with lower educational attainment are, on average, tweeting more actively.

2.4 Training and Evaluation

To train our models, we use a stacking regressor combining five machine learning algorithms: *(i)* elastic net regression, *(ii)* gradient boosting, *(iii)* support vector regression, *(iv)* nearest neighbor regression, and *(v)* a feed-forward neural network (i.e., a multi-layer perceptron). We use grid search to tune the hyperparameter of each model. The performance of the final stacking regressor is evaluated using five-fold cross-validation. We report the cross-validated r^2 for each fold as well as an overall r^2 obtained by combining all cross-validated predictions.

3 Results

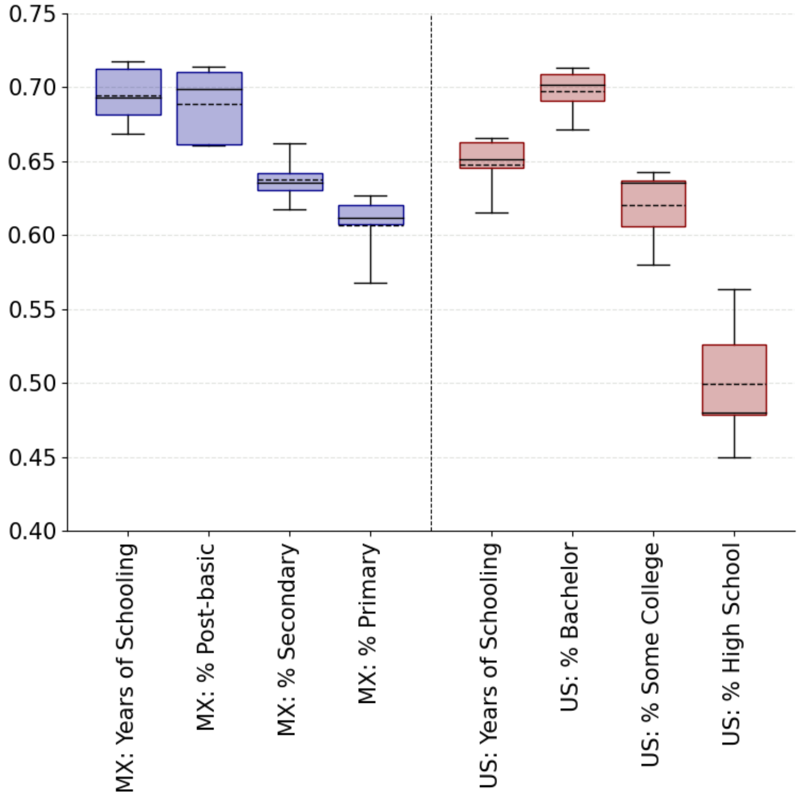
3.1 Main Results

Our final model is able to account for 70 percent of the variation in years of schooling in Mexican municipalities and 65 percent in US counties (see Figure 1). Population-weighted performance estimates are even higher, reaching an r^2 of 0.85 in Mexico and of 0.70 in the United States.[14] A closer look at the predictive power for different educational degrees reveals substantial variation in model performance in both countries.

In Mexico, we report an r^2 of 0.69 for the share of the population holding a post-basic degree (i.e., high school or more), an r^2 of 0.64 for the corresponding share with a secondary degree, and an r^2 of 0.61 when aiming to predict the prevalence of primary school completion. Differences are even more pronounced in the United States, where our model captures 70 percent of the variation in the percentage of adults that hold bachelor's degree, 62 percent for the share that went to college, and 50 percent when focusing on high school completion.

[13] Extreme outliers are defined as values that are lower than $Q_{0.25} - 3\,IQR$ or higher than $Q_{0.75} + 3\,IQR$, with $Q_{0.25}$ and $Q_{0.75}$ as the first and the third quartile and IQR as the interquartile range.

[14] Population weights are not taken into account during training.



MX	r^2
Years of schooling	0.695
	(0.847)
% Post-basic	0.690
	(0.820)
% Secondary	0.638
	(0.815)
% Primary	0.608
	(0.809)

US	r^2
Years of schooling	0.648
	(0.697)
% Bachelor	0.698
	(0.753)
% Some College	0.621
	(0.683)
% High School	0.499
	(0.589)

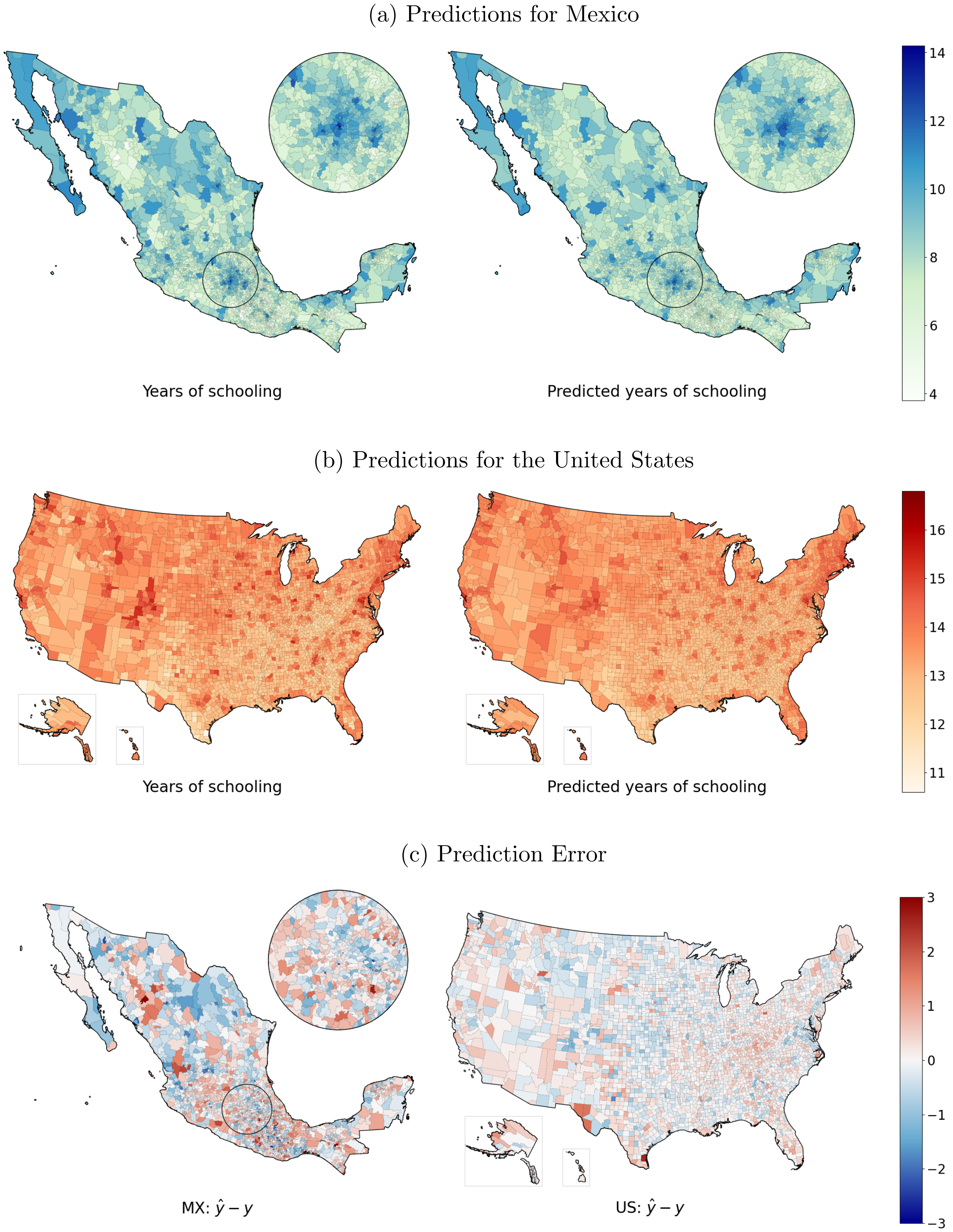
Figure 1: Performance for different educational outcomes in Mexico and the United States

All models are evaluated through five-fold cross-validation. Boxplots show the median (solid line), mean (dotted line), the 20th & 80th percentile (box limits), as well as the minimum & maximum (whiskers) for the r^2 across validation folds for each outcome and country. The table on the right presents the r^2 based on out-of-sample predictions for the full data sets (stacked across folds). Population-weighted r^2 are presented in parentheses. All models are evaluated through five-fold cross-validation.

This suggests that Twitter data is particularly informative about higher education levels and less sensitive to differences at the lower end of the education distribution.

Among the five included models, gradient boosting and support vector machines perform best and, accordingly, receive the highest weights in the final stacking regressor (see Figure A1 and Table A1 in the Appendix). The neural network and the nearest neighbor regressor, on the other hand, perform rather poorly, achieving a lower predictive power than the simple elastic net model (i.e., a regularized linear model). For all outcomes, the ensemble of all models outperforms the best-performing individual model, highlighting the benefits of stacking.

As Figures 3a and 3b show, our model produces the attenuated predictions that are typical for continuous outcomes (Ratledge et al., 2022), meaning that, on average, estimates



(a) Predictions for Mexico

(b) Predictions for the United States

(c) Prediction Error

Figure 2: Maps of true vs. predicted years of schooling

Predicted values for all municipalities and counties are obtained by combining out-of-sample predictions from all folds. In Figure 2c, red indicates overprediction and blue underprediction of true values.

are too high in low-education and too low in high-education areas.[15] This pattern also becomes apparent when comparing maps of true and predicted years of schooling (see Figures 2a and 2b). While spatial patterns look very similar for the two measures, they are slightly less fine-grained in the prediction maps. Similarly, Figure 2c shows that prediction errors tend to be spatially correlated.

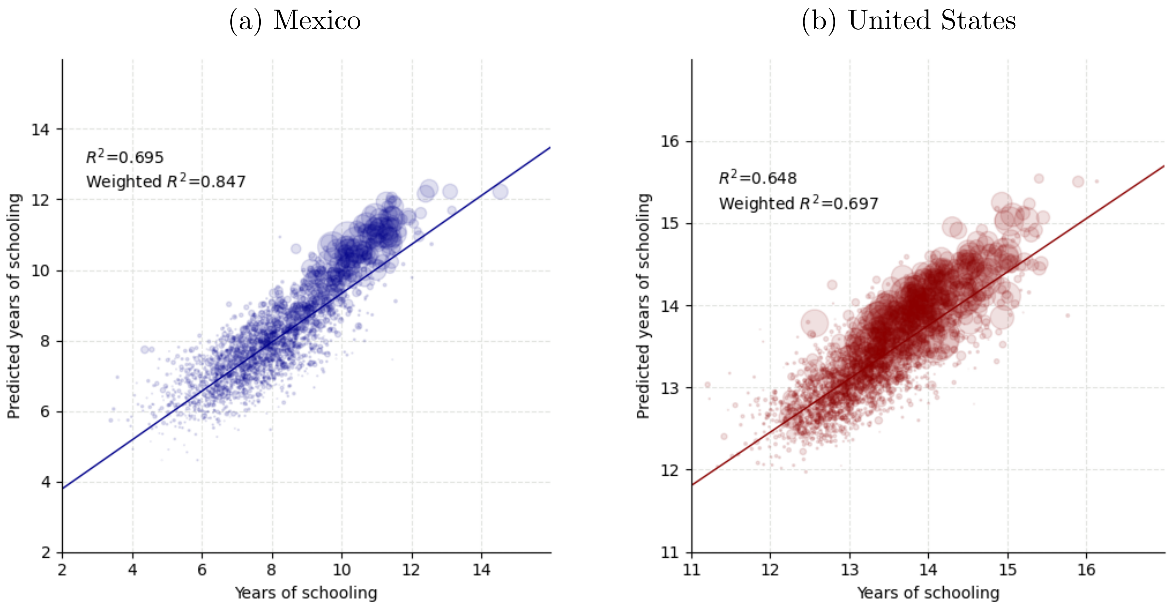


Figure 3: True vs. predicted years of schooling

Predicted values for all municipalities and counties are obtained by combining out-of-sample predictions from all folds. Bubble size is proportional to the population in each unit. r^2 and population weighted r^2 shown. The line indicating the best linear fit is not population-weighted.

3.2 Feature Importance

As our model is based on a limited number of interpretable inputs (see Sections C and D in the Appendix), we can explore how important various types of features are to the success of our approach. Figure 4 shows how different groups of features perform on their own. A model using only population data serves as a benchmark, reaching an r^2 of 0.48 for Mexico and 0.34 for the United States. Simple Twitter penetration data, that is, user and tweet densities/counts, already outperforms the population model, with r^2 values of 0.57 for Mexico and 0.36 for the United States. Particularly in Mexico, knowing where people tweet is thus more informative about human capital concentration than knowing where people live.

[15] The regression line in Figure 3 and Appendix Figure A2 does not take population weights into account. The fact that there are many sparsely populated areas at the lower, and few, but very populous areas at the higher end of the education distribution, creates the illusion that the line does not fit the data.

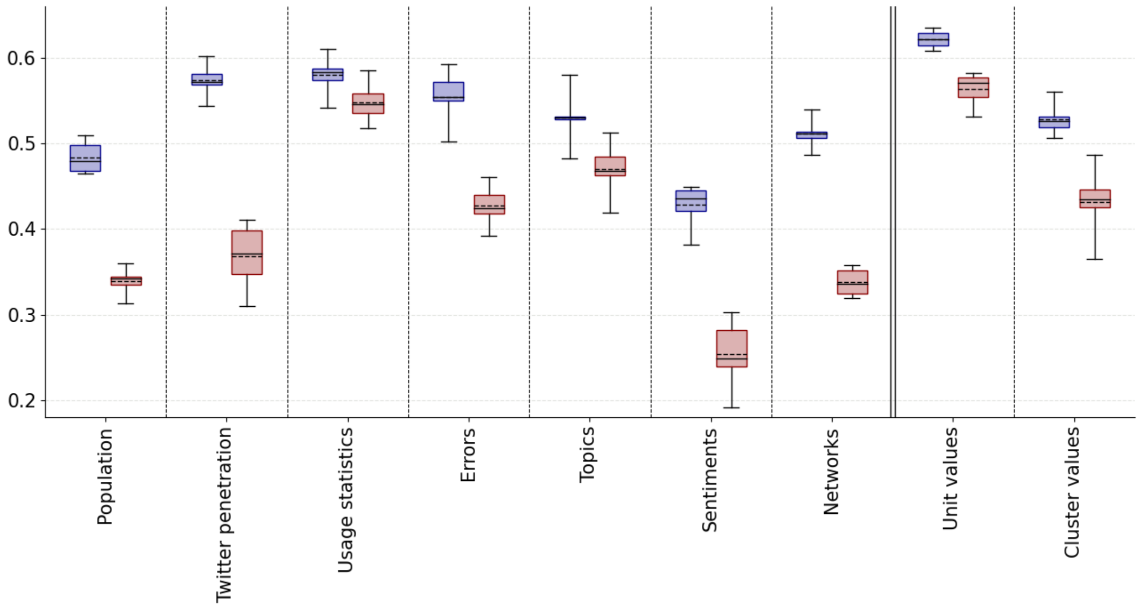


Figure 4: Performance of feature subgroups

Performance of feature subgroups for Mexico (blue) and the United States (red): Population (2x4 features, i.e., 4 at the unit level and 4 at the cluster level), Twitter penetration (2x4 features), usage statistics (2x11 features), spelling mistakes (MX: 2x23 features, US: 2x16 features), topics (2x19 features), sentiment (2x4 features), and networks (2x4 features), as well as all unit level (i.e., municipality or county) and all cluster level (i.e., including spatial neighbors) features. All models are evaluated through five-fold cross-validation. Boxplots show the median (solid line), mean (dotted line), the 20th & 80th percentile (box limits), as well as the minimum & maximum (whiskers) for the r^2 across validation folds for each outcome and country. The outcome is years of schooling in all models.

The performance of usage statistics, that is, features such as the average tweet length or the number of followers, is high in both countries, accounting for 55 to 58 percent of the variance in educational outcomes. The same is true for topic variables, which reach an r^2 around 0.5 in both countries. Error and network statistics, on the other hand, seem to be much more strongly related to human capital in Mexico (r^2 of 0.55 for errors and 0.51 for networks) than in the United States (r^2 of 0.42 for errors and 0.34 for networks). Finally, sentiment features constitute the only group of variables that fails to surpass the benchmark model. Overall, the performance of no single group of features comes close to that of the overall model, suggesting that the different inputs are complementary.

When looking at the contributions of individual features, the user density seems to be the most important predictor in the majority of models (see Appendix Figures A3 and A4).[16] The importance of other features varies more strongly between countries (and measures of feature importance), but network features such as closeness centrality or out degree, simple

[16] The reported feature importances are not based on the final stacking model, but computed separately for (1) the elastic net and (2) the gradient boosting model. Due to the high collinearity between different features, results should be interpreted with care.

usage statistics including the tweet length or the account age, as well as specific topics and errors tend to be very predictive too.

We can also evaluate how our model benefited from including cluster-level features (see Figure 4). When limiting ourselves to unit-level features, we report r^2 values of 0.63 (MX) and 0.56 (US), as opposed to 0.70 (MX) and 0.65 (US) for the full model.[17] Thus, exploiting information from spatial neighbors is critical to the predictive power of our models.

3.3 Performance Heterogeneity

We now explore how our model is affected by the limited number of tweets in sparsely populated areas (Figure 5). In line with expectations, performance is substantially higher when limiting the evaluation to municipalities or counties with more tweets or users. This relationship is even more pronounced when looking at different population thresholds. Particularly in Mexico, model performance increases drastically if we exclude smaller municipalities, where both input and output data is likely to be more noisy. This is consistent with finding that, in both countries, the population-weighted r^2 is substantially higher than the unweighted r^2 for all outcomes.

It is also informative to look at performance by the amount of data we use for the predictions. We streamed Twitter data for two months for our main analyses and used millions of tweets to construct municipality or county-level indicators. To see if similar results can be achieved with a shorter data collection period, we re-run the entire feature engineering and model training procedure on different subsets of our data. As Figure 6 shows, a drastic shorting of the data collection period only marginally reduces performance. This is particularly true in Mexico, where one day of tweets already yields an r^2 of more than 0.65. In the US, on the other hand, about one week of Twitter data is needed to account for 60 percent of the variation in county-level education outcomes. As the curves for both countries flatten out almost completely after a few weeks, extending the data collection period beyond two months is likely to yield only negligible additional performance gains.

[17] This provides a lower bound for the true benefit of exploiting spatial information as cluster-level features are also used to impute missing values and extreme outliers.

[18] Standard errors (shaded area) are computed using $\sqrt{\frac{4r^2(1-r^2)^2(n-k-1)^2}{(n^2-1)(n+3)}}$, where n is the sample size and k is the number of features (Cohen et al., 2013).

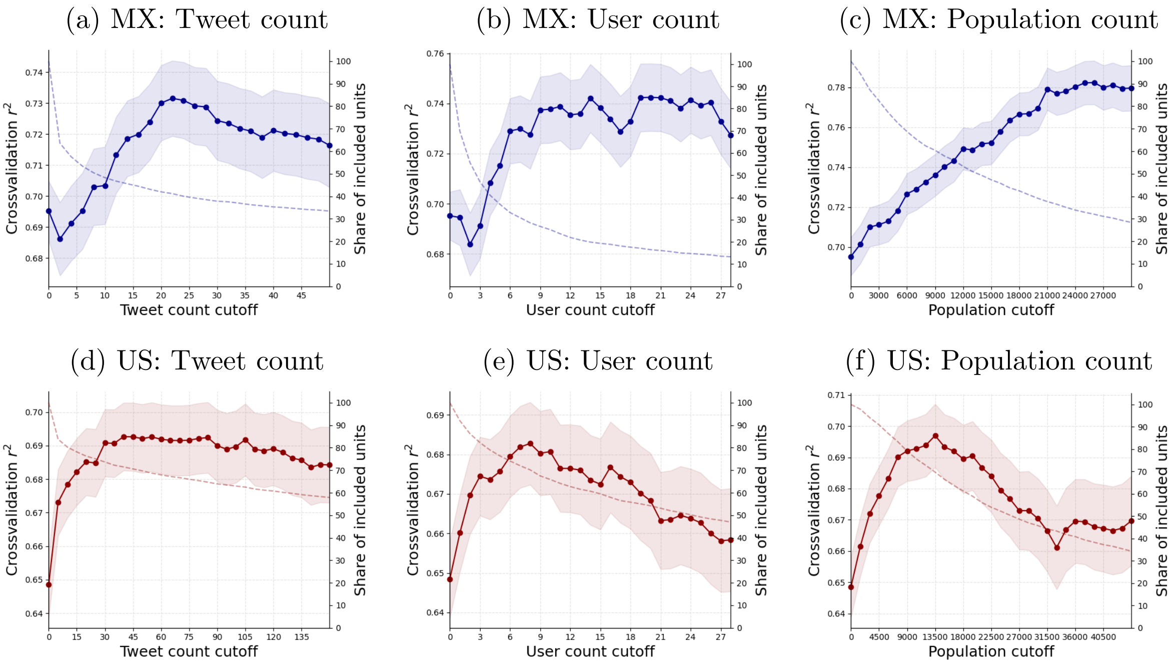


Figure 5: Performance heterogeneity by user, tweet, and population count

The solid line shows the r^2 for units (municipalities or counties) above different tweet, user or population count cutoffs.[18] The proportion of units included at each cutoff is represented through a dashed line.

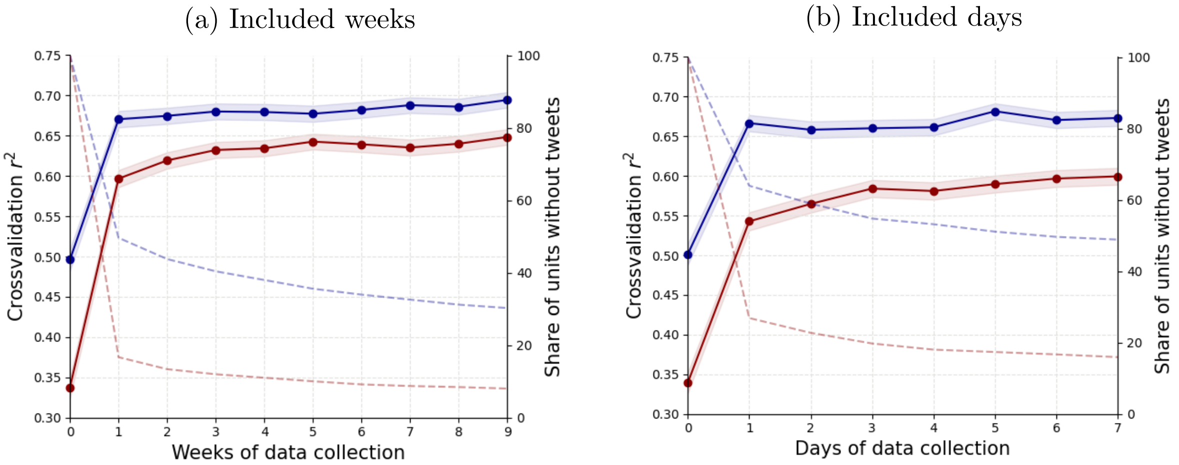


Figure 6: Performance by data collection period

Value for 0 weeks/days corresponds to r^2 of our baseline model using population data only. Standard errors are computed using the same formula as reported in Figure 5.

3.4 Downstream Performance

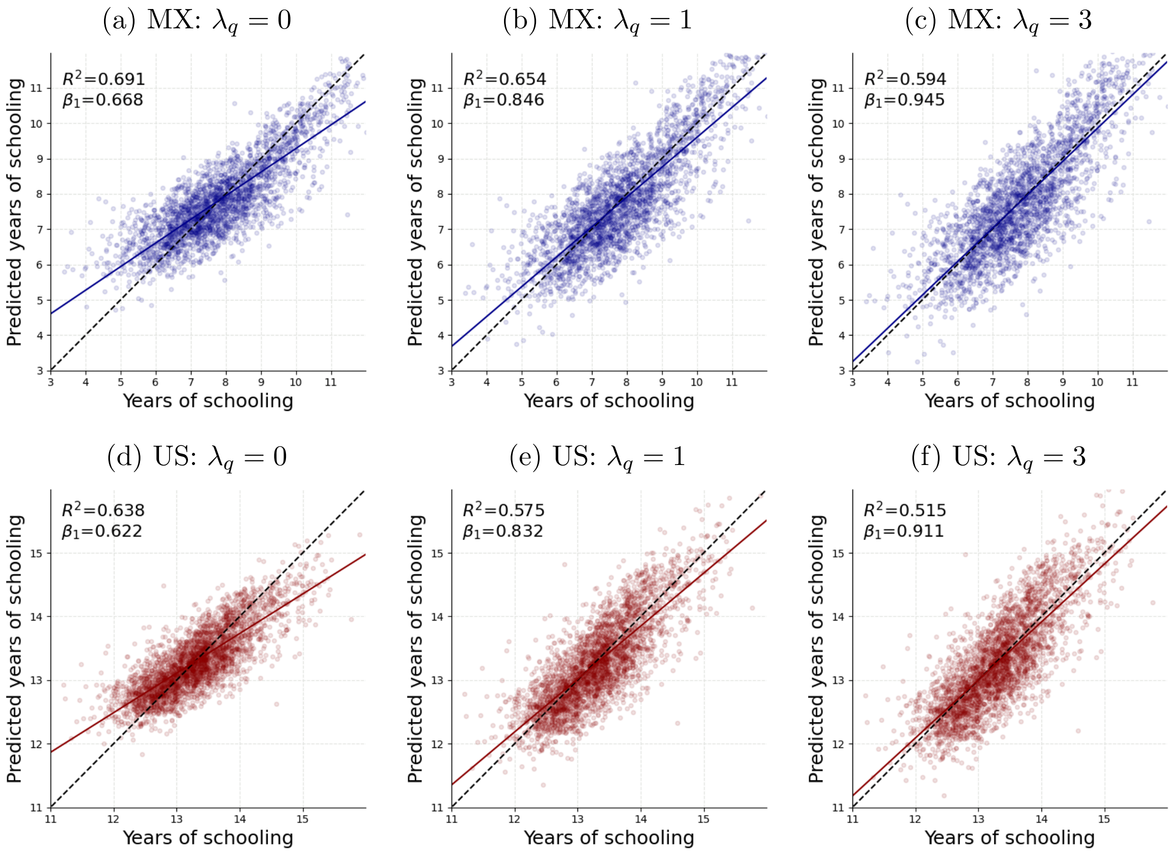
Apart from being directly useful to better understand local patterns in development outcomes and target interventions accordingly, predicted measures may also serve to study relationships with other variables. Using wealth data for Mexico and income data for the United States (see Appendix Table A8), we thus explore how our Twitter-derived indicator performs in downstream regression tasks. The fact that machine-learning-derived indicators are noisy measures gives rise to several potential biases that may jeopardize such applications. If *edu* is the true distribution of the indicator we predicted as $\widehat{edu}$ (e.g., years of schooling), and *econ* is another variable whose relationship to *edu* we would like to study (e.g., wealth), three types of measurement error may occur (see simulations in Appendix Figure A5):

1. Attenuation bias: A random measurement error in $\widehat{edu}$ will cause the correlation between *edu* and *econ* to become diluted. This results in an attenuation bias when regressing *econ* on $\widehat{edu}$, but not in the opposite specification, and decreases precision in both cases (see, e.g., Fuller, 1987).

2. Berkson-type error: A bias that has only recently gained attention (see Ratledge et al., 2022) arises when measurement errors are correlated with *edu*. The typical machine learning model behavior is to overpredict for low and underpredict for high values, a pattern that is very apparent in our application, where the correlation between the prediction error (i.e., $\widehat{edu}$ - *edu*) and *edu* amounts to about -0.6. This does not have an impact on the correlation between *edu* and *econ*, but it distorts coefficients in downstream regressions. Specifically, it leads to a downward bias when $\widehat{edu}$ is used as the outcome variable, and to an upward bias when it acts as the explanatory variable.

3. Correlated learning: If the features used to predict $\widehat{edu}$ contain wealth or income-related information, our model might exploit the correlation between *econ* and *edu* to make better predictions. Indeed, our feature matrix is almost as predictive of economic outcomes ($r2 = 0.64$ for wealth in Mexico and $r2 = 0.62$ for income in the US) as of education.[19] This creates an artificially strong correlation between $\widehat{edu}$ and *econ*. When using $\widehat{edu}$ as the dependent variable, this only leads to overoptimistic standard errors. If $\widehat{edu}$ is the independent variable (and *edu* and *econ* are positively correlated), it additionally induces an upward bias for the point estimate.

[19] This is substantially higher than a model using education only (years of schooling) for the prediction (MX: 0.57, US: 0.50), suggesting that our feature matrix indeed contains wealth and income-related information that is independent of education levels. Estimates are based on re-running the same machine learning procedure we use to predict education for wealth and income.



(a) MX: $\lambda_q = 0$ (b) MX: $\lambda_q = 1$ (c) MX: $\lambda_q = 3$

(d) US: $\lambda_q = 0$ (e) US: $\lambda_q = 1$ (f) US: $\lambda_q = 3$

Figure 7: True vs. predicted values with correction of the Berkson-type error

To correct for the Berkson-type error, we apply an adjusted loss function in the final ridge regression model that performs the stacking. Following Ratledge et al. (2022), we add an additional penalty term to the standard loss function of the ridge regression, which comprises of the mean squared error (MSE) plus an L_2 penalty. The adjusted loss function is thus $MSE + \lambda_l L_2 + \lambda_q Q_{bias}$, where λ_q is the strength of the additional penalty and a hyperparameter that can be tuned. Q_{bias} is the maximum of the squared quintile specific biases, equal to $\max_j(\mathbb{E}[\hat{y}_i - y_i | y_i \in Q_j]^2)$, where $Q_j \in \{Q_1, \dots, Q_5\}$, and $\hat{y}_i$ is the predicted y for observation i. The figure shows the effect of three λ_q parameters on the prediction bias. Solid lines indicate the best linear fit of each model, while dashed black lines represent the expected fit without bias ($\beta_1 = 1$).

With these considerations in mind, we now compare the downstream correlations (Appendix Figure A6) and regression results (Table 2) of $\widehat{edu}$ and $econ$ with the true correlations captured by edu. As Figure A6 in the Appendix shows, the predicted education indicator consistently understates true correlations, suggesting that the attenuation bias dominates over a potential bias due to correlated learning. Table 2 further shows that the slope of the regression coefficients is considerably underestimated for all outcomes when using $\widehat{edu}$ as the dependent variable of the regression and slightly overestimated in the reverse specification, a pattern that is consistent with a Berkson-type error. Hence, it appears that the correlation estimates are mainly affected by attenuation, while biases in regression coefficients are largely driven by a Berkson-type error.

Table 2: Downstream regression results

	Mexico				United States			
	Years of Schooling	Post-Basic	Secondary	Primary	Years of Schooling	Bachelor	College	High School
β_t: $edu \sim econ$	0.740	0.661	0.703	0.728	0.692	0.707	0.655	0.487
	(0.014)	(0.015)	(0.014)	(0.014)	(0.013)	(0.013)	(0.013)	(0.016)
β_p: $\widehat{edu} \sim econ$	0.549	0.499	0.526	0.516	0.496	0.526	0.470	0.320
	(0.013)	(0.014)	(0.012)	(0.012)	(0.011)	(0.012)	(0.011)	(0.011)
β_c: $\widehat{edu_c} \sim econ$	0.748	0.651	0.744	0.687	0.699	0.727	0.661	0.362
	(0.017)	(0.018)	(0.018)	(0.016)	(0.016)	(0.017)	(0.018)	(0.013)
$\beta_t - \beta_p$	-0.191	-0.161	-0.177	-0.212	-0.196	-0.181	-0.185	-0.167
	(0.012)	(0.011)	(0.012)	(0.014)	(0.011)	(0.012)	(0.011)	(0.012)
$\beta_t - \beta_c$	0.008	-0.010	0.041	-0.042	0.007	0.021	0.005	-0.125
	(0.014)	(0.013)	(0.015)	(0.016)	(0.014)	(0.016)	(0.017)	(0.014)
β_t: $econ \sim edu$	0.740	0.661	0.703	0.728	0.692	0.707	0.656	0.488
	(0.014)	(0.015)	(0.014)	(0.014)	(0.013)	(0.013)	(0.013)	(0.016)
β_p: $econ \sim \widehat{edu}$	0.794	0.717	0.826	0.863	0.765	0.738	0.767	0.640
	(0.018)	(0.019)	(0.019)	(0.019)	(0.017)	(0.017)	(0.018)	(0.023)
β_c: $econ \sim \widehat{edu_c}$	0.577	0.539	0.564	0.646	0.535	0.520	0.443	0.515
	(0.013)	(0.015)	(0.013)	(0.015)	(0.012)	(0.012)	(0.012)	(0.019)
$\beta_t - \beta_p$	0.054	0.056	0.123	0.135	0.072	0.031	0.111	0.152
	(0.012)	(0.012)	(0.013)	(0.014)	(0.015)	(0.014)	(0.014)	(0.025)
$\beta_t - \beta_c$	-0.162	-0.122	-0.139	-0.083	-0.157	-0.187	-0.212	0.028
	(0.010)	(0.011)	(0.011)	(0.012)	(0.013)	(0.012)	(0.012)	(0.024)
N	2,457	2,457	2,457	2,457	3,140	3,140	3,140	3,140

The predictions for different educational outcomes, referred to as *edu*, are represented as $\widehat{edu}$, and *econ* is wealth for Mexico and income for the United States. For $\widehat{edu_c}$, we apply a Berkson error correction with λ_q = 3 for years of schooling and λ_q = 15 for all other outcomes (i.e., all percentages). Results are reported in standard deviations ($\widehat{edu}$ and $\widehat{edu_c}$ are standardized using the distribution of *edu*). $\beta_t - \beta_p$ is the original bias and $\beta_t - \beta_p$ is the bias using the predictions based on the adapted loss function. Education is the dependent variable in the upper panel and the independent variable in the lower panel. Standard errors in parentheses.

While in a typical application, we would be unable to quantify the extent of the attenuation bias or avoid correlated learning, it is possible to refine our model in a way that minimizes the Berkson error. Following Ratledge et al. (2022), we add a further penalty term for a quintile-specific bias to the loss function of our final stacking model. If the weight given to this penalty is sufficiently high, the tendency to understate high and overstate low values effectively disappears (see Figure 7), but this comes at the expense of lower overall performance with a decrease in the r^2 by about 10 percentage points. When using this new set of predictions (see Table 2), the bias in the upper panel ($\widehat{edu} \sim econ$) becomes negligible for most outcomes.[20] In the lower panel ($econ \sim \widehat{edu}$) the direction of the bias is reversed as the attenuation bias starts to dominate. This suggests that when appropriately modeled, predicted indicators can produce correct estimates in downstream regression tasks as long as they serve as the outcome and not the treatment variable. Luckily, the former constitutes a much more likely use case, as, for example, it allows to evaluate the effect of interventions or policy changes.

4 Conclusion

Our results show that human capital can be accurately inferred from Twitter data using machine learning. We are able to account for 70 percent of the variation of years of schooling in Mexico and 65 percent in the United States. This is substantially higher than the performance reported in previous attempts to predict human capital, and comparable to the effectiveness of satellite data in predicting wealth. As only a few days of Twitter data are needed to achieve a good performance and the natural language processing tools we use for feature preparation support many different languages, our approach is widely applicable and scalable.

Despite the lower Twitter penetration, our model tends to perform better for Mexico than for the United States, suggesting our approach is also relevant for less affluent regions with lower levels of social media usage. However, within countries, Twitter data appears to be less informative at the lower end of the education distribution. Similarly, the model performs worse in less-populated areas with lower Twitter penetration. An intuitive explanation is that Twitter use is concentrated among the highly educated and thus not particularly well-suited for distinguishing between low and medium levels of education. Including data from other platforms with less selective usage patterns might thus be a promising avenue for future research aiming to further improve predictive performance, particularly in developing countries.

Apart from being directly useful to understand spatial patterns and target interventions, predicted indicators also have the potential to advance scientific research by providing inputs for downstream inference tasks. This paper shows that such applications do not come without caveats. Our data and simulations show that estimates in downstream regression tasks tend to be subject to several biases. We further demonstrate, that these biases can be corrected using an adapted loss function (see Ratledge et al., 2022) if the predicted indicator acts

[20] The bias becomes insignificant for 5 out of 8 outcomes. The correction appears to be particularly effective for outcomes that have a higher initial r^2. In the last model (high school), which is also the one with the lowest initial r^2, the penalized loss function achieves only a limited slope correction under $\lambda_q = 15$ (not shown) and the regression is thus unable to recover the true effect.

as the dependent variable. If carefully tuned, machine learning derived indicators can thus become a valuable data source to study effects on outcomes for which ground truth data are unavailable. However, more research is needed to better understand the empirical relevance of each of the biases, and experiment with the most effective ways of approaching them.

References

Aiken, Emily, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua E. Blumenstock (2022). "Machine learning and phone data can improve targeting of humanitarian aid". *Nature* 603.7903, 864–870.

Barro, Robert J. and Jong Wha Lee (2013). "A new data set of educational attainment in the world, 1950–2010". *Journal of Development Economics* 104, 184–198.

Blumenstock, Joshua E., Gabriel Cadamuro, and Robert On (2015). "Predicting poverty and wealth from mobile phone metadata". *Science* 350.6264, 1073–1076.

Bollen, Johan, Huina Mao, and Xiaojun Zeng (2011). "Twitter mood predicts the stock market". *Journal of Computational Science* 2.1, 1–8.

Burke, Marshall, Anne Driscoll, David B. Lobell, and Stefano Ermon (2021). "Using satellite imagery to understand and promote sustainable development". *Science* 371.6535, eabe8628.

Burke, Marshall and David B. Lobell (2017). "Satellite-based assessment of yield variation and its determinants in smallholder African systems". *Proceedings of the National Academy of Sciences* 114.9, 2189–2194.

Chang, Serina, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec (2021). "Mobility network models of COVID-19 explain inequities and inform reopening". *Nature* 589.7840, 82–87.

Cohen, Jacob, Patricia Cohen, Stephen G. West, and Leona S. Aiken (2013). *Applied multiple regression/correlation analysis for the behavioral sciences.* Routledge.

Curtis, Brenda, Salvatore Giorgi, Anneke E. K. Buffone, Lyle H. Ungar, Robert D. Ashford, Jessie Hemmons, Dan Summers, Casey Hamilton, and H. Andrew Schwartz (2018). "Can Twitter be used to predict county excessive alcohol consumption rates?" *PloS ONE* 13.4, e0194290.

Fuller, Wayne A. (1987). *Measurement error models.* Wiley Series in Probability and Mathematical Statistics. John Wiley Sons, Inc.

Gebru, Timnit, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei (2017). "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States". *Proceedings of the National Academy of Sciences* 114.50, 13108–13113.

Gómez, Juan Carlos, Luis Miguel López Santamarıa, Mario Alberto Ibarra Manzano, and Dora Luz Almanza Ojeda (2021). "Predicción automática del nivel educativo en usuarios de Twitter en México". *Realidad, datos y espacio Revista internacional de estadıstica y geografıa* 12.1.

Head, Andrew, Mélanie Manguin, Nhat Tran, and Joshua E. Blumenstock (2017). "Can human development be measured with satellite imagery?" *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development.* ICTD '17. Lahore, Pakistan: Association for Computing Machinery.

Huang, Xiao, Zhenlong Li, Yuqin Jiang, Xiaoming Li, and Dwayne Porter (2020). "Twitter reveals human mobility dynamics during the COVID-19 pandemic". *PloS ONE* 15.11, e0241957.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon (2016). “Combining satellite imagery and machine learning to predict poverty”. *Science* 353.6301, 790–794.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch (July 2018). “Marian: Fast neural machine translation in C++”. *Proceedings of ACL 2018, System Demonstrations.* Melbourne, Australia: Association for Computational Linguistics, 116–121.

King, Gary, Jennifer Pan, and Margaret E. Roberts (2013). “How censorship in China allows government criticism but silences collective expression”. *American Political Science Review* 107.2, 326–343.

Kuffer, Monika, Karin Pfeffer, and Richard Sliuzas (2016). “Slums from space—15 years of slum mapping using remote sensing”. *Remote Sensing* 8.6, 455.

Lobell, David B. (2013). “The use of satellite data for crop yield gap analysis”. *Field Crops Research* 143, 56–64.

Mboga, Nicholus, Claudio Persello, John Ray Bergado, and Alfred Stein (2017). “Detection of informal settlements from VHR images using convolutional neural networks”. *Remote Sensing* 9.11, 1106.

Mellon, Jonathan and Christopher Prosser (2017). “Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users”. *Research & Politics* 4.3, 2053168017720008.

Poushter, Jacob, Caldwell Bishop, and Hanyu Chwe (2018). *Social media use continues to rise in developing countries but plateaus across developed ones.* Tech. rep. Pew Research Center.

Ratledge, Nathan, Gabe Cadamuro, Brandon de la Cuesta, Matthieu Stigler, and Marshall Burke (2022). “Using machine learning to assess the livelihood impact of electricity access”. *Nature* 611.7936, 491–495.

Sheehan, Evan, Chenlin Meng, Matthew Tan, Burak Uzkent, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon (2019). “Predicting Economic Development Using Geolocated Wikipedia Articles”. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* KDD ’19. Anchorage, AK, USA: Association for Computing Machinery, 2698–2706.

Smirnov, Ivan (2020). “Estimating educational outcomes from students’ short texts on social media”. *EPJ Data Science* 9.1, 1–11.

Stahel, Werner A. (2000). *Statistische Datenanalyse: Eine Einfuehrung fuer Naturwissenschaftler.* Vieweg+Teubner Verlag Wiesbaden.

Statista (2022). *Number of social media users worldwide from 2017 to 2027.* URL: `https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/` (visited on 02/2023).

Stevens, Forrest R., Andrea E. Gaughan, Catherine Linard, and Andrew J. Tatem (2015). “Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data”. *PLoS ONE* 10.2, e0107042.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn (2007). “Bias in random forest variable importance measures: Illustrations, sources and a solution”. *BMC Bioinformatics* 8.1, 1–21.

Sun, Jie, Liping Di, Ziheng Sun, Yonglin Shen, and Zulong Lai (2019). "County-level soybean yield prediction using deep CNN-LSTM model". *Sensors* 19.20, 4363.

Ushio, Asahi and Jose Camacho-Collados (2022). "TweetNLP: Cutting-edge natural language processing for social media". *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Abu Dhabi, U.A.E.: Association for Computational Linguistics.

Wardrop, Nicola A., Warren C. Jochem, Tomas J. Bird, Heather R. Chamberlain, Donna J. Clarke, David Kerr, Linus Bengtsson, Sabrina Juran, Vincent Seaman, and Andrew J. Tatem (2018). "Spatially disaggregated population estimates in the absence of national population and housing census data". *Proceedings of the National Academy of Sciences of the United States of America* 115, 3529–3537.

Wesolowski, Amy, Nathan Eagle, Andrew J. Tatem, David L. Smith, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee (2012). "Quantifying the impact of human mobility on malaria". *Science* 338.6104, 267–270.

Yeh, Christopher, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke (2020). "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa". *Nature Communications* 11.1, 2583.

Yin, Junjun, Yizhao Gao, and Guangqing Chi (2022). "An evaluation of geo-located Twitter data for measuring human migration". *International Journal of Geographical Information Science* 36.9, 1830–1852.

A Appendix

A.1 Main Results

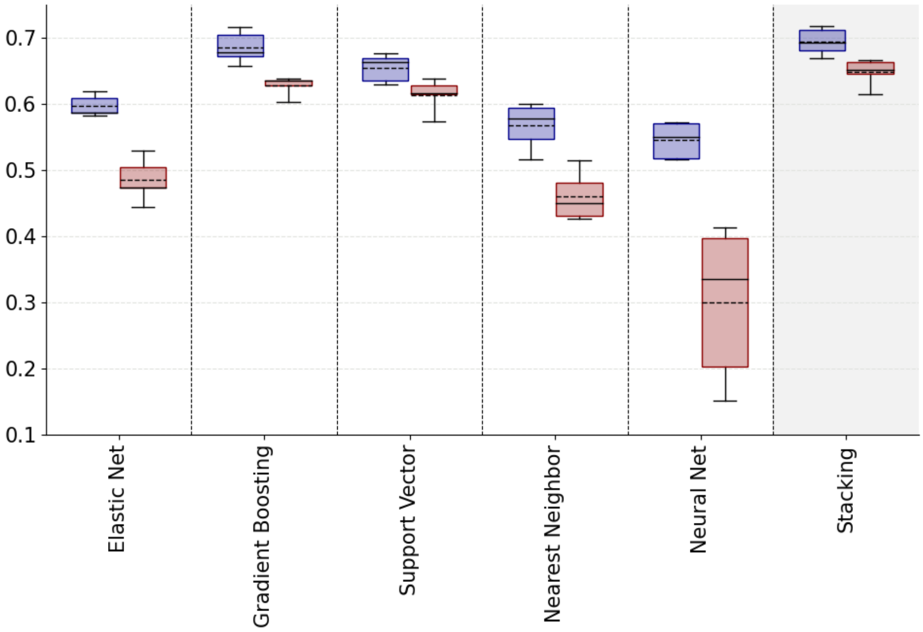


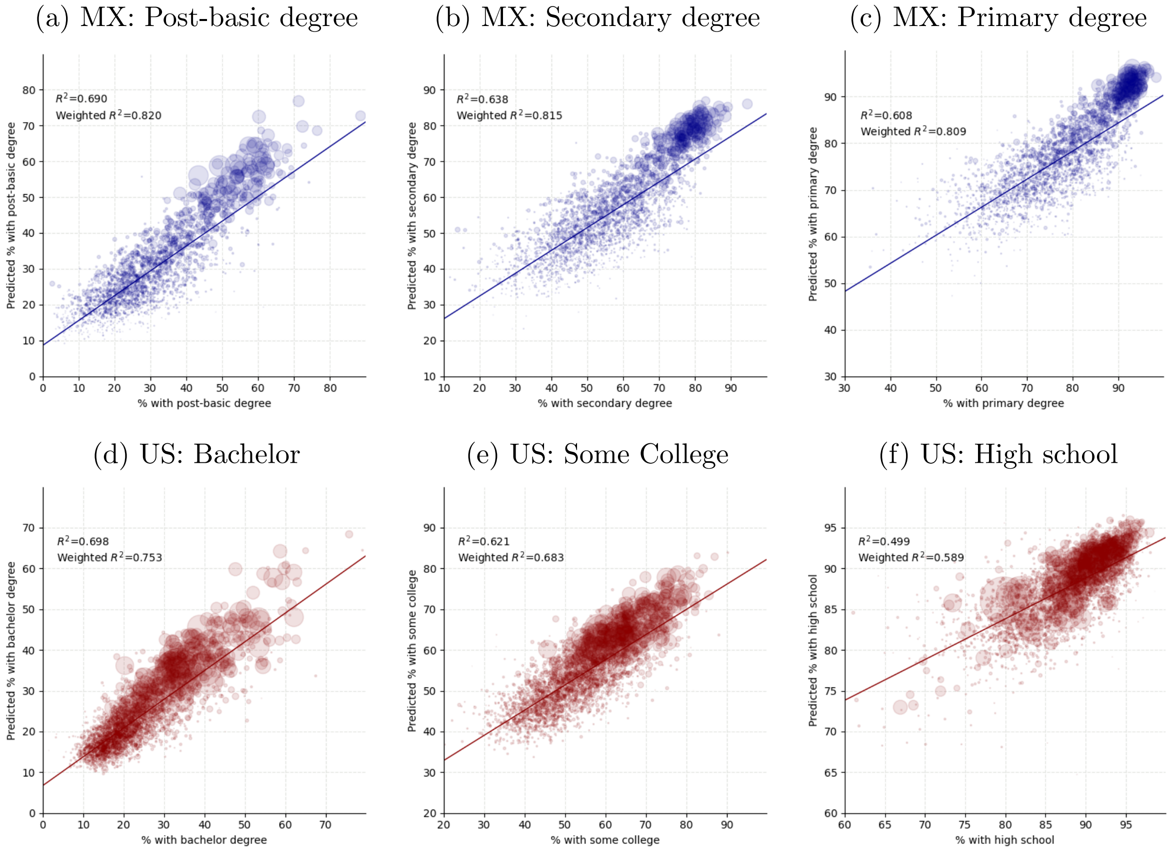
Figure A1: Performance of individual models

Performance of individual models considered in the final stacking model for years of schooling in Mexico (blue) and the United States (red). All models are evaluated through five-fold cross-validation. Boxplots show the median (solid line), mean (dotted line), the 20th & 80th percentile (box limits), as well as the minimum & maximum (whiskers) for the r^2 across validation folds for each outcome and country.

Table A1: Performance of individual models

	Mexico				United States			
	Years of Schooling	Post Basic Education	Secondary Education	Primary Education	Years of Schooling	Bachelor Degree	Some College	Only High School
Elastic Net	0.597 (2.8%)	0.621 (-5.6%)	0.548 (-15.3%)	0.495 (-17.7%)	0.485 (0.2%)	0.522 (-1.4%)	0.461 (4.6%)	0.350 (-4.0%)
Gradient Boosting	0.686 (56.9%)	0.689 (61.6%)	0.638 (62.1%)	0.600 (60.7%)	0.628 (56.5%)	0.674 (52.7%)	0.603 (55.6%)	0.467 (53.0%)
Support Vector Machine	0.655 (29.1%)	0.602 (6.2%)	0.544 (7.7%)	0.459 (-0.1%)	0.614 (41.0%)	0.669 (37.7%)	0.576 (37.1%)	0.457 (42.7%)
Nearest Neighbour Matching	0.567 (0.6%)	0.576 (-0.2%)	0.523 (7.5%)	0.490 (11.5%)	0.461 (3.3%)	0.504 (0.2%)	0.425 (3.3%)	0.359 (15.2%)
Multi-layer Perceptron	0.545 (13.2%)	0.654 (40.3%)	0.590 (41.4%)	0.557 (48.1%)	0.300 (6.2%)	0.627 (16.9%)	0.424 (6.7%)	-0.523 (3.2%)
Stacking	0.695	0.689	0.638	0.607	0.648	0.697	0.620	0.500

Mean r^2 and stacking weights (in parentheses) across folds for different models and outcomes. Note that r^2 values for the final stacking model reported as our main results are computed using the combined out-of-sample predictions of all folds rather than as the mean across folds, and may thus slightly differ.


(a) MX: Post-basic degree

(b) MX: Secondary degree

(c) MX: Primary degree

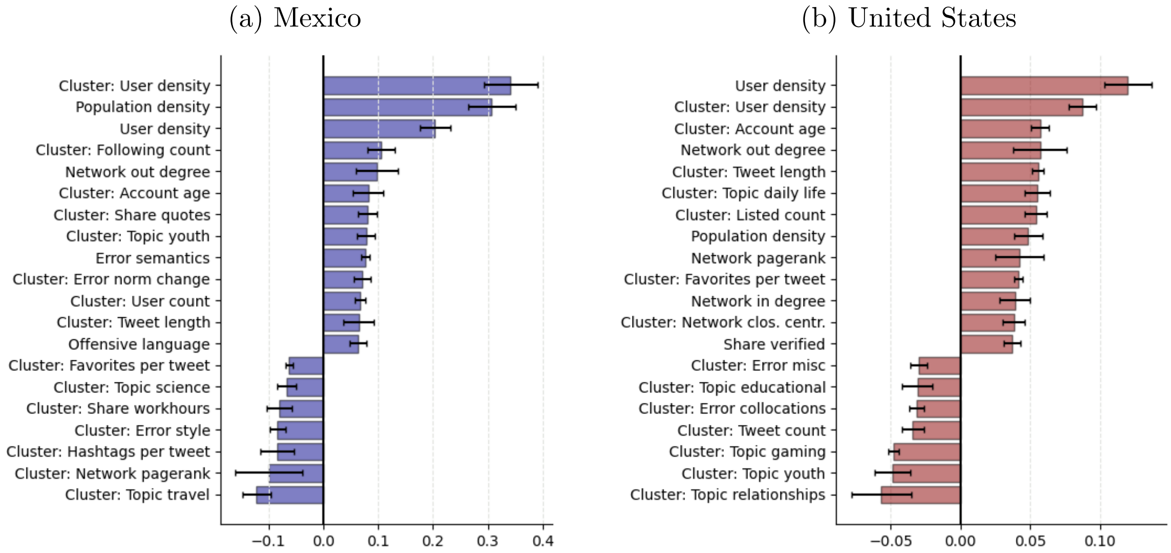
(d) US: Bachelor

(e) US: Some College

(f) US: High school

Figure A2: True vs. predicted values for secondary outcomes

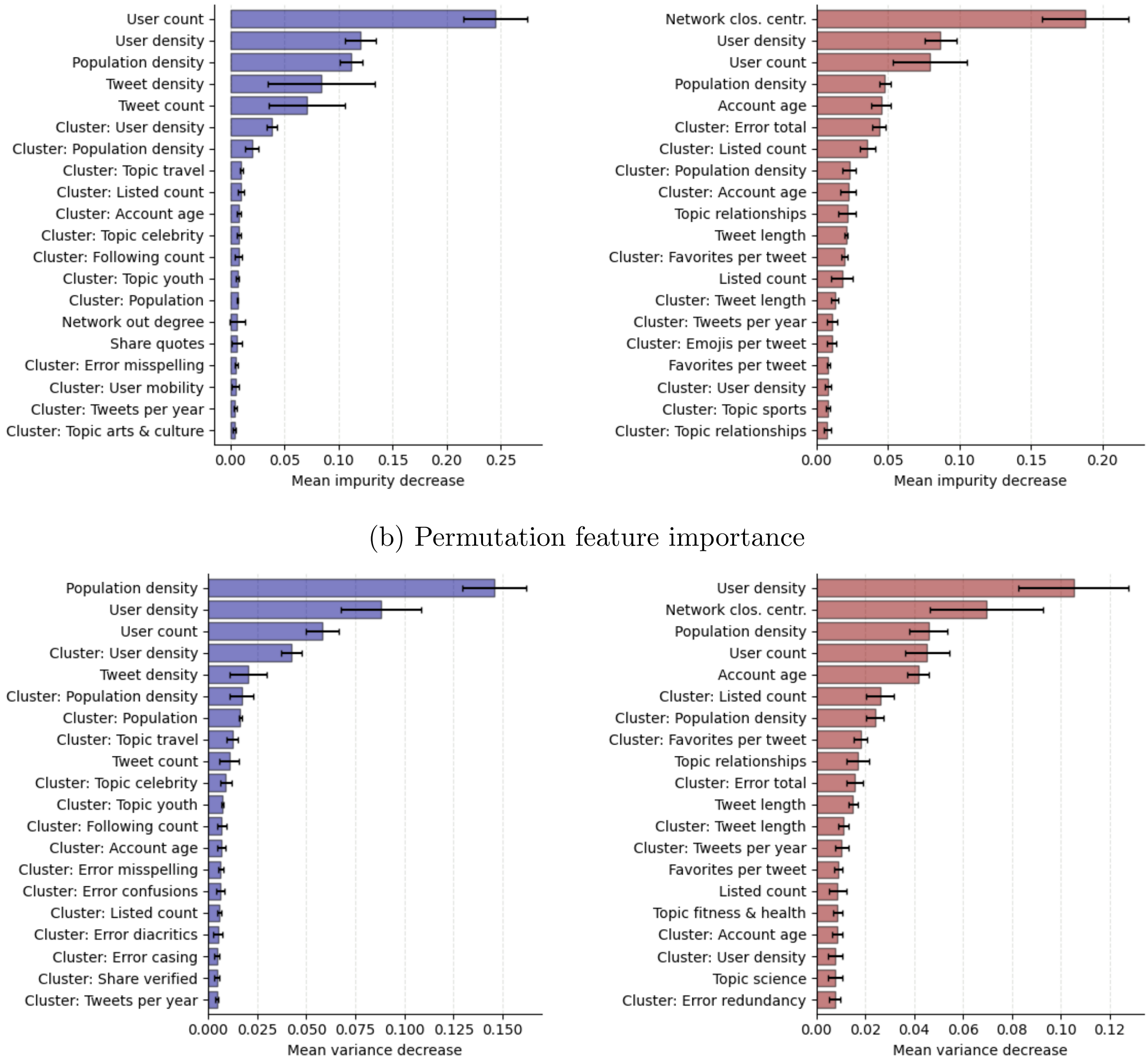
Predicted values for all municipalities and counties are obtained by combining out-of-sample predictions from all folds. Bubble size is proportional to the population in each unit. r^2 and population weighted r^2 shown. Line indicating best linear fit is not population weighted.


(a) Mexico

(b) United States

Figure A3: Feature importance based on elastic net model

Feature importance estimates shown on the x-axis correspond to the standardized regression coefficients in the elastic net model.


(a) Mean impurity decrease

(b) Permutation feature importance

Figure A4: Gradient boosting feature importance

Most important features in gradient boosting regressor for Mexico (blue) and the United States (red). In Figure A4a, feature importances are based on mean impurity decrease. As these can be misleading if features are differently scaled or have varying numbers of categories (Strobl et al., 2007), Figure A4b also presents permutation based feature importances. Note that due to the high correlation between features, estimates should be interpreted with care.

B Bias Correction

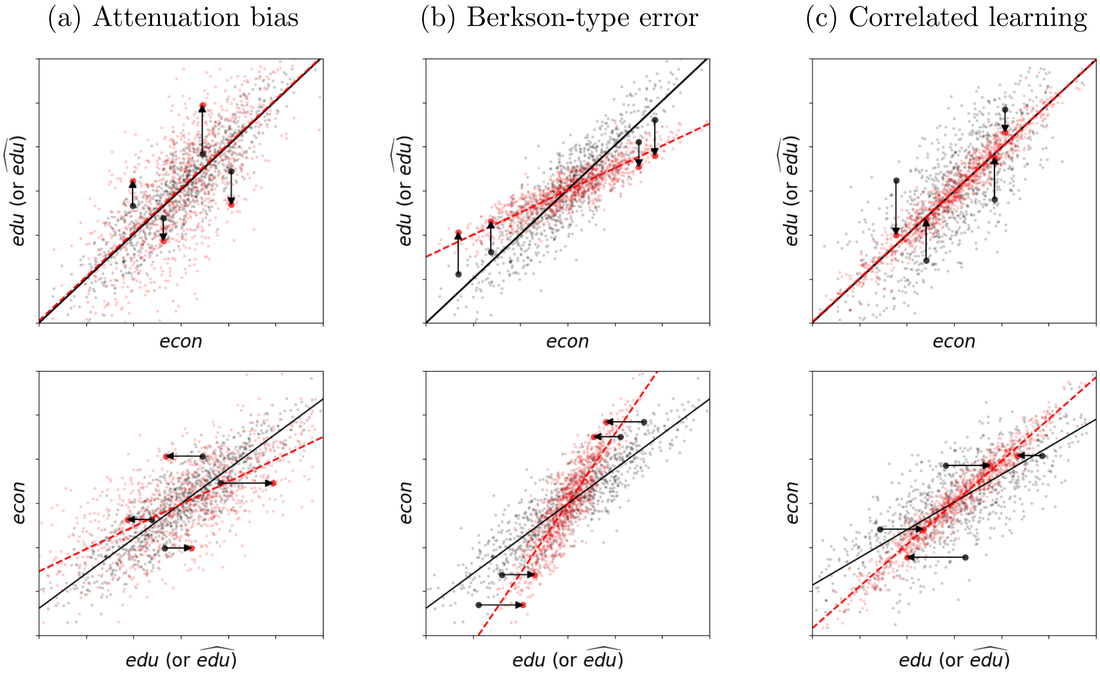


Figure A5: Simulation of different types of biases in downstream regression tasks

Scatter plots and best linear fit for edu (black) and $\widehat{edu}$ (red) with different types of measurement errors. Arrows indicate the movement of typical points as a result of each measurement error. In the upper row, edu (or $\widehat{edu}$) is the outcome of the regression, while it features as the explanatory variable in the lower row.

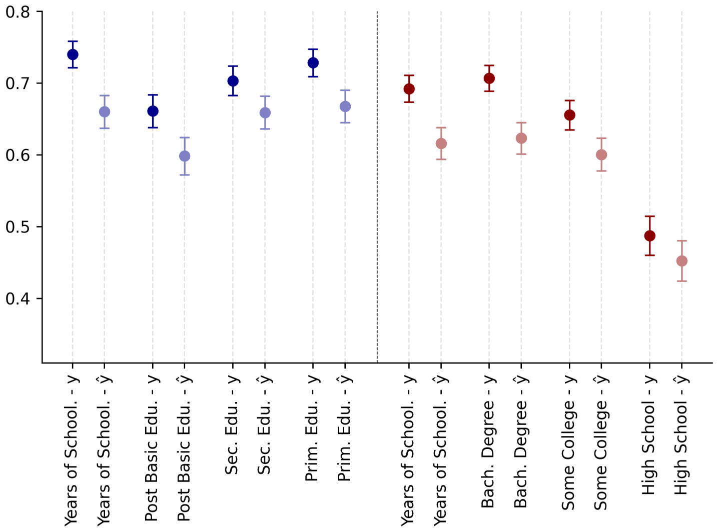


Figure A6: Correlation of observed and predicted education with wealth index and income

Correlations between true and predicted educational outcomes and wealth in Mexico (blue) as well as income in the United States (red). 95% confidence intervals shown.

C Feature Statistics

Table A2: Survey statistics by country

Variable	Country	Mean	SD	Min	Median	Max
Years of Schooling	MX	7.83	1.49	3.40	7.72	14.55
	US	13.30	0.66	9.37	13.28	16.13
Post Basic Education	MX	0.28	0.13	0.03	0.26	0.89
Bachelor Degree	US	22.61	9.71	0.00	20.22	79.14
Secondary Education	MX	0.54	0.14	0.12	0.54	0.95
Some College	US	53.67	10.72	7.41	53.61	90.31
Primary Education	MX	0.76	0.11	0.36	0.76	0.98
High School	US	87.60	6.04	21.85	88.83	98.61
Population	MX	51,173.11	147,322.51	81.00	13,552.00	1,922,523.00
	US	105,661.95	333,146.18	57.00	25,790.00	9,829,544.00
Wealth Index	MX	0.68	0.12	0.07	0.70	0.94
Income	US	57,455.86	14,582.81	22,901.00	55,143.50	160,305.00

Table A3: Twitter penetration and usage statistics by country

Variable	Country	Mean	SD	Min	Median	Max
Tweet count	MX	1,093.25	6,363.61	0.00	8.00	119,126.00
	US	7,195.60	42,124.22	0.00	271.00	1,472,677.00
User count	MX	50.11	269.79	0.00	2.00	5,891.00
	US	299.54	1,548.14	0.00	24.00	52,602.00
Share weekdays	MX	0.70	0.22	0.00	0.72	1.00
	US	0.70	0.15	0.00	0.71	1.00
Share workhours	MX	0.29	0.21	0.00	0.28	1.00
	US	0.31	0.15	0.00	0.31	1.00
Follower count	MX	251.37	1,210.64	0.00	111.83	36,807.50
	US	304.11	754.33	0.00	229.50	24,799.80
Following count	MX	358.47	480.14	0.00	261.71	7,603.16
	US	415.74	556.26	1.00	359.50	25,202.00
Tweet count	US	2,659.14	6,237.16	1.00	1,927.00	183,023.00
	MX	2,405.62	5,565.75	1.00	961.45	79,700.00
User mobility	MX	1.61	0.75	1.00	1.50	10.00
	US	1.76	0.80	1.00	1.71	32.00
iPhone share	US	0.62	0.23	0.00	0.67	1.00
	MX	0.28	0.27	0.00	0.25	1.00
Instagram share	US	0.21	0.24	0.00	0.12	1.00
	MX	0.14	0.23	0.00	0.06	1.00
Favorites per tweet	US	1.73	8.74	0.00	1.35	463.46
	MX	3.76	23.95	0.00	1.25	892.38
Tweets per year	US	495.19	1,703.97	0.39	325.13	52,472.55
	MX	841.93	6,183.31	0.82	260.06	210,975.91
Account age	MX	5.67	2.67	-0.03	5.88	13.56
	US	7.06	1.69	-0.02	7.10	14.65

Variable	Country	Mean	SD	Min	Median	Max
Account age	MX	5.67	2.67	-0.03	5.88	13.56
	US	7.06	1.69	-0.02	7.10	14.65
Listed count	MX	2.86	7.79	0.00	1.00	151.40
	US	12.39	30.79	0.00	6.67	711.60
Followers per following	MX	0.64	2.39	0.00	0.38	68.93
	US	0.71	1.60	0.00	0.60	61.70
Share quotes	MX	0.07	0.11	0.00	0.04	1.00
	US	0.09	0.07	0.00	0.10	1.00
Share replies	MX	0.24	0.22	0.00	0.23	1.00
	US	0.22	0.13	0.00	0.24	1.00
Share verified	MX	0.00	0.03	0.00	0.00	1.00
	US	0.01	0.03	0.00	0.00	1.00
Tweet length	MX	69.94	29.48	4.00	67.53	274.00
	US	80.86	21.74	6.00	79.01	275.00
Hashtags per tweet	MX	0.30	0.51	0.00	0.17	8.00
	US	0.38	0.52	0.00	0.28	8.00
Mentions per tweet	US	0.53	0.31	0.00	0.56	4.44
	MX	0.45	0.45	0.00	0.41	7.00
Urls per tweet	US	0.38	0.52	0.00	0.28	8.00
	MX	0.30	0.51	0.00	0.17	8.00
Emojis per tweet	MX	0.81	0.70	0.00	0.72	6.67
	US	0.55	0.54	0.00	0.50	15.00

Table A4: Error statistics by country

Variable	Country	Mean	SD	Min	Median	Max
Error typography	MX	7.71	9.03	0.00	6.57	170.57
	US	2.55	2.47	0.00	2.18	30.61
Error grammar	MX	0.17	0.54	0.00	0.00	9.80
	US	0.55	0.82	0.00	0.43	15.87
Error confusions	MX	0.14	0.51	0.00	0.00	11.11
	US	0.10	0.26	0.00	0.04	7.44
Error casing	MX	1.29	3.49	0.00	0.18	55.56
	US	1.73	2.74	0.00	1.29	60.61
Error misc	MX	0.12	1.22	0.00	0.00	47.15
	US	0.19	0.44	0.00	0.13	14.71
Error style	US	0.43	0.92	0.00	0.30	23.81
	MX	0.02	0.24	0.00	0.00	7.19
Error repetitions style	US	0.00	0.00	0.00	0.00	0.01
	MX	0.00	0.00	0.00	0.00	0.08
Error semantics	US	0.01	0.11	0.00	0.00	4.57
	MX	0.01	0.06	0.00	0.00	1.94
Error variants	US	0.01	0.06	0.00	0.00	2.51
	MX	0.07	0.72	0.00	0.00	25.64
Error punctuation	US	1.02	1.39	0.00	0.91	32.26
	MX	0.66	1.83	0.00	0.24	35.51
Error typos	US	7.19	5.94	0.00	6.78	181.82
	MX	12.47	13.82	0.00	10.16	285.71
Error total	MX	25.28	17.43	0.00	22.87	285.71
	US	13.87	7.68	0.00	13.16	181.82

Variable	Country	Mean	SD	Min.	Median	Max.
Error expressions	MX	0.02	0.15	0.00	0.00	2.82
Error redundancy	MX	0.00	0.01	0.00	0.00	0.23
Error prepositions	MX	0.00	0.00	0.00	0.00	0.04
Error verb agreement	MX	0.02	0.27	0.00	0.00	10.42
Error misspelling	MX	1.18	4.04	0.00	0.50	125.00
Error proper nouns	MX	0.00	0.01	0.00	0.00	0.34
Error diacritics	MX	1.07	1.60	0.00	0.68	16.67
Error context	MX	0.00	0.01	0.00	0.00	0.38
Error repetitions	MX	0.01	0.14	0.00	0.00	5.38
Error norm change	MX	0.07	0.28	0.00	0.00	5.56
Error noun agreement	MX	0.26	0.80	0.00	0.02	14.93
Error collocations	US	0.01	0.20	0.00	0.00	8.20
Error nonstandard	US	0.00	0.02	0.00	0.00	0.49
Error redundancy	US	0.05	0.42	0.00	0.01	18.87
Errror compounding	US	0.02	0.13	0.00	0.00	5.32

Table A5: Topic statistics by country

Variable	Country	Mean	SD	Min	Median	Max
Topic arts & culture	MX	3.87	2.28	0.30	3.39	24.89
	US	2.97	1.54	0.27	2.68	29.31
Topic business	MX	3.00	3.15	0.26	2.33	46.23
	US	2.79	2.10	0.24	2.55	36.69
Topic celebrity	MX	6.83	5.06	0.31	6.33	49.20
	US	6.10	2.95	0.37	6.33	42.41
Topic daily life	MX	26.03	8.28	1.22	25.48	59.21
	US	21.30	5.91	1.26	20.74	60.28
Topic family	MX	4.03	3.09	0.32	3.52	32.72
	US	3.99	1.89	0.30	3.74	33.24
Topic fashion	MX	1.33	1.47	0.20	1.00	18.35
	US	1.67	1.49	0.19	1.49	38.24
Topic films	MX	4.12	4.23	0.36	3.38	64.47
	US	4.25	3.46	0.26	4.04	67.58
Topic fitness & health	US	2.42	1.61	0.31	2.33	37.45
	MX	2.38	2.36	0.22	1.88	31.72
Topic food & dining	US	3.12	3.35	0.16	2.72	48.61
	MX	2.29	3.29	0.15	1.42	47.21
Topic gaming	MX	1.47	1.55	0.17	1.14	31.18
	US	2.24	1.31	0.29	2.07	19.91

Variable	Country	Mean	SD	Min	Median	Max
Topic educational	MX	1.45	1.65	0.15	1.10	24.93
	US	1.80	1.65	0.16	1.52	30.81
Topic music	MX	4.67	5.57	0.13	3.74	69.11
	US	4.01	3.14	0.12	3.85	66.19
Topic news	MX	13.00	9.50	0.34	11.69	83.39
	US	12.29	5.85	0.40	12.00	80.56
Topic hobbies	MX	7.31	3.21	0.49	6.92	34.64
	US	5.12	2.06	0.34	4.82	26.46
Topic relationships	MX	6.27	3.70	0.29	5.78	27.32
	US	4.76	1.98	0.30	4.58	21.19
Topic science	MX	1.87	2.79	0.23	1.22	54.54
	US	1.69	1.71	0.29	1.49	46.50
Topic sports	MX	5.87	6.44	0.31	4.54	65.84
	US	15.14	9.31	0.12	14.36	85.78
Topic travel	MX	3.28	3.37	0.24	2.31	30.65
	US	2.95	2.37	0.23	2.19	30.52
Topic youth	MX	0.93	1.26	0.18	0.68	22.34
	US	1.40	1.34	0.18	1.15	23.65

Table A6: Sentiment statistics by country

Variable	Country	Mean	SD	Min	Median	Max
Sentiment negative	MX	0.16	0.12	0.00	0.16	0.95
	US	0.16	0.08	0.00	0.17	0.91
Sentiment positive	MX	0.38	0.18	0.01	0.37	0.99
	US	0.50	0.13	0.01	0.48	0.99
Hate speech	MX	0.04	0.03	0.01	0.04	0.42
	US	0.05	0.02	0.01	0.04	0.33
Offensive language	MX	0.15	0.07	0.03	0.15	0.89
	US	0.16	0.06	0.03	0.16	0.83

Table A7: Network statistics by country

Variable	Country	Mean	SD	Min	Median	Max
Network in degree	MX	0.14	0.87	0.00	0.00	15.17
	US	0.36	2.17	0.00	0.01	67.43
Network out degree	MX	0.14	0.79	0.00	0.00	14.65
	US	0.36	1.91	0.00	0.01	56.15
Network clos. centr.	MX	0.16	0.18	0.00	0.00	0.55
	US	0.34	0.16	0.00	0.40	0.68
Network pagerank	MX	0.00	0.00	0.00	0.00	0.06
	US	0.00	0.00	0.00	0.00	0.05

D Feature Descriptions

Table A8: Survey indicator description

Label	Description
Years of Schooling	Average years of schooling in municipality (MX) or county (US) according to census. We approximate years of schooling for the US by attainment statistics (see main text)
Post Basic Education	Share of population with post basic education
Secondary Education	Share of population with secondary education
Primary Education	Share of population with primary education
Wealth Index	Index based on share of households that have 13 wealth related items according to the Mexican census, sum across standardized items
Bachelor Degree	Share of county level population with some college level education
Some College	Share of population with a bachelor degree
High School	Share of population with high school education
Income	Income statistics provided by US census
Population	Population counts according to census

Table A9: Network indicator description

Label	Description
Network in degree	Number outgoing references measured by mentions and quotes (log scale)
Network out degree	Number incoming references measured by mentions and quotes (log scale)
Network clos. centr.	Pagerank for municipalities (MX) or counties (US) according to respective network based on mentions and quotes (log scale)
Network pagerank	Closeness centrality for municipalities (MX) or counties (US) according to respective network based on mentions and quotes (log scale)

Table A10: Twitter penetration and usage indicator description

Label	Description
Tweet count	Number of tweets
User count	Number of users
Share weekdays	Share of tweets created during weekdays (Monday-Friday)
Share workhours	Share of tweets created during workhours (Monday-Friday, 8:00am-4:00pm))
Follower count	Median number of followers per user (log scale)
Following count	Median number of friends per user (log scale)
Tweet count	Median number of tweets per user (log scale)
User mobility	Average number of municipalities (MX) or counties (US) users tweet from (log scale)
iPhone share	Share of tweets sent from an iPhone
Instagram share	Share of tweets sent via Instagram (log scale)
Favorites per tweet	Number of likes per tweet, median (log scale)
Tweets per year	Median number of tweets per year (log scale)
Account age	Age of average account

Table A11: Twitter penetration and usage indicator description

Label	Description
Account age	Age of average account
Listed count	Average number of public lists user is a member of (log scale)
Followers per following	Number of followers divided by number of accounts a user follows, median (log scale)
Share quotes	Share of tweets that are quotes (log scale)
Share replies	Share of tweets that are replies (log scale)
Share verified	Share of verified users (log scale)
Tweet length	Average number of characters per tweet (log scale)
Hashtags per tweet	Average number of hashtags per tweet (log scale)
Mentions per tweet	Average number of mentions per tweet (log scale)
Urls per tweet	Average number of urls per tweet (log scale)
Emojis per tweet	Number of emoji per tweet (log scale)

Table A12: Error indicator description (countries' joint errors)

Label	Description
Error total	Number of errors per character (log scale)
Error casing	Casing error (log scale)
Error confusions	Word confusions (log scale)
Error grammar	Grammar error (log scale)
Error variants	Errors regarding American and British English (log scale)
Error misc	Miscellaneous error (log scale)
Error punctuation	Punctuation error (log scale)
Error repetitions style	Style error related to repetitions (log scale)
Error semantics	Semantic error (log scale)
Error style	Style error (log scale)
Error typography	Typography error (log scale)
Error typos	Typo (log scale)

Table A13: Error indicator description (countries' disjoint errors)

Label	Description
Error noun agreement	Noun verb agreement error (log scale)
Error verb agreement	Verb subject agreement error (log scale)
Error norm change	Deviation from linguistic norms (log scale)
Error collocations	Collocation error (log scale)
Error compounding	Compounding error (log scale)
Error context	Context dependent error (log scale)
Error diacritics	Errors regarding accents (diacritic marks, log scale)
Error expressions	Incorrect expression (log scale)
Error misspelling	Misspelling (log scale)
Error nonstandard	Error related to non-standard English (log scale)
Error prepositions	Error related to prepositions (log scale)
Error proper nouns	Error related to proper nouns (log scale)
Error redundancy	Redundancy in text (log scale)
Error redundancy	Redundancy in text (log scale)
Error repetitions	Repetition in text (log scale)

Table A14: Topic indicator description

Label	Description
Topic arts & culture	Share of tweets classified into the arts & culture topic (log scale)
Topic business	Share of tweets classified into the business & entrepreneurs topic (log scale)
Topic celebrity	Share of tweets classified into the celebrity & pop culture topic (log scale)
Topic daily life	Share of tweets classified into the diaries & daily life topic (log scale)
Topic family	Share of tweets classified into the family topic (log scale)
Topic fashion	Share of tweets classified into the fashion & style topic (log scale)
Topic films	Share of tweets classified into the films, tv & video topic (log scale)
Topic fitness & health	Share of tweets classified into the fitness & health topic (log scale)
Topic food & dining	Share of tweets classified into the food & dining topic (log scale)
Topic gaming	Share of tweets classified into the gaming topic (log scale)

Table A15: Topic indicator description

Label	Description
Topic educational	Share of tweets classified into the learning & educational topic (log scale)
Topic music	Share of tweets classified into the music topic (log scale)
Topic news	Share of tweets classified into the news & social concern topic (log scale)
Topic hobbies	Share of tweets classified into the other hobbies topic (log scale)
Topic relationships	Share of tweets classified into the relationships topic (log scale)
Topic science	Share of tweets classified into the science & technology topic (log scale)
Topic sports	Share of tweets classified into the sports topic (log scale)
Topic travel	Share of tweets classified into the travel & adventure topic (log scale)
Topic youth	Share of tweets classified into the youth & student life topic (log scale)

Table A16: Sentiment indicator description

Label	Description
Sentiment negative	Average share of tweets with negative sentiment in contrast to positive and neutral
Sentiment positive	Average share of tweets with positive sentiment in contrast to negative and neutral
Hate speech	Score indicating hate speech, average (log scale)
Offensive language	Score indicating offensive language, average (log scale)

Chapter 2

Inadequate Teacher Content Knowledge and What Could Be Done About It: Evidence from El Salvador

JOURNAL OF DEVELOPMENT EFFECTIVENESS
https://doi.org/10.1080/19439342.2023.2213681

Routledge
Taylor & Francis Group

Inadequate teacher content knowledge and what could be done about it: evidence from El Salvador

Aymo Brunetti[a], Konstantin Büchel[a], Martina Jakob[b], Ben Jann[b] and Daniel Steffen[c]

[a]Department of Economics, University of Bern, Bern, Switzerland; [b]Institute of Sociology, University of Bern, Bern, Switzerland; [c]Institute of Financial Services, Lucerne University of Applied Sciences and Arts, Luzern, Switzerland

ABSTRACT

Good teachers are the backbone of a successful education system. Yet, in developing countries, teachers' content knowledge is often inadequate. This study documents that primary school maths teachers in the department of Morazán in El Salvador only master 47 percent of the curriculum they teach. In a randomised controlled trial with 175 teachers, we further evaluate a computer-assisted learning (CAL) approach to address this shortcoming. After a five months in-service training combining CAL-based self-studying with monthly workshops, participating teachers outperformed their peers from the control group by 0.29σ, but this effect depreciated by 72 percent within one year. Our simulations show that the program is unlikely to be as cost-effective as CAL interventions directly targeting students.

ARTICLE HISTORY
Received 28 September 2022
Accepted 31 March 2023

KEYWORDS
Education quality; teacher performance; teacher training; student learning; basic maths education; computer-assisted learning

1 Introduction

In light of the persistently low learning levels in many developing countries, it is critical to gain a better understanding of the binding constraints to effective teaching. While various aspects of educational systems such as material inputs, pedagogical practices or teacher incentives have been extensively studied (e.g. Kremer, Brannen, and Glennerster 2013; Glewwe and Karthik 2016), one indispensble precondition to successful instruction has been largely neglected: teachers' content knowledge. Consequently, little is known about the extent to which teachers master the curriculum they have to convey to their students and how to effectively narrow potential knowledge gaps. This paper addresses both questions (see Figure 1). In the first part of the study, we assess the content knowledge of Salvadoran maths teachers based on a representative sample of primary school teachers in the department of Morazán. In the second part of the study, we experimentally evaluate an intervention aiming to improve teachers' content knowledge through computer-assisted learning (CAL).

Recent evidence suggests that many primary school teachers may not possess sufficient mastery of the concepts they have to teach. For a sample covering seven sub-Saharan nations, Bold et al. (2017) asked teachers to mark mock tests and then estimated that only two-thirds of primary school teachers possess minimum proficiency in their subject. In *the first part of our study*, we directly measure teachers' content knowledge through an exam-type assessment with a representative sample of 224 primary school maths teachers in the department of Morazán in El Salvador. The average primary school teacher in our sample was able to answer 47 percent of grade two to grade six questions correctly and only 14 percent of the teachers possessed minimum subject proficiency as defined by Bold et al. (2017). For

CONTACT Konstantin Büchel konstantin.buechel@unibe.ch Department of Economics, Schanzeneckstrasse 1, University of Bern, Bern CH-3001, Switzerland

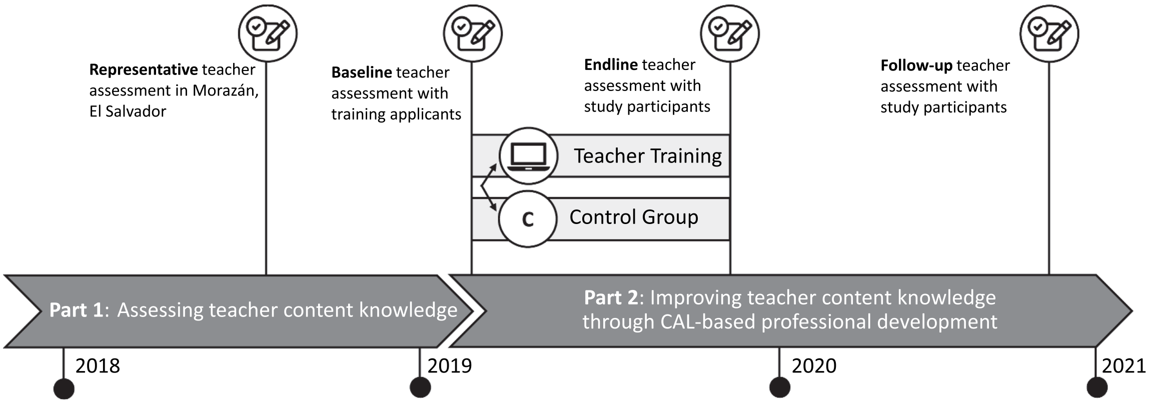

Representative teacher assessment in Morazán, El Salvador
Baseline teacher assessment with training applicants
Endline teacher assessment with study participants
Follow-up teacher assessment with study participants
Teacher Training
C
Control Group
Part 1: Assessing teacher content knowledge
Part 2: Improving teacher content knowledge through CAL-based professional development
2018
2019
2020
2021

Figure 1. Timeline of the study.

example, 43 percent of the teachers correctly computed the area of a rectangle, 36 percent were able to add two fractions and a mere 25 percent could retrieve information from a descriptive chart.

Teachers' content knowledge matters. Previous findings suggest that a 1σ increase in teacher content knowledge is associated with a 0.09σ gain in annual student learning (Metzler and Woessmann 2012; Bau and Das 2020). Bold et al. (2019) further document that gaps in the content knowledge of African teachers account for 30 percent of the shortfalls in student learning relative to the curriculum. But how can teachers' subject mastery be improved? Unfortunately, there is little evidence on how to strengthen teacher content knowledge and the impact thereof.[1] A growing strand of literature documents the success of technology-based instruction with students (for reviews see Escueta et al. 2020; Rodriguez-Segura 2022), and the targeted use of technology may also entail considerable advantages for teacher professional development. For instance, a successful CAL-based training would be relatively easy to replicate and scale, and hence mitigate concerns about the sensitivity of program effectiveness to details of design and implementation (see Kerwin and Thornton 2021).

To assess the value of CAL software in teacher professional development, *the second part of this study* presents a randomised controlled trial implemented with 175 maths teachers in El Salvador. The treatment consisted of a five-month in-service teacher training program that combined CAL-based self-studying with monthly revision workshops. The self-studying modules were financially incentivised CAL-assignments based on learning videos and quizzes developed by KHAN ACADEMY and administered via the offline application KOLIBRI. To measure teachers' content knowledge, we conducted assessments based on the local primary school maths curriculum before, shortly after, and one year after the program. An initially planned student assessment one year after the program's conclusion could not be carried out due to the COVID-19 pandemic.

We find that immediately after the intervention, program teachers outperformed their peers from the control group by 0.29σ or 5.52 percentage points ($p < 0.01$), but this effect diminished by 72 percent one year later. The data further reveals sizeable heterogeneity in treatment effects. In the short term, the program was particularly successful in raising test scores among teachers under 40 (0.53σ, $p < 0.01$), and regarding more advanced concepts from grades five to six (0.31–0.35σ, $p < 0.01$), but even these effects became insignificant after one year.

Investments in teacher competencies have the potential to be highly cost-effective. If teachers retain their acquired skills, further student cohorts will benefit after the intervention period. This stands in contrast to student-centred interventions such as remedial CAL classes that often require continued investments. A unique feature of this study is that the results from the CAL-based teacher training can be directly compared to findings from student-centred CAL lessons implemented in the same context and by the same NGO. Based on the parameters obtained through our experiment, we simulate the long-term cost effectiveness of our teacher intervention and compare it to the cost-effectiveness of remedial CAL lessons we experimentally evaluated with third to sixth graders (see Büchel et al. 2022). Our benchmark findings indicate that an annual retention rate of at least 55 percent among treated teachers is required so that the CAL training with teachers would be more effective than CAL lessons with students. In our experiment, we observe a retention rate of 28 percent, suggesting that the long-term effectiveness of the teacher program is lower than that of the student intervention.

The high depreciation rate of effects at the teacher level is in line with the sparse evidence on this topic. Bando and Xia (2014) find substantial gains in competencies of Mexican teachers from a intensive training in English skills and instructional methods, but the gap between the treatment group and the control group faded after 12 months. Similarly, Cilliers et al. (2019, 2020) report substantial short term gains for two pedagogy-centred teacher training programs, but only for one of the programs these effects were found to persist. Our study complements these findings by showing that steep depreciation rates in newly acquired skills are also a key challenge in purely content-related teacher training programs focusing on primary school mathematics. Considering the lack of evidence on the sustainability of professional development programs and the relevance of the topic for educational policy, this likely remains an important avenue for future research.

2 Assessing teacher content knowledge in El Salvador

Despite impressive improvements in the accessibility of primary education, the quality of schooling often remains alarmingly low in developing countries. According to statistics by the World Bank (2018), less than 40 percent of students in a typical lower-middle income country pass minimum thresholds in mathematics by the end of primary school, and this rate drops to 14 percent in low income countries.

While many features of the schooling system affect the learning achievements of students, teachers are widely considered the most important input to the educational production function (Baumert et al. 2013; Hanushek 2011), and their salaries account for the bulk of education spending (Bold et al. 2017). Barber and Mourshed (2007) conclude from their study of high-performing school systems that 'the quality of an education system cannot exceed the quality of its teachers'. Hence, it is critical to understand how well teachers are prepared for the challenging task that awaits them in the classroom. One essential pre-requisite for effective teaching is a sound mastery of the concepts to be taught. However, recent evidence from African countries and India points to alarmingly low levels of teacher content knowledge. Most notably, Bold et al. (2017) report results on teacher skills based on a large-scale assessment across seven countries in Sub-Saharan Africa. According to their definition, a teacher possesses minimum subject knowledge in mathematics if she or he is able to mark at least 80 percent of items on a mock test for fourth graders correctly. On average, only two thirds of the teachers met this low requirement, with estimates ranging from 93 percent in Kenya to 49 percent in Togo. They 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. Similar results are reported for the Indian province Bihar, where only 34 percent of the teachers were able to solve a perimeter problem corresponding to grade five (Sinha, Banerji, and Wadhwa 2016).

Our research adds to the still sparse evidence on teachers' content knowledge by conducting a representative assessment in the department of Morazán in El Salvador, a lower middle-income country in Central America. According to recent World Bank data, both the access to and the quality of primary schooling in El Salvador is below the average for lower middle-income countries.[2] Morazán is located in southeastern El Salvador and is one of the poorest regions of the country. In national assessments, its secondary students perform at the country average.[3]

The first part of our study is based on a representative sample of 231 maths teachers from public primary schools in the department of Morazán who were asked to participate in an exam-type assessment.[4] Overall, 224 teachers (97%) complied with our invitation and took part in a 90-minute paper-and-pencil test comprising 50 items from the Salvadorian primary school maths curriculum. The weighting of questions across the three domains *Number Sense and Elementary Arithmetic* (∼65%),

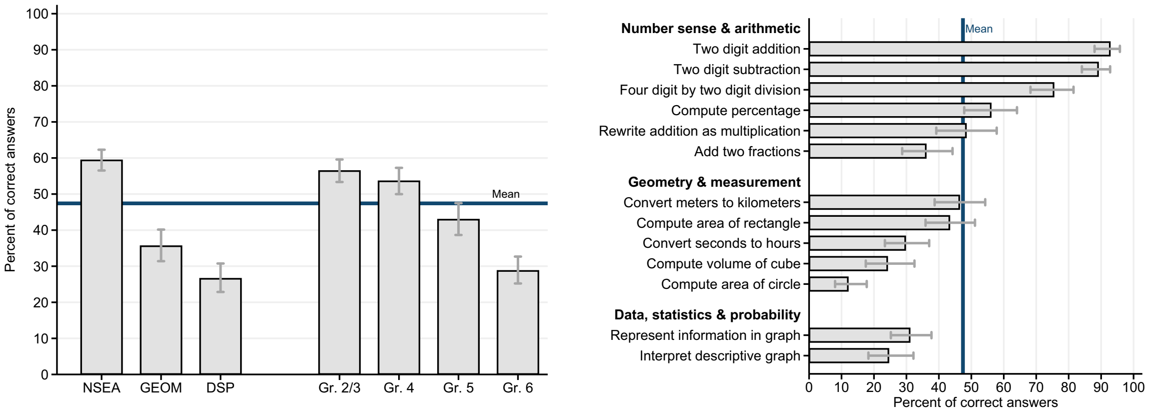


Figure 2. Content knowledge of maths teachers in Morazán, El Salvador. Note: *N*= 224; survey design taken into account; capped spikes indicate 95% confidence intervals. Results in sub Figure 2a are divided into the domains Number Sense and Elementary Arithmetic (NSEA), Geometry and Measurement (GEOM) and Data, Statistics and Probability (DSP).

Geometry and Measurement (∼30%), and *Data, Statistics and Probability* (∼5%) was closely aligned with the national curriculum. The test covered concepts taught in grade 2 (6 items), grade 3 (13 items), grade 4 (10 items), grade 5 (11 items), and grade 6 (10 items). To make sure that the items were suitable for the Salvadorian context, the assessment was reviewed by local teaching experts and the local education ministry.[5]

Figure 2 presents the results by subject domain and item difficulty (sub Figure 2a) and for selected example items (sub Figure 2b). The average teacher is able to answer 47 percent of grade two to six questions correctly, and performance is poor across all tested subject domains. Learning shortfalls are most apparent in *Data, Statistics and Probability* (27% correct answers) and *Geometry and Measurement* (36% correct answers), and least pronounced regarding *Number Sense and Elementary Arithmetic* (59% correct answers). Many teachers not only struggle with the more advanced items pertaining to grade six (29% correct answers), but even with items covering the basic materials from grades two and three (57% correct answers). While most teachers can handle basic operations such as additions or subtractions (about 90%), only 56% are able to solve a simple operation involving percentages, less than half (46%) can convert metres to kilometres or compute the area of a rectangle (43%), about a third can add two fractions (36%), and only one in four (25%) can retrieve information from a descriptive chart. Teachers instructing pupils up to grade 6 (54% correct answers) achieve somewhat better results than teachers instructing pupils up to grade 5 (46% correct answers), grade 4 (41% correct answers), and grade 3 (40% correct answers). Applying the minimum proficiency threshold advocated by Bold et al. (2017), our assessment suggests that only 14 percent of teachers possess sufficient content knowledge to effectively teach maths at the primary school level.

These results are particularly striking given that 97 percent of the teachers in our sample possess a university degree, meaning that they have either completed a teaching degree (2 to 3 years, 70% of teachers) or a bachelor's degree (5 to 6 years, 27% of teachers). Hence, despite 13 to 17 years of formal education, most primary schools teachers are confronted with the daunting task of teaching what they don't know. If quality education for all is to be achieved, it is thus critical to find effective ways to sustainably improve teacher skills.

3 Improving teacher content knowledge through computer-assisted learning

3.1 Intervention

For the second part of this study, we cooperated with the Swiss-Salvadoran NGO Consciente to implement an in-service teacher training program between April and August 2019. The intervention targeted 87 primary school maths teachers and consisted of two elements: *(i)* computer-assisted self-studying at home, and *(ii)* monthly revision workshops.

Self-Studying. Drawing on the extensive materials of the learning software Khan Academy, 16 study modules covering selected contents of the Salvadoran primary school maths curriculum were curated by the implementing organisation *Consciente*. In accordance with the official curriculum, the main focus of the training program was on *Number Sense and Elementary Arithmetic*, but concepts pertaining to *Geometry and Measurement* and *Data, Statistics and Probability* were 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.[6] 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 (weight: 0.85) and on quiz performance at the end of each module (weight: 0.15). 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. The maximum compensation a teacher could receive during the program was 348 USD, which roughly corresponds to 40 percent of the average monthly gross salary for Salvadoran primary school teachers.[7] Throughout the intervention, the software monitored teachers' progress and participants received regular reminders and individual support in case of technical problems.

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 in the self-studying modules 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, as they took place during work hours, teachers were only compensated for travel expenses.

3.2 Experimental design

To evaluate the impact of the program on teachers' maths performance, we set up a randomised controlled trial, where applicants to the teacher training program were randomly assigned to either the treatment or the control group. Before, shortly after, and one year after the intervention teachers were administered a comprehensive maths assessment. A comparison between the two groups allows us to track the causal effect of the program on teacher content knowledge over time.

Sampling and Randomisation. To recruit the study participants, our partner NGO visited 253 primary schools throughout Morazán and distributed registration sheets to all grade three to grade six maths teachers. In total, 313 teachers from 186 schools initially registered for the program/study and 274 teachers from 175 schools confirmed their application by attending a sensitisation meeting. In 108 out of 175 schools (i.e. 62%) only one teacher applied; in schools with multiple applications, every applicant was invited to the baseline assessment and only the worst-performing applicant of each school was selected for study participation yielding a final sample of 175 teachers from 175 different schools. Note that this part of the sampling procedure was not communicated to applicants to avoid misaligned incentives before or after the assessments. 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, randomisation was stratified by baseline score and gender.

Data and Measurement. The CAL-based intervention and the assessments were developed independently to avoid teaching-to-the-test artefacts. The primary objective of the mathematics assessments is to measure the maths competencies by teachers as laid out in the Salvadoran primary school curriculum. To that end, each assessment round comprised 50 different items from various international and Salvadoran sources and were specifically designed by the research team to emulate the Salvadoran maths curriculum for grades two to six covering *Number Sense and Elementary Arithmetic* (~60–65%), *Geometry and*

Measurement (~30%), and *Data, Statistics and Probability* (~5–10%) (see appendices B.3 and B.4 for more information). Hence, despite using different items, the three assessment rounds cover roughly the same level of difficulty and the same curricular content; for comparability, the baseline assessment is identical to the representative teacher assessment discussed in section 2. The assessments were administered during regional teacher meetings and had to be completed in 90 minutes using paper and pencil. We further collected data on teacher characteristics through a brief survey we administered directly after the baseline assessment. All participants were informed about how the collected data is used for implementation and research purposes. Our teacher data is complemented by administrative data on school characteristics provided by the education ministry as well as monitoring data on module completion and workshop attendance collected during the intervention.

Baseline Characteristics. Table A1 in the appendix shows that baseline characteristics are well-balanced across the two experimental groups. In both the treatment and the control group, the average teacher scored 43 percent correct answers and is thus slightly below the the regional average of 47 percent (p-value = 0.053). Table A2 in the appendix further shows that the average teacher in the experimental sample is 44 years old (compared to a regional average of 44.4 years, p-value = 0.68), 64 percent of the study participants are female (regional average: 60%, p-value = 0.47) and 24 percent completed a Bachelor's or Master's degree beside a teaching diploma (regional average: 30%, p-value = 0.24).

Compliance and Attrition. Good completion rates for modules (74%) and high attendance rates at workshops (85%) show that teachers complied well with the experimental protocol (see Figure A.1 in the appendix). While all 175 teachers participated in the baseline assessment, 164 teachers took the endline assessment shortly after the intervention (6% attrition), and 136 teachers participated in the follow-up assessment one year later (22% attrition). Table A3 compares attrition rates across experimental groups and provides no indication that participation in assessments correlated significantly with the treatment status.

Results

We estimate the intent-to-treat (ITT) treatment effect of being randomly assigned to the treatment group at endline (i.e. *EL*=one month after the intervention) or follow-up (i.e. *FU*=one year after the intervention) based on

$$Y_{jk}^{wave} = \alpha + \beta\, Treat_{jk} + \delta Y_{jk}^{BL} + X_{jk}^{'}\gamma + S_{jk}^{'}\rho + \phi_k + \in_{jk} \text{ for } wave \in [\mathrm{EL}, \mathrm{FU}], \tag{1}$$

where Y_{jk}^{wave} represents the endline (or follow-up) maths score of teacher *j* in stratum *k* and is either measured as the percentage share of correct answers or the standardised share of correct answers such that the control group's mean in a given wave is zero ($\mu_{control}^{wave} = 0$) and the standard deviation is one (i.e. $\sigma_{control}^{wave} = 1$). The main variable of interest is the binary indicator $Treat_{jk}$ that equals one if teacher *j* belongs to the treatment group. Y_{jk}^{BL} denotes the baseline test score and X_{jk} are additional pre-determined teacher attributes including age, gender, highest educational degree, years since graduation, maths specialisation and commuting time to school. S_{jk} captures covariates at the school level including an equipment and an infrastructure index, travel time to the department's capital as well as binary indicators for the availability of a computer lab, gang activities on school grounds and location in a rural area. Finally, ϕ_k denotes stratum fixed effects and $\in_{jk}$ is the error term. In the following, we report results based on equation (1) as well as a sparse specification excluding the pre-determined teacher attributes X_{jk} and school level characteristics S_{jk}.

Immediate Program Effect. Table 1 displays the benchmark estimates for the effect of the program on teachers' content knowledge.[8] In columns (1) to (4), we estimate the short-term program effects measured one month after the program ended. Columns (1) and (2) show that the evaluated teacher training program raised the share of correct answers by 5.38 to 5.52 percentage points (p-value < 0.01). This translates to an impact of 0.28σ to 0.29σ (p-value < 0.01) when the program effect is estimated based on standardised scores as in columns (3) and (4).

Table 1. ITT-estimates for the program effects on teachers' maths scores.

	Immediate effect				Effect after one year			
	Percent	correct	Standardized		Percent	correct	Standardized	
Dependent variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	5.38 ***	5.52 ***	0.28 ***	0.29 ***	0.61	1.48	0.03	0.08
	(1.46)	(1.49)	(0.08)	(0.08)	(1.78)	(1.77)	(0.10)	(0.10)
Baseline score	0.90 ***	0.85 ***	0.92 ***	0.86 ***	0.77 ***	0.64 ***	0.82 ***	0.68 ***
	(0.09)	(0.10)	(0.09)	(0.10)	(0.12)	(0.13)	(0.13)	(0.14)
Adjusted R^2	0.80	0.81	0.80	0.81	0.69	0.71	0.69	0.71
Observations	164	164	164	164	136	136	136	136
Teacher controls[a]	No	Yes	No	Yes	No	Yes	No	Yes
School controls[b]	No	Yes	No	Yes	No	Yes	No	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

The *immediate effect* is estimated based on the endline data collected in September 2019 about one month after the intervention concluded. The persistency of the *effect after one year* is estimated based on the follow-up data collected in September 2020. The mean (standard deviation) of the dependent variable for control units equals 45.4 (19.2) in *columns 1 and 2* and 54.9 (18.3) in *columns (5) and (6)*. In *columns (3), (4), (7), and (8)*, the share of correct answers is standardised to have a wave-specific mean of zero and a wave-specific standard deviation of one in the control group. *a: Teacher level controls* include age, educational degree, years since graduation, commuting time to school as well as binary indicators for gender and maths specialisation. *b: School level controls* are an infrastructure index, an equipment index, travel time to the department's capital as well as binary indicators for the availability of a computer lab, exposure to gang activities, and location in a rural area. Huber-White robust standard errors in parentheses. *$p<0.10$, **$p<0.05$, ***$p<0.01$.

An authoritative assessment of our results against previous findings is difficult because only few studies quantify the impact of teacher training programs on teacher content knowledge: Bando and Xia (2014) report standardised treatment effects of 0.35σ on teachers' English proficiency from a six-month professional development program in Mexico, whereas Zhang et al. (2013) find no significant impact on teachers' English skills from an intensive three-week program in Chinese migrant schools. Taking experimental impact evaluations on children's learning outcomes as a benchmark, the program's immediate impact of 0.29σ on teachers' content knowledge is sizeable. Even for well-proven types of educational interventions, such as remedial education or computer-assisted learning, systematic reviews report average effect sizes on children's maths scores below 0.2σ (e.g. Snilstveit et al. 2015; McEwan 2015).

Persistency of the Program Effect. How persistent is the program effect in the long run? Columns (5) to (8) of Table 1 indicate that less than one third of the impact remains after one year. The effect estimates based on the follow-up assessment vary between 0.6 percentage points (column 5, no controls) and 1.5 percentage points (column 6, with controls) or 0.03σ (column 7, no controls) and 0.08σ (column 8, with controls) and are imprecisely estimated with p-values between 0.40 and 0.73. Hence, the reported immediate gains in teachers' content knowledge were rather elusive, as the short-term effect depreciated by more than two-thirds after one year.

How do these findings compare to other studies? The evidence base on the long-term sustainability of teacher training programs is still surprisingly scarce, but so far it confirms that achieving persistent effects through teacher training programs is challenging. In line with our results, Bando and Xia (2014) find that the gap in English proficiency between participants and the control group documented immediately after the intervention disappeared when they reassessed the Mexican teachers twelve months later. Research by Cilliers et al. (2019, 2020) examines the sustainability of two teacher development programs focusing on teaching techniques instead of content knowledge. Their findings suggest that professional development programs are able to produce sustainable improvements among participants, but that persistent program effects cannot be taken for granted, even when a sizeable short-term impact has been achieved.[9] Similar to Cilliers et al. (2019, 2020) and Bando and Xia (2014), the results presented in Table 1 underscore that short-term gains do not necessarily translate to a sustained impact that persists in the long-run.

Effect Heterogeneity and Robustness. To gain a more nuanced understanding of the program's impact on teachers, we first explore several dimensions of effect heterogeneity and then asses whether the follow-up results may be driven by selective attrition.

Table A4 in the appendix shows that the immediate effect as well as the persistency of the impact did not vary by item domain: The teacher training program was equally effective in producing short-term gains in *Number Sense & Elementary Arithmetic* (0.25σ, $p < 0.01$) and in *Geometry, Measurement, Data & Statistics* (0.30σ, $p < 0.01$), and the effects across both domains largely disappear after one year. While we find no heterogeneity along domain, Table A5 shows that the program was about twice as effective at improving the participants' proficiency in concepts from grade levels five and six ($0.31–0.36\sigma$, $p < 0.01$) compared to concepts from grade levels three and four (both 0.19σ, $p = 0.06–0.13$). Yet, the ITT-estimates become insignificant across items of all grade levels at the follow-up assessment one year after the conclusion of the program.

Testing for effect heterogeneity along teacher characteristics in Table A6 and Figure A.3 shows that older teachers were significantly less perceptive than their younger colleagues. For instance, participants older than 50 gained on average 0.04σ ($p = 0.81$) at endline, while their youngest colleagues (≤ 40) experienced average gains of 0.53σ ($p < 0.01$); the gap between these two age groups is significant at the 5% level (p-value = 0.02). But even for teachers younger than 40, we do not obtain a significant program impact after one year (0.15σ, $p = 0.42$). We also find that participants with the lowest baseline score gained the least, but the effect differences along baseline ability are not statistically significant.

While the data reveal effect heterogeneity across several dimensions, all of the sub-analyses replicate the substantial deterioration in program effects after 12 months. Since only 136 teachers participated in the follow-up assessment (compared to 175 teachers at baseline and 164 teachers at endline), one may be concerned about bias induced by selective attrition. To understand the relevance of potential selection effects at the follow-up assessment, Table A7 restricts the analysis to the subsample of 131 teachers who participated in all three assessments. This leaves the point estimates unaltered compared to the benchmark analysis. For instance, the full specification for standardised scores at endline yields an effect of 0.29σ ($p < 0.01$) in column (4) of Table 1 compared to 0.28σ ($p < 0.01$) based on the restricted sample in column (4) of Table A7. Similarly, the changes in impact estimates for standardised scores after one year are negligible when we use the restricted sample (0.08σ, $p = 0.40$) instead of the benchmark specification (0.09σ, $p = 0.38$). Finally, Table A8 re-estimates the benchmark specification weighting observations by their inverse probability of selection into the endline or follow-up assessment. To be precise, we use entropy balancing (see Hainmueller 2012; Jann 2021) to reweight both the treatment group and the control group in a way such that the distribution of covariates in both groups (and during all assessments) is identical to the pooled distribution in the overall sample of 175 teachers. Again we do not find any indication that the deterioration in program effects after one year are driven by selective attrition along observable attributes.

3.4 Discussion: learning gains among students and the cost-effectiveness of the program

Two questions that naturally arise are *(a)* to what extent increased content knowledge of teachers transmits to their students, and *(b)* whether it is more cost-effective to organise CAL sessions for students or their inadequately prepared teachers. Originally, the field experiment was designed to address both questions *directly*, but school closures as a response to the COVID-19 pandemic disrupted in-class transmission from teachers to students for several months and infection risks inhibited large-scale assessments with children.

Teacher Content Knowledge & Learning Gains among Students. Due to COVID-19 related constraints, we quantify the transmission of teacher content knowledge to student learning from *observational* Salvadoran data, and then compare our results to quasi-experimental estimates from related international research. Our data on teacher content knowledge can be combined with information on student learning from a field experiment conducted in the same context (see Büchel et al. 2022). We merge the teacher and student data collected in 2018 using accurate class assignment information (i.e. school + grade + stream) to estimate the impact of teachers' content knowledge on learning gains of their students over the course of one school year. Our results in Table 2 suggest that a 1σ better teacher knowledge is associated with a 0.09σ to 0.12σ gain in student learning. Since the assignment of teachers

Table 2. Relation between teachers' content knowledge and students' learning during one school year in a sample of Salvadorian primary school classes of grades 3 to 6.

	Standardized student learning gains			
	(1)	(2)	(3)	(4)
Standardized teacher score	0.098 *** (0.031)	0.091 *** (0.032)	0.097 *** (0.031)	0.116 *** (0.036)
Grade level fixed effects	Yes	Yes	Yes	Yes
Class level controls	No	Yes	Yes	Yes
School level controls	No	No	Yes	Yes
Teacher controls	No	No	No	Yes

Number of observations: 2786 students, 120 teachers, 48 schools. *Teacher controls* comprise age, sex, highest degree, experience as a maths teacher, and travel time to school. *Class level controls* are class size, sex ratio, avg. household size, avg. household wealth, avg. maternal literacy rate within the class, and a binary indicator for afternoon classes. *School level controls* encompass an infrastructure index, an equipment index, travel time to the department's capital as well as binary indicators for student access to a computer lab, exposure to gangs, and location in a rural area. As the student data was collected for an experimental evaluation of a computer-assisted learning intervention (see Büchel et al. 2022), all models control for the treatment assignment of classes. School-level clustered standard errors in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

to classes was not experimentally manipulated, these estimates may be biased. To get a sense of the potential bias, Table A10 in the appendix compares our observational estimates to quasi-experimental evidence for primary schools in Peru (Metzler and Woessmann 2012), several African countries (Bietenbeck, Piopiunik, and Wiederhold 2018; Bold et al. 2019), and Pakistan (Bau and Das 2020). The international estimates for the transmission of teachers' content knowledge to student learning in mathematics are closely aligned with our most conservative estimate in column 2 of Table 2 and consistently suggest that a 1σ increase in teacher content knowledge is associated with an annual gain in students' maths scores of 0.09σ. Applying this transmission parameter, the immediate program effect on teachers' content knowledge of 0.29σ translates to a very small gain in average student learning of 0.026σ (i.e. $0.09 \times 0.29\sigma$), which is equivalent to 0.08 additional years of schooling (for details, see appendix section A.9).

Cost-Effectiveness of the Program. A fundamental advantage of teacher training programs are potential long-term cascade effects: CAL interventions targeting students require the *continuous* maintenance of large computer labs, whereas improving *one* teacher's content knowledge enhances the learning experience of *many* children *every* year. This brings about two favourable implications: First, the program costs per (indirectly) targeted child during a teacher training are considerably lower than the program costs per child for additional CAL lessons. Second, the costs of additional CAL lessons to children accrue periodically, while a one time investment in teacher skills produces recurrent gains – although these likely fade out as the treatment effect on teachers' content knowledge depreciates.

With these considerations in mind, we calculate the cost-effectiveness of the CAL-based teacher training combining four elements: *(i)* the immediate impact of the CAL training on teacher content knowledge, namely estimates from column (4) of Table 1; *(ii)* the annual depreciation in the program effect on teacher content knowledge combining the long-term effect estimates in column (8) of Table 1 with the immediate impact estimates; *(iii)* one-time implementation costs per (indirectly targeted) student calculated to 12 USD using the guidelines by Dhaliwal et al. (2014); *(iv)* the transmission of teacher content knowledge to students' learning gains, as discussed above.

One particularly valuable feature of this study is that its results can be compared to a companion paper evaluating CAL lessons offered to pupils of grades three to six (see Büchel et al. 2022). Importantly, the two field experiments were conducted in the same environment (i.e. primary schools in the Salvadorian department Morazán), using the identical CAL software for teaching basic mathematics (i.e. KHAN ACADEMY content via an offline application), and both interventions were implemented together with the same partner organisation (i.e. the Swiss-Salvadoran NGO CONSCIENTE).[10]

Figure 3a depicts cost-effectiveness estimates from Büchel et al. (2022) and cost-effectiveness estimates for the CAL-based teacher training based on 1000 random draws. We model uncertainty by

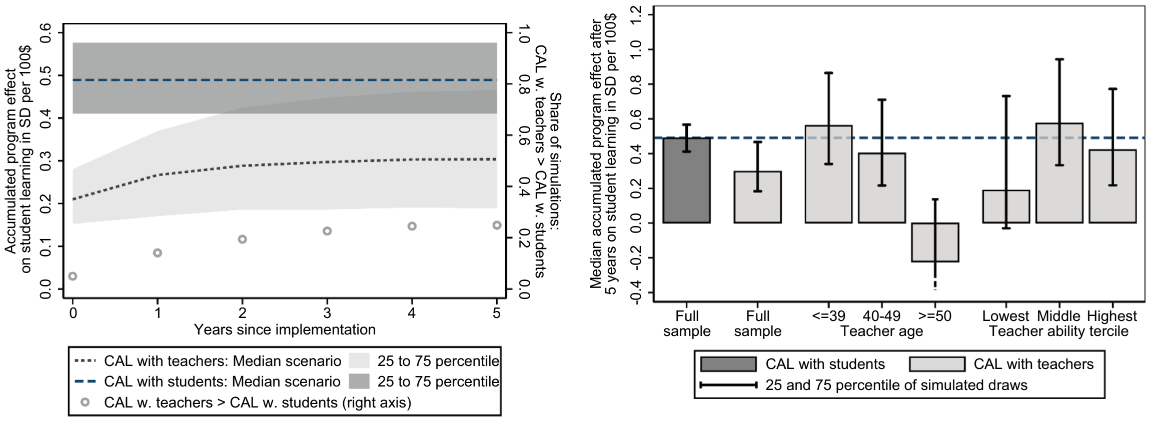


Figure 3. Simulated cost-effectiveness of CAL with teachers and CAL with students. Note: Figure 3a is based on 1000 random draws using the following parameters: Distribution of immediate program effect based on Table 1$\sim\mathcal{N}(0.29, 0.08^2)$; distribution of program effect after one year based on Table 1$\sim\mathcal{N}(0.08, 0.10^2)$; covariance between immediate and follow-up effect=0.004; distribution of effect transmission from teachers to students based on Tables 2 and A.10 $\sim\mathcal{N}(0.09, 0.03^2)$; program costs of CAL directed at teachers=12$ per student; distribution of annual effect of CAL program directed at students based on Büchel et al. (2022) $\sim\mathcal{N}(0.21, 0.05^2)$; program costs of CAL directed at students based on Büchel et al. (2022)=43$ per student. Figure 3b depicts the accumulated effect on students after five years for scenarios targeting different teacher groups that were analyzed in Section A.6: the plotted bars correspond to the median value after 5 years, while the capped spikes reproduce the 25th and 75th percentiles represented as shaded areas in Figure 3a.

drawing each cost-effectiveness parameter (except the implementation costs per student) from a normal distribution with a mean equal to the parameter's point estimate and a variance equal to the point estimate's squared standard error. The left-hand axis shows the accumulated program effect on student learning (measured in σ) per 100 USD, whereas the right-hand axis depicts the share of simulations yielding larger accumulated effects for CAL-based teacher trainings than for CAL-based lessons directed at students. The results suggest that the evaluated CAL lessons for students were likely more cost-effective than the evaluated CAL-based teacher trainings. In the median scenario, a 100 USD investment in CAL lessons for students increases learning gains by 0.49σ compared to 0.31σ for the same 100 USD investment in CAL-based teacher trainings. These impact estimates correspond to 1.5 school year equivalents for CAL directed at students, and 1 school year equivalent when CAL training is provided to teachers (for details, see appendix section A.9).

The shaded areas in Figure 3a represent draws between the 25th and 75th percentile, and indicate that the variance in the cost-effectiveness estimates for CAL-based teacher trainings is substantial. Yet, even when taking this large variance into account, CAL-based teacher trainings outperform CAL lessons for students in only 25 percent of the simulated scenarios, as depicted by the grey hollow circles. Most of the uncertainty in the cost-effectiveness simulations for the teacher trainings arises from the imprecise estimates on the persistency of program effects. If we remove this uncertainty by feeding the simulation with constant follow-up estimates (i.e. mean = 0.08 and sd = 0), the share of draws where CAL-based teacher trainings outperform additional CAL lessons for students decreases from 25 to 14 percent. To make the CAL training with teachers at least as cost-effective as CAL lessons with students, our simulations suggest that a retention rate of 55 percent among treated teachers would be required, which is twice as high as the retention rate observed in the experiment.

While these results are not in favour of software-based professional development programs, systematically targeting the most perceptive teachers would likely improve the cost-effectiveness of the policy. Figure 3b presents simulation results for the accumulated program effect on student learning after 5 years under different targeting regimes. The plotted bars correspond to the median scenario after 5 years, and the capped spikes reproduce the 25th and 75th percentiles represented as shaded areas in Figure 3a. The results in Figure 3b highlight that targeting young and middle-aged teachers would likely lift the teacher training's cost-effectiveness close to or even beyond the cost-effectiveness of additional CAL lessons for students. The same conclusion applies to targeting teachers in the second and third ability terciles. Having said that, a better understanding on how to achieve more persistent impacts in teacher trainings is arguably the key to unlocking the policy's full potential and increasing cost-effectiveness manifold.

4 Conclusion

Well qualified teachers are an essential requirement to achieve *quality education for all*, as envisioned by the *2030 Agenda for Sustainable Development*. Drawing on data from a representative maths assessment, this study documents that primary school maths teachers in northeastern El Salvador only master 47 percent of the curriculum they teach. This number is based on a direct assessment of teacher skills and is considerably lower than previous estimates for other developing countries relying on an indirect assessment through the grading of mock student tests.

Our field experiment shows that targeted teacher training using CAL software can produce substantial short-term gains in teachers' content knowledge. After a five-month teacher training program, we observe an average intention-to-treat effect of 0.29σ, with estimates ranging from effectively zero for teachers over the age of 50 to 0.52σ for teachers under 40. However, achieving sustained improvements in teacher skills proved to be more challenging. Learning gains at the teacher level depreciate by 72 percent to a mere 0.08σ one year after the treatment.

The unique setting of our experiment allowed us to compare the cost-effectiveness of CAL for teachers with that of an analogous CAL experiment directly targeting students. Teacher-centred initiatives are generally seen as a highly sustainable educational investment because they potentially benefit all future student cohorts a teacher instructs. Our simulations suggest that this assumption only holds if learning gains at the teacher level can be largely maintained over time. Based on the empirical parameters of the two experiments, we estimate that the retention rate of the effect on teacher knowledge should be at least 55 percent to guarantee that CAL for teachers is more cost-effective than CAL for students. With the actual retention rate of 28 percent we observed in our teacher experiment, the student-centred approach can be considered superior.

Our findings illustrate the importance of going beyond short term gains when evaluating the effectiveness of policies and interventions. While some programs can only be expected to have an impact on the cohort that was directly treated, others may induce sustained changes that can substantially increase the overall cost-effectiveness. Future research should appreciate this and help identify effective ways of ensuring the persistency of the achieved gains.

Notes

1. In a thorough literature search, we identified 28 experimental and quasi-experimental studies analysing the impact of teacher professional development in low and middle-income countries. Among those, seven training approaches include a significant content-knowledge component, and only three out of the seven studies actually assess the impact of the treatment on the content-knowledge of teachers (Antonio, Diosdado, and Moral 2011; Zhang et al. 2013; Bando and Xia 2014). The three cited studies evaluate training programs that combine pedagogical and content-related elements, and their findings are mixed. Bando and Xia (2014) document short term gains on the English skills of teachers and students in Mexico, whereas Antonio, Diosdado, and Moral (2011) find a positive impact on maths skills of Philippine teachers but not their students. Zhang et al. (2013) study the impact of a three-week training in English for teachers in Chinese schools, and report insignificant treatment effects for both teachers and students. A potentially promising approach to teacher professional development is the use of CAL software.
2. Measures for learning outcomes based on various international assessments at the primary school level were recently harmonised by Angrist et al. (2021). These harmonised learning outcomes are provided online by the World Bank, as are net primary school enrolment statistics: https://data.worldbank.org/indicator.
3. In 2019, Morazán ranked seventh among the 14 Salvadoran departments in the 'PAES' examination, a standardised test administered to all secondary school students throughout the country (MINED, Ministerio de la Educación Ciencia y Tecnología de El Salvador 2019).

4. The sample covers 98 of a total of 302 public primary schools in the department of Morazán, with an estimated population of about 650 teachers teaching at least one maths class between grades 3 and 6. For details on the design of the sample see Appendix section B.1. Taking the survey design into account, our sample can be considered representative for primary school maths teachers of grades 3 to 6 in the department of Morazán, but not necessarily for primary school teachers in the country as a whole.
5. More details on the assessment design are provided in the Appendix section B.3. We discuss further results from the assessment along a teacher opinion survey in an early draft of the working paper (see Brunetti et al. 2020).
6. KHAN ACADEMY is free of charge and features maths content in more than 30 languages. Like in many developing countries, poor internet coverage is a challenge in El Salvador. We therefore deployed an open-source platform, Kolibri, designed to make offline learning with content from Khan Academy and other CAL-sources possible.
7. The compensation was designed to promote program compliance, that is to incentivise teachers to expose themselves to the maths contents conveyed through the professional development program. Importantly, the payments were *not* conditional on teachers' performance in the baseline, endline, or follow-up assessment.
8. In the appendix section A.4, we present density plots for the participants' share of correct answers at the baseline, endline, and follow-up assessments disaggregated by treatment status.
9. One intervention arm studied by Cilliers et al. (2019, 2020) was delivered in the form of a four days training workshop designed to demonstrate how participants can teach a language and literacy curriculum effectively (training-based approach), while the second intervention arm was built around monthly visits from coaches who provided feedback on the participants' pedagogical techniques (coaching approach). For the *training-based* intervention, the authors report a substantial fade-out for the program's effect on teacher behaviour: Depending on the measured outcome, the program's effects declined by 50% to 90% over one year and became statistically insignificant. The *coaching-based* intervention produced more sustainable impacts on teacher behaviour, as between 66% and 100% of the immediate program effects carried over to the follow-up survey after one year.
10. The effect of CAL-based mathematics lessons for pupils in El Salvador is similar in magnitude to experimental impact estimates for CAL approaches in other low- and middle-income countries. Büchel et al. (2022) review nine experimental impact evaluations of CAL approaches with a total of eleven different CAL-based treatment arms. In ten out of eleven treatment arms the reported impact estimates are positive and statistically significant at the 10%-level. The average ITT endline effect across those eleven treatment arms is 0.24σ, while Büchel et al. (2022) estimate an ITT effect of CAL lessons of 0.21σ in the context of El Salvador. A broader and more recent review on the effectiveness of educational technology in developing countries by Rodriguez-Segura (2022) also suggests that CAL produces medium to large learning gains when used for self-led practicing.
11. The analysis draws on the same sample of students as the cited field experiment, except for the following limitations: Four teachers did not attend the assessment so that we drop their classes from the sample. We also eliminate five classes that were re-assigned to a new teacher during the school year 2018. Finally, we only include teachers who provided information on all covariates; this excludes another eleven classes from the sample.
12. Teacher controls comprise age, sex, highest degree, experience as a maths teacher, and travel time to school. Class level controls are class size, sex ratio, avg. household size, avg. household wealth, avg. maternal literacy rate within the class, and a binary indicator for afternoon classes. School level controls encompass an infrastructure index, an equipment index, travel time to the department's capital as well as binary indicators for student access to a computer lab, exposure to gangs, and location in a rural area.
13. A sub-group of the applicants took the maths assessment in the context of the representative maths assessment (see section B.1). In March 2019, the same assessment was administered to all other applicants. 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. In both cases, the assessment was unannounced.
14. Further information on the Standardized Testing and Reporting (STAR) program in California is available online: www.cde.ca.gov/re/pr/star.asp. VERA is coordinated by the Institut für Qualitätsentwicklung im Bildungswesen (IQB), see www.iqb.hu-berlin.de/vera. SAT is an acronym for standardised assessment tests coordinated by the UK's Standards and Testing Agency, see https://www.gov.uk/government/organisations/standards-and-testing-agency.

Acknowledgements

We are grateful to David Burgherr, Malin Frey, Christoph Kühnhanss, and Amélie Speiser who provided excellent research assistance. The project further benefited from invaluable feedback by Mauricio Romero and participants at the SSES Annual Congress 2021 and the German Development Economics Conference 2021. We further acknowledge generous funding by the IMG Stiftung and the Faculty of Business, Economics and Social Sciences of the University of Bern. Martina Jakob discloses that she serves on a voluntary basis as president of *Consciente – Unterstützungsverein für El Salvador (Schweiz)*. We received IRB approval from the Faculty of Business, Economics and Social Sciences at the University of Bern (ID = 052019). A randomized controlled trials registry entry is available at: https://www.socialscienceregistry.org/trials/4092.

Disclosure statement

Martina Jakob declares that she is the founder and president of Consciente – Unterstützungsverein für El Salvador (Schweiz), the NGO that implemented the educational interventions that are evaluated in this study. Martina Jakob contributes on a voluntary basis to Consciente and does not receive financial compensation for her work.

Funding

The work was supported by the University of Bern Research Foundation (IMG Stiftung) .

Notes on contributors

Aymo Brunetti is Professor of Economics at the University of Bern. His research interests are in development economics, economic policy and financial regulation.

Konstantin Büchel is Research Specialist at Youth Impact (Botswana) and Research Fellow at the Center of Regional Economic Development (CRED), University of Bern.

Martina Jakob is a PhD student at the Institute of Sociology of the University of Bern. In her research, she uses randomized controlled trials and machine learning to study topics such as education, public goods, environmental behavior, and inequality.

Ben Jann is Professor of Sociology at the University of Bern, Switzerland. His research interests include social-science methodology, statistics, social stratification, and labor market sociology. He is principal investigator of TREE, a large-scale multi-cohort panel study in Switzerland on transitions from education to employment (www.tree.unibe.ch).

Daniel Steffen is a lecturer and researcher at the Institute of Financial Services Zug of the Lucerne University of Applied Sciences and Arts. His research interests are in real estate and development economics.

References

Angrist, N., S. Djankov, P. Goldberg, and H. Patrinos. 2021. “Measuring Human Capital Using Global Learning Data.” *Nature* 592 (7854): 403–408. doi:10.1038/s41586-021-03323-7.

Antonio, S., N. M. Diosdado, and L. Moral. 2011. “Module-Based Professional Development for Teachers: A Cost-Effective Philippine Experiment.” *Teacher Development* 15 (2): 157–169. doi:10.1080/13664530.2011.571496.

Bando, R., and L. Xia 2014. The Effect of In-Service Teacher Training on Student Learning of English as a Second Language. IDB Working Paper Series No. 529.

Barber, M., and M. Mourshed. 2007. How the World’s Best-Performing Schools Systems Come Out on Top. Mc Kinsey & Company Report, online avaiable, URL: https://www.mckinsey.com/ .

Bau, N., and J. Das. 2020. “Teacher Value-Added in a Low-Income Country.” *American Economic Journal: Economic Policy* 12 (1): 62–96. doi:10.1257/pol.20170243.

Baumert, J., and K. Mareike. 2013. “The COACTIVE Model of Teachers’ Professional Competence.” In *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers*, edited by Mareike Kunter, et al., 25–48. New York: Springer.

Bietenbeck, J., M. Piopiunik, and S. Wiederhold. 2018. “Africa’s Skill Tragedy: Does teachers’ Lack of Knowledge Lead to Low Student Performance?” *The Journal of Human Resources* 53 (3): 553–578. doi:10.3368/jhr.53.3.0616-8002R1.

Bold, T., D. Filmer, G. Martin, E. Molina, C. Rockmore, B. Stacy, J. Svensson, and W. Wane. 2017. What Do Teachers Know and Do? Does It Matter? Evidence from Primary Schools in Africa. World Bank Policy Research Working Paper No. 7956.

Bold, T., D. Filmer, G. Martin, E. Molina, B. Stacy, C. Rockmore, J. Svensson, and W. Wane. 2017. “Enrollment Without Learning: Teacher Effort, Knowledge and Skill in Primary Schools in Africa.” *Journal of Economic Perspectives* 31 (4): 185–204. doi:10.1257/jep.31.4.185.

Bold, T., D. Filmer, E. Molina, and J. Svensson. 2019. The Lost Human Capital: Teacher Knowledge and Student Achievement in Africa. World Bank Policy Research Working Paper No. 8849.

Brunetti, A., K. Büchel, M. Jakob, B. Jann, C. Kühnhanss, and D. 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.

Büchel, K., M. Jakob, K. Christoph, D. Steffen, and A. Brunetti. 2022. “The Relative Effectiveness of Teachers and Learning Software. Evidence from a Field Experiment in El Salvador.” *Journal of Labor Economics* 40 (3): 737–777. doi:10.1086/717727.

Cilliers, J., B. Fleisch, J. Kotze, M. Mohohlwane, and S. Taylor. 2019. The challenge of sustaining effective teaching: Spillovers, fade-out, and the cost-effectiveness of teacher development programs. Unpublished manuscript.

Cilliers, J., B. Fleisch, C. Prinsloo, and S. Taylor. 2020. "How to Improve Teaching Practice? An Experimental Comparison of Centralized Training and In-Classroom Coaching." *The Journal of Human Resources* 66: 203–213.

Dhaliwal, I., D. Esther, G. Rachel, and T. Caitlin. 2014. "Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education." In *Education Policy in Developing Countries*, edited by Paul Glewwe, 285–338. Chicago and London: University of Chicago Press.

Escueta, M., A. Nickow, P. Oreopoulos, and V. Quant. 2020. "Upgrading Education with Technology: Insights from Experimental Research." *Journal of Economic Literature* 58 (4): 897–996. doi:10.1257/jel.20191507.

Glewwe, P., and M. Karthik. 2016. "Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps and Policy Implications." In *Handbook of the Economics of Education*, edited by Eric Hanushek, Stephen Machin, and Ludger Woessmann, 653–743. Amsterdam: Elsevier.

Hainmueller, J. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20 (1): 25–46. doi:10.1093/pan/mpr025.

Hanushek, E. 2011. "The Economic Value of Higher Teacher Quality." *Economics of Education Review* 30 (3): 466–479. doi:10.1016/j.econedurev.2010.12.006.

Jackson, K., J. Rockoff, and D. Staiger. 2014. "Teacher Effects and Teacher-Related Policies." *Annual Review of Economics* 6 (1): 801–825. doi:10.1146/annurev-economics-080213-040845.

Jann, B. 2021. Entropy Balancing as an Estimation Command. University of Bern Social Sciences Working Paper No. 39.

Kerwin, J., and R. Thornton. 2021. "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures." *The Review of Economics and Statistics* 103 (2): 251–264. doi:10.1162/rest_a_00911.

Kremer, M., C. Brannen, and R. Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science* 340 (6130): 297–300. doi:10.1126/science.1235350.

McEwan, P. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." *Review of Educational Research* 85 (3): 353–394. doi:10.3102/0034654314553127.

Metzler, J., and L. 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. doi:10.1016/j.jdeveco.2012.06.002.

MINED, Ministerio de la Educación Ciencia y Tecnología de El Salvador. 2019. Informe de resultados: Paes 2019. Online avaiable, URL: https://www.mined.gob.sv.

Rodriguez-Segura, D. 2022. "EdTech in Developing Countries: A Review of the Evidence." *The World Bank Research Observer* 37 (2): 171–203. doi:10.1093/wbro/lkab011.

Sinha, S., R. Banerji, and W. Wadhwa. 2016. *Teacher Performance in Bihar, India: Implications for Education*. Washington D. C: The World Bank.

Snilstveit, B., J. Stevenson, D. Phillips, M. Vojtkova, E. Gallagher, T. Schmidt, H. Jobse, M. Geelen, M. Pastorello, and J. Eyers. 2015. "Interventions for Improving Learning Outcomes and Access to Education in Low- and Middle- Income Countries: A Systematic Review." 3ie Systematic Review 24.

World Bank. 2018. *World Development Report 2018: Learning to Realize Education's Promise*. Washington D.C: World Bank.

Zhang, L., F. Lai, X. Pang, Y. Hongmei, and S. 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. doi:10.1080/19439342.2013.807862.

A Appendix: Analysis

A.1 *Characteristics at Baseline*

Table A1. Baseline characteristics by treatment status.

	Treatment group (1)	Control group (2)	p-value (3)
Panel A: Baseline maths scores (*N* = 175)			
%-Share correct answers	43.26	43.27	1.00
	(2.94)	(2.07)	
Standardized maths 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)	
Academic degree[b]	0.23	0.25	0.76
	(0.06)	(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)	
Panel C: School level 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[d]	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 area	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)	

This table presents the mean and standard error of the mean (in parentheses) for baseline characteristics by treatment status. 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*: A dummy variable indicating whether the teacher took the baseline in September 2018 (0) or March 2019 (1). *b*: Binary indicator whether teacher has completed an academic degree (1), i.e. *licenciatura* (5–6 years of tertiary education, equiv. to a bachelor's degree) or *maestria* (equiv. to a master's degree), rather than just a teaching degree (*profesorado*) (2–3 years of tertiary education) or high school (*bachillerato*) (0). *c*: Respondent teaches maths only (1) or various subjects (0). *d*: For each school a list covering twelve technical equipments and eleven facilities is available. The equipment and infrastructure indices refer to the share of items or facilities on this list that a school possesses.

Table A2. Comparison between representative and experimental sample.

	Represent. sample (1)	Experim. sample (2)	p-value (3)
Panel A: Baseline maths score			
%-Share correct answers	47.40	43.27	0.05
	(1.54)	(1.47)	
Panel B: Sociodemographics			
Age	44.43	44.07	0.68
	(0.64)	(0.60)	
Female	0.60	0.64	0.47
	(0.04)	(0.04)	
Academic degree[a]	0.30	0.24	0.24
	(0.04)	(0.03)	
Years since highest degree	19.66	19.29	0.68
	(0.67)	(0.61)	
Math specialization[b]	0.06	0.07	0.64
	(0.02)	(0.02)	
Travel time to school (min.)	54.65	65.58	0.10
	(4.43)	(4.94)	
Observations	224	175	

This table presents the mean and standard error of the mean (in parentheses) for baseline characteristics for the representative and the experimental sample. Column 3 shows the p-value (based on two-sided two-sample t-tests) from testing whether the mean is equal across the two samples. As no school-level information is available for the representative sample, we can only compare teacher-level variables. *a*: Binary indicator whether teacher has completed an academic degree (1), i.e. *licenciatura* (5–6 years of tertiary education, equiv. to a bachelor's degree) or *maestria* (equiv. to a master's degree), rather than just a teaching degree (*profesorado*) or high school (*bachillerato*) (0). With very few exceptions, teachers have either completed *licenciatura* or *profesorado*. *b*: Respondent teaches maths only (1) or various subjects (0).

A.2 *Program Compliance*

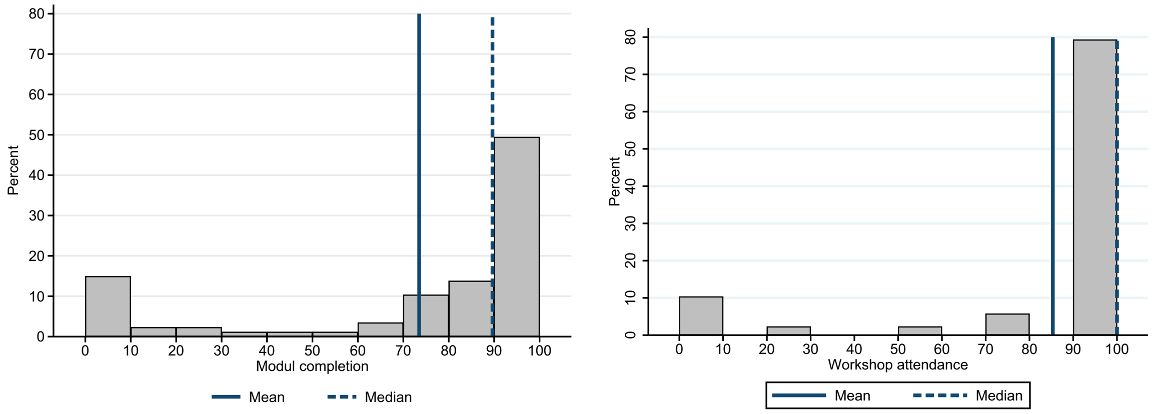


Figure A1. Compliance of teachers with the program. *Note*: The overall compliance rate is the weighted average of the module completion rate in Figure A.1a and the attendance rate in Figure A.1b with an average of 75 percent and a median of 91 percent.

A.3 *Attrition*

Eleven teachers (6.3%) did not take part in the endline assessment, and 39 teachers (22.2%) missed the follow-up assessment conducted during the first year of the COVID-19 pandemic. Table A3 reports estimates from linear probability models (LPM) that test whether the attrition status of participants is correlated with the treatment assignment. The estimates in columns (1) to (3) include different set of control variables and unambiguously suggest that attrition at the endline assessment is uncorrelated with treatment status (p= 0.56 − 0.78). Similarly, columns (4) to (6) yield an insignificant correlation between the treatment status and attendance at the follow-up assessment (p= 0.56 − 0.60). The same conclusions hold if the correlation between treatment status and attrition is estimated with a Logit model (results not shown).

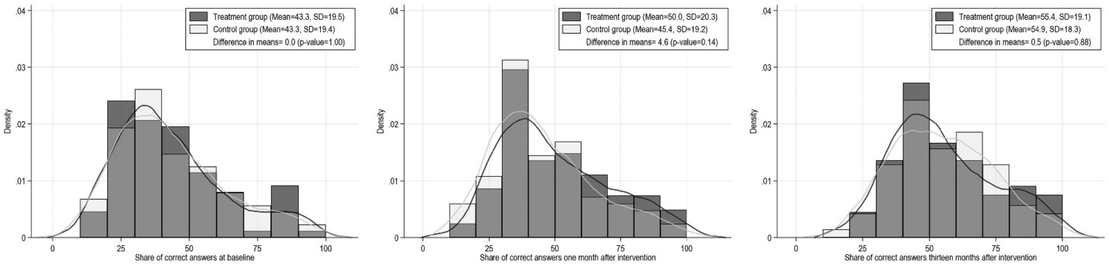
Table A3. Linear probability model for attrition by treatment status.

	Endline assessment			Follow-up assessment		
Attrition status at:	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.012	0.011	0.023	0.037	0.033	0.020
	(0.037)	(0.039)	(0.039)	(0.063)	(0.064)	(0.064)
Baseline score		−0.003 *	−0.004		0.000	−0.001
		(0.002)	(0.002)		(0.005)	(0.006)
Observations	175	175	175	175	175	175
Teacher controls	No	No	Yes	No	No	Yes
School controls	No	No	Yes	No	No	Yes
Stratum FE	No	Yes	Yes	No	Yes	Yes

Huber-White robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.4 *Distribution of Correct Answers by Wave and Treatment Status*

Figure A.2 presents density plots for the participants' share of correct answers at the baseline, endline, and follow-up assessments disaggregated by treatment status. Before the implementation of the program in spring 2019, the distributions of correct answers given by the treatment group and by the control group closely coincide (difference in means = 0.0, p-value = 1.00). At the endline assessment, about one month after the professional development program ended, we observe an increase in the share of correct answers in the treatment group compared to the control group. The difference in means at endline is 4.6 percentage points with a p-value of 0.14. One year later, at the follow-up assessment, the two distributions again largely overlap with a difference in means of 0.5 percentage points (p-value = 0.88).



(a) Baseline (March 2019) (b) Endline (Sept. 2019) (c) Follow-up (Sept. 2020)

Figure A2. Share of correct answers by treatment assignment.

A.5 Program Effects by Subtopic

Table A4. ITT-Estimates on the effects on teacher's standardised maths scores by subtopic.

	Immediate effect (in σ)		Effect after one year (in σ)	
Dependent variable:	NSEA	GEOM & DSP	NSEA	GEOM & DSP
Subject domain of items:	(1)	(2)	(3)	(4)
Treatment	0.25 ***	0.30 ***	0.09	0.06
	(0.09)	(0.09)	(0.11)	(0.11)
Baseline score	0.63 ***	0.49 ***	0.42 ***	0.48 ***
	(0.10)	(0.11)	(0.13)	(0.11)
Adjusted R^2	0.73	0.72	0.60	0.67
Observations	164	164	136	136
Teacher controls	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes
Stratum FE	Yes	Yes	Yes	Yes

The *immediate effect* is estimated based on the endline data collected in Sept. 2019 about one month after the intervention concluded. The persistency of the *effect after one year* is estimated based on the follow-up data collected in September 2020. *NSEA*: Items covering number sense and elementary arithmetics; *GEOM & DSP*: Items covering geometry and measurement as well as data, statistics, and probability. In all columns, the share of correct answers (by subject domain) is standardised to have a wave-specific mean of zero and a wave-specific standard deviation of one in the control group. Huber-White robust standard errors in parentheses.

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

A.5.1 Program Effects by Grade Level

Table A5. ITT-Estimates on the effects on teacher's maths scores by grade level of items.

	Immediate effect (in σ)				Effect after one year (in σ)			
Dependent variable:	Gr. 2/3	Gr. 4	Gr. 5	Gr. 6	Gr. 2/3	Gr. 4	Gr. 5	Gr. 6
Grade level of items:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.19	0.19 *	0.31 ***	0.36 ***	0.12	−0.09	0.09	0.16
	(0.12)	(0.10)	(0.10)	(0.11)	(0.14)	(0.12)	(0.12)	(0.11)
Baseline score	0.35 ***	0.08	0.58 ***	0.38 ***	0.29 **	0.15	0.52 ***	0.36 ***
	(0.09)	(0.10)	(0.11)	(0.09)	(0.12)	(0.12)	(0.14)	(0.11)
Adjusted R^2	0.58	0.63	0.68	0.60	0.38	0.56	0.54	0.64
Observations	164	164	164	164	136	136	136	136
Teacher controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

The *immediate effect* is estimated based on the endline data collected in Sept. 2019 about one month after the intervention concluded. The persistency of the *effect after one year* is estimated based on the follow-up data collected in Sept. 2020. In all columns, the share of correct answers (by grade level of items) is standardised to have a wave-specific mean of zero and a wave-specific standard deviation of one in the control group. Huber-White robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.6 Program Effects by Teachers' Baseline Ability and Age

We estimate the following regression equation:

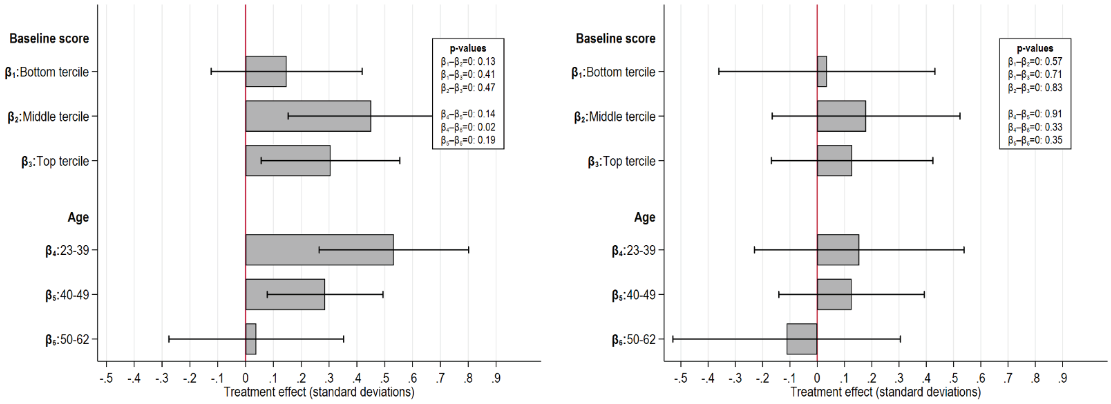
$$Y_{jk}^{EL} = \alpha + \beta\, Treat_{jk} + \lambda(Treat_{jk} \times Covariate_{jk}) + \delta Y_{jk}^{BL} + X_{jk}^{'}\gamma + S_{jk}^{'}\rho + \phi_k + \epsilon_{jk}, \qquad (A.1)$$

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 (1).

Table A6. Effect heterogeneity along baseline ability and age.

	Immediate effect (in σ)		Effect after one year (in σ)	
Dependent variable:	Baseline score	Age	Baseline score	Age
Covariates:	(1)	(2)	(3)	(4)
Treatment	0.29 ***	0.28 ***	0.08	0.07
	(0.08)	(0.08)	(0.10)	(0.10)
Covariate	0.85 ***	−0.01 *	0.67 ***	−0.02
	(0.11)	(0.01)	(0.14)	(0.01)
Treatment x covariate	0.03	−0.02 *	0.03	−0.02
	(0.07)	(0.01)	(0.10)	(0.01)
Adjusted R^2	0.80	0.81	0.71	0.72
Observations	164	164	136	136
Baseline maths score	Yes	Yes	Yes	Yes
Teacher controls	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes
Stratum FE	Yes	Yes	Yes	Yes

The *immediate effect* is estimated based on the endline data collected in Sept. 2019 about one month after the intervention concluded. The persistency of the *effect after one year* is estimated based on the follow-up data collected in Sept. 2020. In all columns, the share of correct answers is standardised to have a wave-specific mean of zero and a wave-specific standard deviation of one in the control group. Huber-White robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.



(a) Immediate program effects (in σ)

(b) Program effects after one year (in σ)

Figure A3. Effect heterogeneity by baseline score and age. *Note*: Same set of controls as in Table A7. Spikes show 95% confidence intervals.

A.7 *Program Effects Estimated with Fully Balanced Sample*

Table A7. ITT-estimates for the program effects on teachers' maths scores with constant sample.

	Immediate effect				Effect after one year			
	Percent correct		Standardized		Percent correct		Standardized	
Dependent variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	5.21 ***	5.28 ***	0.27 ***	0.28 ***	0.71	1.58	0.04	0.09
	(1.61)	(1.66)	(0.08)	(0.09)	(1.83)	(1.80)	(0.10)	(0.10)
Baseline score	0.83 ***	0.73 ***	0.84 ***	0.75 ***	0.76 ***	0.60 ***	0.81 ***	0.64 ***
	(0.09)	(0.11)	(0.09)	(0.11)	(0.12)	(0.13)	(0.13)	(0.14)
Adjusted R^2	0.80	0.81	0.80	0.81	0.70	0.73	0.70	0.73
Observations	131	131	131	131	131	131	131	131
Teacher controls	No	Yes	No	Yes	No	Yes	No	Yes
School controls	No	Yes	No	Yes	No	Yes	No	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

The *immediate effect* is estimated based on the endline data collected in September 2019 about one month after the intervention concluded. The persistency of the *effect after one year* is estimated based on the follow-up data collected in September 2020. In *columns (3), (4), (7), and (8)*, the share of correct answers is standardised to have a wave-specific mean of zero and a wave-specific standard deviation of one in the control group. Huber-White robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.8 *Program Effects Estimated with Entropy Balancing*

Table A8. ITT-estimates for the program effects on teachers using entropy balancing.

	Immediate effect				Effect after one year			
	Percent correct		Standardized		Percent correct		Standardized	
Dependent variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	5.33 ***	5.33 ***	0.28 ***	0.28 ***	1.35	1.35	0.07	0.07
	(1.46)	(1.46)	(0.08)	(0.08)	(1.77)	(1.77)	(0.10)	(0.10)
Baseline score	0.88 ***	0.85 ***	0.89 ***	0.86 ***	0.74 ***	0.67 ***	0.79 ***	0.71 ***
	(0.09)	(0.10)	(0.09)	(0.10)	(0.12)	(0.13)	(0.13)	(0.14)
Observations (weighted)	175	175	175	175	175	175	175	175
Observations (unweighted)	164	164	164	164	136	136	136	136
Teacher controls	No	Yes	No	Yes	No	Yes	No	Yes
School controls	No	Yes	No	Yes	No	Yes	No	Yes
Stratum FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

The *immediate effect* is estimated based on the endline data collected in September 2019 about one month after the intervention concluded. The persistency of the *effect after one year* is estimated based on the follow-up data collected in September 2020. In *columns (3), (4), (7), and (8)*, the share of correct answers is standardised to have a wave-specific mean of zero and a wave-specific standard deviation of one in the control group. Robust standard errors accounting for uncertainty induced by entropy balancing in parentheses, see Jann (2021) for methodological details. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.9 *The Effect of Teacher Content Knowledge on Student Learning*

This section comprehensively discusses evidence on the effect of teacher content knowledge on student learning: *First*, we present estimates based on Salvadoran data. *Second*, we summarise international evidence from Asia, Africa, and South America. *Third*, we discuss a possible quantification in terms of school year equivalents.

Estimates based on Salvadoran Data. Our data on teacher content knowledge can be combined with data on student learning outcomes collected during a field experiment in 2018 (for details see Büchel et al. 2022).[11] The teacher survey was administered towards the end of the school year 2018, which also marked the end of the aforementioned experiment. However, as the assignment of teachers to classes was not experimentally manipulated, we do not claim that the reported correlations are causal. In line with standard practice, we specify the basic model for estimating the relation between teacher content knowledge and students' learning as

$$\Delta\tilde{Y}_i = \alpha + \beta\tilde{S}_j + G_i\gamma + T_i\delta + \in_i.$$

The dependent variable, $\Delta\tilde{Y}_i$, is student i's learning defined as $\Delta\tilde{Y}_i = \tilde{Y}_i^2 - \tilde{Y}_i^1$ with $\tilde{Y}_i^1 = (Y_i^1 - \bar{Y}^1)/\hat{\sigma}_{Y^1}$ and $\tilde{Y}_i^2 = (Y_i^2 - \bar{Y}^1)/\hat{\sigma}_{Y^1}$, where Y_i^1 and Y_i^2 are the student's IRT scores in wave 1 and wave 2, respectively, and $\bar{Y}^1$ and $\hat{\sigma}_{Y^1}$ are the mean and standard deviation of the scores in wave 1. The predictor of interest is $\tilde{S}_j$, the standardised knowledge score of teacher j (who teaches student i), defined as $\tilde{S}_j = (S_j - \bar{S})/\hat{\sigma}_S$ where S_j is the percentage of correct answers that teacher j achieved in the assessment. The model further includes an indicator vector for the student's grade, G_i, since teacher knowledge is correlated with grade and ability improvements are smaller among higher-grade students. Furthermore, the treatments imposed as part of the field experiment did affect learning so that teacher effects are evaluated within treatment groups as captured by the indicator vector T_i.

This basic model corresponds to specification (1) in Table A9. Additional specifications include class-level controls (columns 2–4), school-level controls (columns 3 & 4), and additional teacher characteristics (column 4).[12] All specifications yield a positive relation between teacher content knowledge and student learning. Quantitatively, a 1σ increase in teacher knowledge is associated with a 0.09σ to 0.12σ gain in student learning.

International Evidence. In Table A10, we compare our estimates reported in Table A9 to recent evidence reported for primary schools in (Metzler and Woessmann 2012), Africa (Bietenbeck, Piopiunik, and Wiederhold 2018; Bold et al. 2019), and Pakistan (Bau and Das 2020). While the evidence unambiguously demonstrates that better content knowledge of teachers improves student learning, the effect magnitude varies by subject. Studies distinguishing between maths and language find that teachers' content knowledge plays a more decisive role in the instruction of maths. Evidence for Peru, Pakistan and El Salvador consistently suggest that a 1σ increase in teacher content knowledge is associated with an annual gain in students' maths scores of about 0.09σ. With respect to language, less evidence is available and the correlation is weaker. The estimated coefficients vary between 0.03 (insig.) and 0.06 for languages, and between 0.03 and 0.06 when the effect of teacher content knowledge is estimated across multiple subjects.

The finding that content knowledge of teachers has a stronger impact on learning outcomes in maths is consistent with studies from OECD countries reporting greater variance in teacher effects on achievement in maths than language. One reason may be that maths is almost exclusively learned in the classroom, while languages are learned to a great extent outside of school (e.g. Jackson, Rockoff, and Staiger 2014).

Converting Salvadoran Estimates to School Year Equivalents. To assess the magnitude of the relation between teachers' content knowledge and student learning, it is informative to express learning gains in school year equivalents. To do so, we use our Salvadoran data introduced above and compute each student's difference in IRT scores between wave 1 and 2 and divide it by the average score difference between grades, so that results are expressed in units of children's average learning gains during one school year. Formally, we replace the dependent variable $\Delta\tilde{Y}_i$ with

$$\Delta Y_i^E = (Y_i^2 - Y_i^1)/\hat{\gamma}$$

where $\hat{\gamma}$ is the slope coefficient of student's grade $\tilde{G}_i$ in model

$$Y_i = \alpha + \gamma\tilde{G}_i + T_i\delta + \in_i$$

estimated using data from wave 2. Treatment indicator vector T_i is included in the model to eliminate a biasing effect of

Table A9. Relation between teacher's test score and students' learning over an eight month evaluation period and in a sample of Salvadorian primary school classes of grades 3 to 6.

	Student learning gains							
	Standardized (σ)				School year equivalents			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Standardized teacher score	0.098 *** (0.031)	0.091 *** (0.032)	0.097 *** (0.031)	0.116 *** (0.036)	0.276 *** (0.083)	0.256 *** (0.085)	0.274 *** (0.080)	0.324 *** (0.091)
Grade level fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Class level controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
School level controls	No	No	Yes	Yes	No	No	Yes	Yes
Teacher controls	No	No	No	Yes	No	No	No	Yes

Notes: Number of observations: 2786 students, 120 teachers, 48 schools. *Teacher controls* comprise age, sex, highest degree, experience as a maths teacher, and travel time to school. *Class level controls* are class size, sex ratio, avg. household size, avg. household wealth, avg. maternal literacy rate within the class, and a binary indicator for afternoon classes. *School level controls* encompass an infrastructure index, an equipment index, travel time to the department's capital as well as binary indicators for student access to a computer lab, exposure to gangs, and location in a rural area. As the student data was collected for an experimental evaluation of a computer-assisted learning intervention (see Büchel et al. 2022), all models control for the treatment assignment of classes. School-level clustered standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A10. Evidence on the effect of teacher content knowledge on student learning in developing countries.

	Metzler and Woessmann (2012)	Bietenbeck, Piopiunik, and Wiederhold (2018)	Bold et al. (2019)	Bau and Das (2020)	results, Table A.9
Main effect (per year)					
+1σ teacher test score	*Math*: 0.09	*Mixed*: 0.03	*Mixed*: 0.07	*Math*: 0.09	*Math*: 0.09–0.12
on student test scores (in σ)	*Lang.*: 0.03 (insig.)			*Language*: 0.06	
Sample					
Country and region	Peru	6 East African countries	7 African countries	Pakistan, Punjab	El Salvador, Morazán
Subjects	Math Language	*Mixed*: Math and language	*Mixed*: Math and language	Math Language	Math
Level of education	Primary school	Primary school	Primary school	Primary school	Primary school
	(Grade 6)	(Grade 6)	(Grade 4)	(Grades 3–5)	(Grades 3–6)
Empirical strategy	Teacher FE +	Teacher FE +	Teacher FE +	Teacher value-	various
	Student FE	Student FE	Student FE	added approach	controls

Sources for estimates reported in first row: Metzler and Woessmann (2012): Table 2, column 1. Bietenbeck, Piopiunik, and Wiederhold (2018): Table 3, column 5. Bold et al. (2019): Table 4, column 3. Bau and Das (2020): Table 3 (columns 2–6), Table 4 (column 7).

the CAL intervention that took place between wave 1 and wave 2.

Replicating columns (1) to (4) in Table A9 with student learning measured in school year equivalents suggests that a 0.3σ increase in a teacher's maths score is associated with 0.08 to 0.1 additional years of schooling (see columns 5 to 8 in Table A9). Accordingly, shifting a student from a teacher at the lowest to one at the highest decile would yield 0.7 to 0.9 additional years of schooling, and hence almost double the students' annual progress in maths.

B Appendix: Methods

B.1 *Sampling for the Representative Teacher Assessment in 2018*

Our base population encompasses all primary school maths teachers teaching at least one class between grades 3 and 6 in one of the 302 public primary schools in the department of Morazán, El Salvador. Six out of the 302 public schools in Morazán registered zero students in these grades and were excluded, leaving 296 schools in the population. Since the teacher assessment took place in the context of a randomised controlled trial on a computer assisted learning (CAL) (Büchel et al. 2022), our sample is drawn from two strata of schools.

(1) *Schools that were eligible for the CAL project*: Of the 296 public primary schools with classes in grades 3 to 6 in Morazán, 57 schools fulfilled the eligibility criteria for the CAL project (defined in terms of school size, security situation, accessibility, and electrification). In these 57 schools, 198 classes from grades three to six were randomly chosen to be part of the CAL experiment. All maths teachers instructing at least one of these classes are included in the target sample of the present study (138 teachers; 4 of them did not participate). Teachers from this stratum of schools had a probability of 65.7% of becoming part of our sample and are thus over-sampled relative to the base population.

(2) *Schools that were not eligible for the CAL project*: Among the remaining 239 schools, 50 schools were randomly selected, stratified by 16 geographical regions, and all maths teachers in grades 3 to 6 in these schools were invited to participate in the assessment (93 teachers; 3 of them did not participate). Teachers from this stratum of schools had a sampling probability of 21%.

In our data analyses, we take account of the described stratification, the unequal sampling probabilities, as well as the fact that schools, not teachers, are the primary sampling unit (using Taylor-linearisation for variance estimation).

B.2 *Sampling and Randomization for the Field Experiment in 2019/2020*

As illustrated in Figure B.1, the sampling and randomisation procedure for the field experiment consisted of five steps.

(1) All public primary schools with students in grades three to six in Morazán serve as the starting point for the sampling process.
(2) 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.
(3) The NGO sent out invitations to all grade three to six maths teachers in eligible schools. Overall, 313 teachers from 186 schools applied to participate in the program/study and 274 teachers from 175 schools confirmed their interest by attending a sensitisation meeting.
(4) Before the start of the intervention, all candidates took an unannounced baseline assessment.[13] Based on the results of this assessment, 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.
(5) In a final step, 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, randomisation 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 randomised at the teacher level, the pre-selection left only one teacher per school in the sample. This prevents potentially biased estimates due to spillover effects within schools.

B.3 *Assessment Design*

To design the maths tests for the representative teacher assessment and the three assessments for the field experiment, we proceeded as follows.

(1) We first summarised the Salvadoran maths curriculum for grades two to six along the three topics *Number Sense & Elementary Arithmetic* (NSEA), *Geometry & Measurement* (GEOM), and *Data, Statistics & Probability* (DSP).
(2) For the assessments, we then mapped test items from various sources on the Salvadoran curriculum. These sources include official textbooks of El Salvador, publicly available items from the STAR evaluations in California, publicly available items from the VERA evaluations in Germany, and publicly available items from the SAT assessments in Britain.[14]
(3) We then designed paper and pencil maths 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 are meant to be completed in 90 minutes. The relative weighting of the three main domains emulates the weighting in the national primary school maths curriculum. To make sure that questions are suitable for the Salvadoran context, assessments were reviewed by local teaching experts and the local education ministry. Moreover, the exam lasted a generous 90 minutes to guarantee that every participant had enough time to carefully draft the answers so that wrong answered cannot be attributed to time pressure.
(4) Based on these assessments, we used two different main outcome measures at the teacher level: the share of correctly answered questions and standardised test scores. All results in the field experiment are based on double coded data by pre-trained staff in El Salvador (batch 1) and Switzerland (batch 2 plus harmonisation of batches 1 and 2).

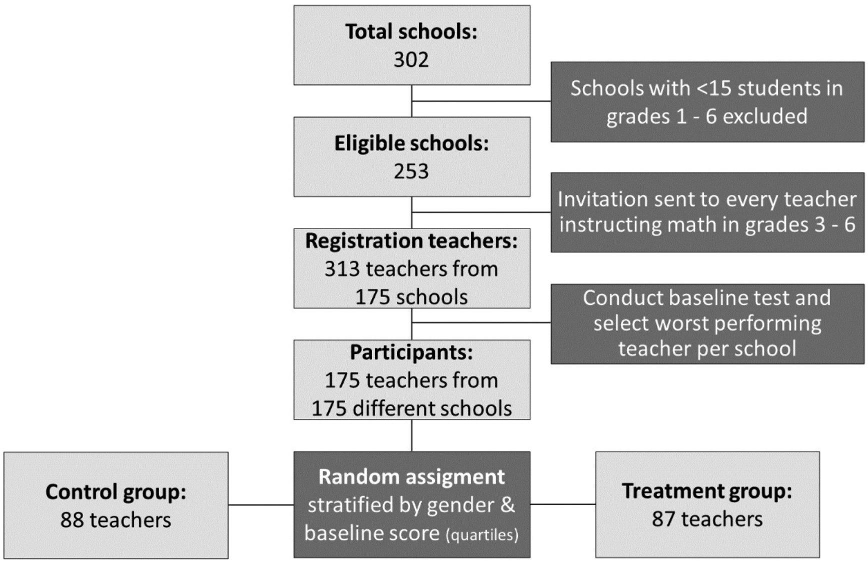
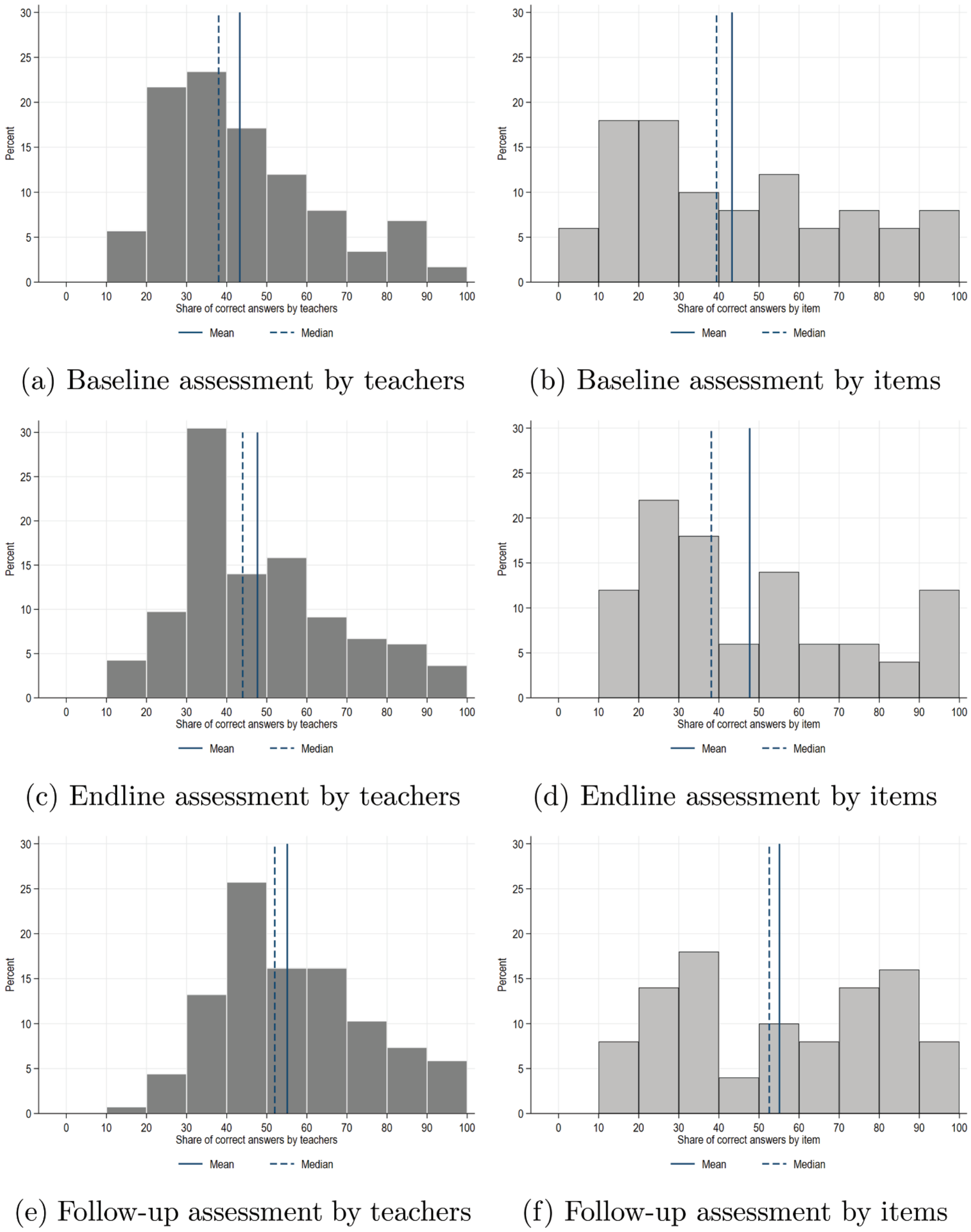

Total schools:
302
Schools with <15 students in grades 1 - 6 excluded
Eligible schools:
253
Invitation sent to every teacher instructing math in grades 3 - 6
Registration teachers:
313 teachers from 175 schools
Conduct baseline test and select worst performing teacher per school
Participants:
175 teachers from 175 different schools
Control group:
88 teachers
Random assigment stratified by gender & baseline score (quartiles)
Treatment group:
87 teachers

Figure B1. Sampling and randomisation scheme.

B.4 *Assessment Diagnostics*

Figure B.2 presents the distribution of correct answers by teachers and by items for the baseline, endline and follow-up assessments. The histograms show that there are neither floor nor ceiling effects. Teachers were able to answer at least 10 percent of the items in the baseline, 16 percent in the endline and 18 percent in the follow-up assessment. 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 share of correct answers across all waves and items is 3 percent) or an item which was solved successfully by all teachers (maximum share of correct answers is 98 percent across all waves and items).



(a) Baseline assessment by teachers

(b) Baseline assessment by items

(c) Endline assessment by teachers

(d) Endline assessment by items

(e) Follow-up assessment by teachers

(f) Follow-up assessment by items

Figure B2. Share of correct answers across items and teachers by assessment.

Chapter 3

Participatory Teaching Improves Learning Outcomes: Evidence from a Field Experiment in Tanzania

Participatory Teaching Improves Learning Outcomes: Evidence from a Field Experiment in Tanzania*

Martina Jakob[a,°], Konstantin Büchel[a,°],
Daniel Steffen[b], Aymo Brunetti[a]

[a]University of Bern
[b]Lucerne University of Applied Sciences and Arts

October 29, 2023

Abstract

Participatory teaching methods have been shown to be more successful than traditional rote learning in high-income countries. It is, however, less clear if they can help address the learning crisis in low- and middle-income countries, where classes tend to be large and teachers have fewer resources at their disposal. Based on a field experiment with 440 teachers from 220 schools in Tanzania, we use official standardized student examinations to assess the impact of a pedagogy-centered intervention. A five-day in-service teacher training on participatory and practice-based methods improved students' test scores 18 months later by 0.15σ. The additional provision of laptops with a learning software allowing teachers to refresh their content knowledge did not yield further learning gains for students. Complementary results from qualitative surveys and interviews suggest that the program was highly appreciated by different stakeholders, but that participants are unable to assess its impact along different dimensions, giving equally positive evaluations of its successful and its less successful elements.

JEL classification: C93, I21, J24, O15.
Keywords: productivity in education, participatory teaching, teacher content knowledge, computer-assisted learning, development economics.

*The authors acknowledge generous funding by the IMG Stiftung. This study received IRB approval from the Faculty of Business, Economics and Social Sciences at the University of Bern on November 4, 2019 (serial number: 122019). A randomized controlled trials registry entry is available at: `https://www.socialscienceregistry.org/trials/4959`.

Contact Details (°shared first authorship):
°Jakob: Univ. of Bern, Inst. of Sociology, Fabrikstr. 8, CH-3012 Bern, martina.jakob@unibe.ch
°Büchel: Univ. of Bern, Dept. of Economics, Schanzeneckstr. 1, CH-3001 Bern, konstantin.buechel@unibe.ch
Steffen: Lucerne University of Applied Sciences and Arts, Suurstoffi 1, CH-6343 Rotkreuz, dani.steffen@hslu.ch
Brunetti: Univ. of Bern, Dept. of Economics, Schanzeneckstr. 1, CH-3001 Bern, aymo.brunetti@unibe.ch

1 Introduction

Only 4 percent of students in low-income countries, compared to 95 percent in high-income countries, reach minimum literacy skills towards the end of primary school (World Bank, 2018, p. 8). To narrow the global learning gap, we need to rethink the strategies that teachers in developing countries use in the classroom. While schools in high-income countries have increasingly adopted participatory pedagogical approaches with a high degree of student engagement, more teacher-centered approaches such as lecturing and rote learning are still the norm in many low- and middle-income countries. Modern pedagogy takes a clear stance and considers student engagement a vital component of effective teaching, a view that is corroborated by vast evidence from high-income countries (e.g. Cornelius-White, 2007; Seidel and Shavelson, 2007; Harbour et al., 2015). However, it is not clear if this insight can be transferred to low- and middle-income countries, where teachers often have to manage very large classrooms and have few teaching aids at their disposal. Under such constraints, switching to more demanding teaching strategies could even prove detrimental (e.g., Berlinski and Busso, 2017). Moreover, in light of recent evidence on insufficient subject mastery among many teachers in disadvantaged regions (e.g., Sinha et al., 2016; Bold et al., 2017a; Brunetti et al., 2023), it remains an open question whether improving pedagogy alone is effective or if shortfalls in teachers' content knowledge need to be tackled simultaneously.

To address these questions, we conducted a randomized controlled trial (RCT) with 440 math teachers and more than 25,000 students from 220 schools in Tanzania. With an average of 51 students per teacher and a persistent shortage of classrooms and teaching aids, Tanzania faces resource constraints that are typical for many education systems in low-income countries (UNESCO, 2022). The intervention we study consisted of a five-day in-service program where teachers learned how to engage their students more actively in classes, bring their teaching closer to every-day live, and collaborate in teams to handle large classrooms and exchange on teaching techniques. After the initial five-day workshop, all teachers were invited to half-yearly refresher meetings to revise concepts and discuss implementation issues. Half of the teachers in the treatment group were randomly selected to further receive a laptop with a computer-assisted learning (CAL) software enabling them to refresh their content knowledge. The learning software consisted of short math videos and quizzes from "Khan Academy", and teachers participated in additional sessions to familiarize themselves with the program and discuss their progress. Both versions of the treatment were administered by Swiss NGO Helvetas that has implemented teacher training programs in Tanzania since 2000.

We scraped student-level data from standardized assessments published by National Examinations Council of Tanzania (NECTA) to estimate the impact of the program on students, and used data from our own assessments to study intermediate effects on teachers. Our design allows us to analyze direct effects on participating teachers and their students as well as spillover effects on peer teachers and students in treated schools. To better understand the mechanisms behind potential effects, we complemented the experimental data with classroom observations, surveys, and in-depth interviews.

Our analysis establishes four sets of findings: First, switching to participatory pedagogy successfully improved overall student tests scores two years later by 0.15σ (p-value=0.018), and the share of students with top grades increased by 6 percentage points from 16 to 22 percent (p-value=0.013). Point estimates for pass rates are positive too, but do not reach statistical significance (p-value=0.117). These effects are particularly remarkable considering that we used data from official national tests that were not specifically tailored to the intervention. Our complementary data shows that treatment teachers did indeed apply a wide range of the participatory pedagogical strategies taught in the training, such as group work (observed in 87% of classroom visits), games (28%) or dialogue (26%), and expressed great enthusiasm for the program in in-depth interviews.

Second, students who were taught by teachers equipped with laptops and CAL software did not outperform students whose teachers only participated in the pedagogical intervention. Point estimates for the difference between the teacher in-service training with and without supplying the CAL software are small and statistically insignificant. While teachers receiving the laptop with CAL software markedly improved their understanding of concepts related to the subdomain of number sense and arithmetic by 0.22σ (p-value=0.058), the effect on an overall score of math proficiency is statistically insignificant (p-value=0.135). The average teacher achieved 78 percent correct answers at baseline, suggesting that many teachers were already sufficiently proficient in their subject before the intervention.[1] This is in line with results from our heterogeneity analysis showing that the CAL based refresher was significantly more effective for teachers with low content knowledge at baseline.

Third, we do not find evidence for spillovers on indirectly exposed teachers and students in treatment schools, even though the program was specifically designed to produce such externalities. Although trained teachers and their peers self-reported that they engaged in cascading activities such as model lessons and peer learning groups, estimates for spillover effects at both the student and the teacher level are close to zero and statistically insignificant (p-value=0.403).

Fourth, the data from our complementary analyses allows us to compare participants' views about impacts of the program with the actual causal estimates from the RCT. We observe that participants' survey and interview responses are not very informative about what aspects of the program did or did not work, as respondents gave equally positive evaluations for all of them. For example, while 74 percent of the trained teachers strongly agree with the statement that the program improved their pupils' math skills, so do 78 percent of their indirectly exposed colleagues, even though we do not find any indication for such spillovers in our experimental data.

Our study contributes to a growing body of literature on how to address the learning crisis in developing countries. A vast spectrum of approaches has been evaluated in recent decades (see, e.g., Kremer et al., 2013; Glewwe and Muralidharan, 2016; World Bank, 2018, for an overview), but one key factor has received surprisingly little attention: teachers. Closing the global learning gap will crucially depend on how teachers in low- and middle-income countries perform in the classroom. The pivotal role of teachers in developing countries has been appreciated by recent studies focusing on the role of teacher incentives and pay, including De Ree et al. (2018), Duflo et al. (2012), Mbiti et al. (2023), and Muralidharan and Sundararaman (2011). Yet, the teacher performance not only depends on the economic incentives instructors face, but also on the repertoire of teaching strategies they have at their disposal.

A common strategy pursued by many development agencies is the promotion of a more student-centered pedagogy. Our study provides support for this approach, suggesting that attending five days of training in participatory pedagogy can be enough for teachers to restructure their classes and achieve higher learning gains for their students – even when their classes are large and few teaching aids are readily available. Promoting more engaging teaching strategies in low- and middle-income

[1] It is noteworthy that this substantially higher than the performance of teachers in El Salvador who averaged 47 percent on an almost identical assessment (Brunetti et al., 2020).

countries may thus be an essential element in the global quest for "inclusive and equitable quality education" (UN, 2015).

Our paper also ties into a nascent strand of literature studying complementarities in the educational production function (e.g., Mbiti et al., 2019). Our findings suggest that shortfalls in teacher content knowledge are unlikely to constitute a binding constraint to effective teaching in Tanzanian primary schools. Teachers already exhibited considerable subject mastery, and the pedagogy intervention was at least equally successful in improving student learning without simultaneously addressing shortfalls in content knowledge.

We also add to the literature on treatment externalities. The canonical example for treatment externalities in education was documented by Miguel and Kremer (2004), where treating students with de-worming pills produced large spillovers on non-targeted children such as younger siblings. Such treatment externalities can drastically boost the cost-effectiveness of an educational program, a fact that has given rise to so called *cascading models* to deliberately include the promotion of spillovers in program designs. Our findings suggest that in the context of pedagogical interventions, achieving such externalities may not be straightforward. A possible explanation is that teachers need a considerable degree of (first-hand) exposure to the new teaching strategies to be able and willing to effectively restructure their classes.

Finally, this paper contributes on the methodological discussion on how best to evaluate programs (Banerjee and Duflo, 2009; Garbarino and Holland, 2009). While qualitative methods such as surveys and interviews provide important insights and fruitfully complement experimental data, our findings suggest that they may be ill-equipped to assess the impact of a program and distinguish between its successful and less successful elements. This highlights the importance of quantitative analysis to learn what actually works rather than relying on people's self-reports about it.

2 Context and Intervention

Our study is set in Tanzania, a lower-middle income country in East Africa. Tanzania's education system faces several challenges that are typical for developing countries. The massive expansion of schooling starting in the late nineties has put considerable strain on schools throughout the country, and resulted in shortages of teachers, classrooms and teaching materials. Consequently, the pupil-teacher ratio in primary schools stands at 51 students per instructor (UNESCO, 2022). In this context, the country has struggled to translate enrollment into learning. For example, about sixty percent of students in grade 3 are unable to read and understand a simple paragraph (Sumra et al., 2015). Learning outcomes crucially depend on what teachers do in the classroom. However, a recent study finds that only 36 percent of teachers in Tanzania possess the minimum pedagogical knowledge needed for effective teaching (Bold et al., 2017a).

The program we study in this paper was implemented by Helvetas, a large Swiss development organization focusing on building capacity in Africa, Asia, Latin America and Eastern Europe. Helvetas has been active in Tanzania for more than 50 years with projects in a broad range of fields including agriculture, youth employment, and education. After several years of piloting teacher professional development at small scale, Helvetas, the Tanzanian Teachers' Union (TTU), and the Ministry of Education jointly launched the SITT program (Inclusive School-Based In-Service Teachers Training) aiming at transforming pedagogy in Tanzanian classrooms. Prior to the experimental evaluation we discuss in this paper, the program had already been rolled out in $1,430$ schools throughout North-eastern Tanzania.

The aim of the program is to promote a more *student-centered approach* to teaching that fosters active participation among pupils. This involves activities such as group work or students taking turns with the teacher to explain concepts in front of the class. To make classes more accessible and relevant to students, teachers are encouraged to incorporate practical examples from everyday life. Through the use of inexpensive local materials such as berries, stones or toothpicks, teachers also learn how to address shortages in high-quality teaching aids. These strategies are conveyed to teachers and to the responsible government officials through a centrally organized five-day workshop. After the initial training, teachers are invited to participate in biannual two-day refresher meetings, where the application of the strategies is discussed and experiences are shared. As a guide throughout the school year, each teacher receives a comprehensive manual summarizing the teaching strategies.

In the spirit of a *cascading model*, participating teachers are also encouraged to share their knowledge with all other teachers in their schools through different collaborative activities. Most importantly, they are expected to invite their colleagues to model lessons to showcase the new teaching methods in action. Trained teachers also have to organize peer learning groups where their peers can discuss their impressions from the model lessons and share their experience with the new pedagogical techniques in their own teaching. Finally, teachers are encouraged to manage large classes as a team to promote cooperative behavior and joint learning. The implementation of the new teaching strategies and the cascading activities is overseen by government quality assurance officers and the Helvetas team through monitoring visits to targeted schools. As an indirect monitoring tool, teachers are added to a "WhatsApp" group where they are expected to share their experiences.

In 2020, the intervention was supplemented by additional activities to address potential shortfalls in teachers' content knowledge. In this context, half the teachers participated in an extended version of the program where they received a laptop equipped with a *computer-assisted learning* software. Learning materials included video content and short quizzes in Swahili produced by Khan Academy and were provided through the offline-first learning platform Kolibri developed by Learning Equality. Learning videos were typically around 5 to 10 minutes long and structured into three broad themes, *(i)* Number Sense and Elementary Arithmetics (NSEA, 80 videos), *(ii)* Geometry and Measurement (GEOM, 80 videos), and *(iii)* Data, Statistics and Probability (DSP, 11 videos). Videos were shared through a user-friendly interface and complemented with short quizzes. Each quiz drew on a basis of roughly 20 items that were presented in random order. Upon submitting an answer, users received instant feedback. The software tracked performance and awarded badges of success for quizzes with at least five correct answers. Previous studies have shown computer-assisted studying with Khan Academy to be effective at improving test scores of both students Büchel et al. (2022) and teachers (Brunetti et al., 2023).

3 Research Design

3.1 Sampling and Randomization

To assess the impact of the in-service teacher training, we conducted a randomized controlled trial with a sample of 220 public primary schools in the Tanzanian districts in of Mbulu DC, Mbulu TC, Karatu and Siha, where the program had not been introduced yet. The implementing organization adopted a selection protocol similar to earlier implementation phases by excluding the best performing and the geographically least accessible schools in each district.

The experimental design allows to distinguish between *direct effects* on participating teachers and their pupils as well as *cascading effects* on peer teachers and their pupils. Specifically, selected schools nominated two teachers for the study: one *targeted teacher* for possible program participation and one *peer teacher* who was included for the estimation of spillovers. The selection of both targeted and peer teachers was done in coordination with the district education office and tied to the conditions *(i)* that both teachers should instruct math, and that *(ii)* the *targeted teacher* should teach math to sixth grade pupils in 2020 and seventh grade pupils in 2021. This procedure yielded a total sample of 440 teachers from 220 schools.

After the selection of schools and teachers, the research team randomly assigned each of the 220 schools to one out of three experimental conditions (see Figure 1):

- Pedagogy (65 schools, 130 teachers): Targeted teachers participated in the pedagogy training and were instructed to share their knowledge with their colleagues at their school.

- Pedagogy + Content (65 schools, 130 teachers): Targeted teachers participated in the pedagogy training and were instructed to share their knowledge with their colleagues at their school. They also obtained a laptop with computer-assisted learning software to self-study math.

- Control (90 schools, 180 teachers): Targeted teachers did not participate in any intervention activities.

Randomization was conducted after the nomination of teachers and the baseline data collection, and was stratified along three dimensions: district of school, baseline performance of pupils (i.e., school average in the standard 4 national examinations in 2018), and baseline performance of targeted teachers (i.e., math assessment conducted in November 2019).

3.2 Data

We rely on nationally standardized tests to measure effects on students, and conducted our own assessments to study intermediate effects on teachers. This experimental data is complemented with qualitative data we collected through classroom observations, surveys, and interviews in the treatment group.

Student assessments. The National Examinations Council of Tanzania (NECTA) conducts two standardized national student assessments that can be leveraged for this study: the *Primary School Leaving Examination* (PSLE) administered in grade 7, and the *Standard Four National Assessment* (SFNA) administered in grade 4. These yearly assessments are conducted with the entire student population in the respective grades and have high stakes: failing SFNA requires pupils to repeat grades, and passing PSLE is mandatory for admission in secondary school. Both assessments cover various subjects, but we rely on math scores for the main analysis. The math module in PSLE consists of 45 items that need to be completed in two hours, and SFNA includes 25 math questions students

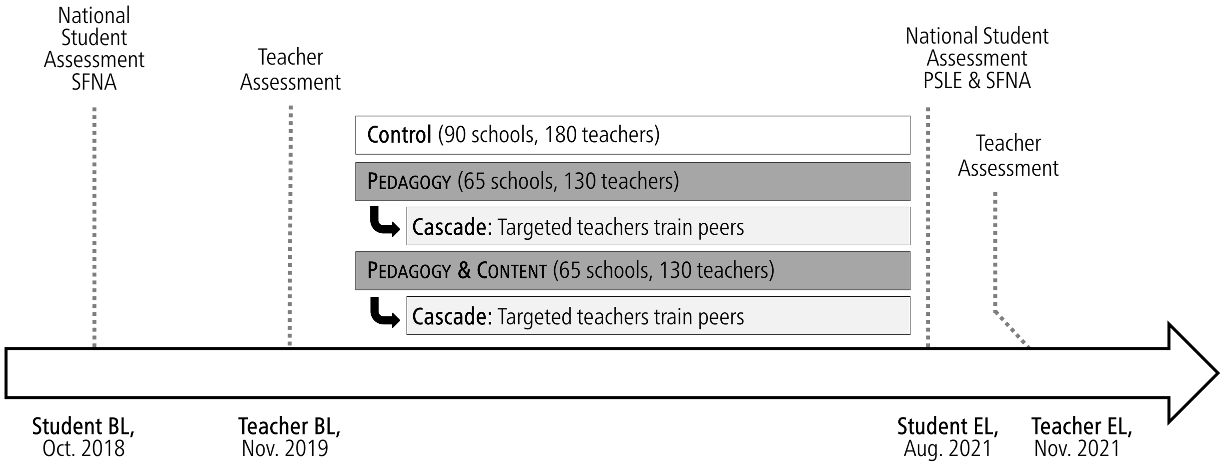

National Student Assessment SFNA
Teacher Assessment
National Student Assessment PSLE & SFNA
Teacher Assessment
Control (90 schools, 180 teachers)
PEDAGOGY (65 schools, 130 teachers)
Cascade: Targeted teachers train peers
PEDAGOGY & CONTENT (65 schools, 130 teachers)
Cascade: Targeted teachers train peers
Student BL, Oct. 2018
Teacher BL, Nov. 2019
Student EL, Aug. 2021
Teacher EL, Nov. 2021

Figure 1: Timeline of the study.
The main intervention event is a five day workshop for all treated teachers and was conducted in February 2020. Afterwards, teachers implemented the new strategies and share them with their colleagues, participated in biannual meetings, and were visited by quality assurance officers of the Ministry of Education. The *National Standard Four Assessment* (SFNA 2018, SFNA 2021) and the *Primary Standard Leaving Examination* (PSLE 2021) are conducted by the Tanzanian government and the results are published online, see `https://onlinesys.necta.go.tz/`.
Source: Own representation.

need to answer in 90 minutes (NECTA, 2018, 2020). Assessment data is publicly available at the student level.

Our *main outcome measure* is the PSLE math score of seventh graders in 2021, the cohort taught by targeted (and potentially trained) teachers in 2020 and 2021. Pupils' PSLE scores can be merged with their SFNA scores from three years earlier (i.e., 2018) to establish a pupil-level baseline score. To assess spillover effects through *cascading*, the SFNA math scores from grade four pupils in 2021 can be used, as these pupils were taught by peer teachers in the same school who were exposed to cascading activities.[2] No baseline data is available in this case. As both PSLE and SFNA results are published online, we use web scraping to obtain the student-level data. Our final sample consists of 10,101 seventh graders to assess the direct effects of the programs and 15,023 fourth graders to estimate spillovers.

Teacher assessments. To measure teacher content knowledge in math, all 440 study participants were invited to two comprehensive math assessments conducted before and after program implementation. The assessments were designed to mirror the Tanzanian primary school curriculum between grade 2 and grade 7 and covered the domains Number Sense & Elementary Arithmetics (NSEA, about 60%), Geometry & Measurement (GEOM, about 35%), and Data, Statistics, & Probability (DSP, about 5%). Assessments were administered as paper-and-pencil tests in regional meet-ups and had to be completed in 90 minutes.

[2] Note that standard 4 pupils were not necessarily taught by the one peer teacher who was chosen to participate in the teacher assessments. However, this is irrelevant for the study of pupils' learning outcomes as cascading activities are explicitly targeted at all teachers in a school and hence should impact learning across all grades and classrooms in a program school.

Complementary qualitative data. We collected three different types of qualitative data to get deeper insight into how switching to participatory pedagogy was viewed and put in practice by treated teachers. First, all teachers had to fill in a short survey about their evaluations of the program and their perceptions about how it had impacted them and their students. The survey primarily included single-choice questions, where respondents could rate certain elements or indicate whether they agreed or disagreed with a given statement, but also featured space for written feedback and suggestions. Survey forms were administered during the endline math assessment to all teachers and tailored to the different experimental groups.[3] Second, to better understand how teachers incorporated the new methods into their classes, quality assurance officers of the education ministry conducted classroom observations in lessons of program participants. Based on the Teach tool proposed by the World Bank (2019), a monitoring questionnaire was designed and government officials were briefed on how to conduct the classroom observations. Overall, 112 visits to treated teachers were conducted. Third, to complement the surveys and interviews, six participants of the Pedagogy intervention (about 120 min. audio recordings), six teachers from the Pedagogy & Content group (about 120 min. audio recordings), six peer teachers (about 70 min. audio recordings), and twelve government or TTU officials (about 150 min. audio recordings) participated in *semi-structured interviews*.[4]

3.3 Baseline characteristics, compliance, and attrition

Table A.1 in the appendix shows that *baseline characteristics* are well-balanced across the three experimental groups. The average teacher in our sample scored 78 percent correct answers on the math test we administered prior to the intervention. As the test was designed to cover the Tanzanian primary school curriculum, this suggests that, on average, teachers master three quarters of the materials they have to teach. About 4 in every 10 teachers in our sample are female and the average teacher is 38 years old. Panel 2 on school characteristics shows that the typical class size is about 40 students.[5] The number of students that took the SFNA exam, roughly 50 per school, provides a proxy for the number of students per grade. As this figure is not much higher than the average class size, most schools can be assumed to have only one class per grade. Most importantly, pupils' baseline scores are well-balanced across experimental groups. On average, about 67 percent of students passed the baseline math exam, and 40 percent of students scored one of the two top grades (A or B).

Our monitoring data suggests that *compliance* with the treatment assignment was very high. All

[3] We designed four different questionnaires: (1) a questionnaire for teachers in the Pedagogy treatment with items about the training and the implementation of the new methods, (2) a similar questionnaire for the Pedagogy + Content group with additional questions about the content training with the laptops, (3) a questionnaire for peer teachers asking about casacading activities, and (4) a short questionnaire for the control group with questions about the evaluation process. With the exception of the control group, the different survey versions followed the same basic structure and had many common items, allowing for comparison across different groups.

[4] During these conversations, the interviewees were asked *(i)* to share their general impression of the intervention, *(ii)* to explain their view on the main elements of the Pedagogy intervention, *(iii)* to share their assessment on the impact of the program on teachers' math and teaching skills as well as the learning outcomes of children, and *(iv)* to give feedback on selected activities and program inputs; additionally, officials were asked *(v)* to compare the pedagogical intervention with similar educational initiatives by other organizations, and *(vi)* to comment on their attitude towards rigorous program evaluation. Table D.1 in the appendix section D provides an overview of statements by topic and type of interviewee.

[5] While information on the number of pupils per classroom is difficult to collect, the number of pupils per *stream* can serve as a proxy. In Tanzania the concept of a "class" is surprisingly blurry because several streams of pupils can be instructed in one classroom (and effectively become one class) if schools do not have enough classrooms or teachers to teach streams separately.

teachers in the treatment group participated in the five-day teacher training, and 94 percent of the teachers in the Pedagogy & Content group report having used the laptops for content revision. To be able to assess the impact of the program using students' tests scores in grade 7, targeted teachers had to teach math to all sixth graders in their school in 2020 and to all seventh graders in 2021. Our data collected during the endline teacher survey shows that 85 percent of the students in the treatment group were indeed taught by targeted teachers. This share does not differ significantly between experimental groups.

Tables A.2 and A.3 examine patterns of *attrition* for teachers and students respectively. At the teacher level, 99 percent of the selected teachers took part in the baseline assessment, and attrition for the endline assessment was about 15 percent and evenly distributed across experimental groups. This yields a total sample size of 368 teachers. At the student level, we start with baseline data for 12, 657 pupils from 220 schools. About 17 percent of these students either dropped out of school between grade 4 and grade 7, missed the endline examination, or could not be matched between the two examination rounds. Moreover, one school dropped out because the targeted teacher missed both the base- and endline data collection. Finally, an estimation sample with 10, 101 seventh graders from 219 schools remains. Both for teachers and pupils, attrition was unrelated to the experimental assignment. For the estimation of spillovers, we can use a sample of 15, 023 grade 4 students from 220 schools. Due to the unavailability of baseline data, we cannot study the attrition for this cohort of students.

4 Results

4.1 Did promoting participatory teaching strategies improve learning?

We estimate the *intent to treat* (ITT) effect on students of directly targeted teachers with the following benchmark equation

$$Y_{isk}^{PSLE} = \beta Treatment_s + X_i^{'}\gamma + V_s^{'}\lambda + \phi_k + \epsilon_{isk}, \tag{1}$$

where Y_{isk}^{PSLE} is the standardized math PSLE score of student i in school s and stratum k at endline, and *Treatment* is a binary indicator that takes the value of 1 if a school was assigned to the treatment group and is 0 otherwise. Student level controls, X_i, comprise sex, baseline math score, and average baseline score across all subjects taken from the SFNA baseline assessment. V_s represents a vector of school-level controls including the number of students who took the baseline assessment, the average PSLE score at baseline[6], the driving distance to the district headquarters and the class size, as well as the math score, sex, and age of the targeted teacher. ϕ_k stands for k strata fixed effects, and ϵ_{isk} represents the error term.

The results in Table 1 document that students in treated schools significantly outperformed the control group by 0.15σ (column 2). Pupils in program schools were also up to 6 percentage points more likely to achieve a top grade (i.e., A or B) than their peers in control schools (columns 3 and 4). This corresponds to an increase in top grades by 36 percent. Estimates in columns 5 and 6 further

[6] Note that this is not the average score of the cohort we study, but that of a previous cohort of seventh graders in the school.

Table 1: Overall program effect on the math score of pupils

	Standardized		Scored A or B		Passed	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.107^{+}	0.145^{*}	0.046^{*}	0.056^{*}	0.023	0.036
	(0.062)	(0.061)	(0.023)	(0.022)	(0.023)	(0.023)
Pupil baseline math score	0.466^{**}	0.327^{**}	0.121^{**}	0.082^{**}	0.210^{**}	0.155^{**}
	(0.017)	(0.021)	(0.008)	(0.008)	(0.008)	(0.010)
Mean of dep. variable	-0.008	-0.008	0.155	0.155	0.592	0.592
Observations	10101	10101	10101	10101	10101	10101
Adjusted R^2	0.252	0.295	0.146	0.180	0.202	0.224
Controls	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is pupils' standardized math score for columns (1) and (2), a binary variable indicating whether a pupil scored A or B (highest grades) in math for columns (3) and (4), and a binary variable indicating whether a pupil passed the math exam for columns (5) and (6). *Pupil baseline math score* is a pupil's score in the SFNA exam administered in grade 4. Controls include *(i) pupil-level controls* for average SFNA baseline score across all subjects and sex, *(ii) school-level controls* for average PSLE baseline score (all subjects), class size, and number of pupils in grade 4 and *(iii) teacher-level controls* for sex, age, and math performance at baseline. Huber-White robust standard errors, clustered at the school level, in parentheses. + p<0.10, * p<0.05, ** p<0.01.

suggest that the program induced a 2 to 4 percentage point increase in pass rates, but these effects are not statistically significant at conventional levels.

Table A.5 in the appendix examines effects on students' average score across all subjects rather than their math score. Results are very similar, with estimated effects of 0.12σ and an increase in top grades by 7 percentage points or 30 percent. This suggest that although the pedagogical training was tailored to math, teachers were able to transfer the methods to other subjects.

Overall, the observed impacts are comparable to effects documented in RCTs of similar programs (see Snilstveit et al., 2015; McEwan, 2015). Unlike most other studies, our analyses are based on standardized national assessments that are not tailored to the intervention under study, which strengthens their external validity.

Our causal estimates are consistent with insights from our complementary data sources. Classroom observations point to a widespread use of the participatory teaching strategies advertised through the training program. As Figure C.1 in the appendix shows, treated teachers frequently applied methods such as group work (87% of visits), games (28%), student presentations (28%), and dialogues (26%). Treatment teachers also used a wide range of teaching materials, including daily life objects (66% of visits), textbooks (46%), and flash cards (20%). The survey data further shows that 96 percent of treated teachers rate the participatory teaching model as excellent (75%) or good (21%). Similarly, 96 percent of targeted teachers strongly (74%) or rather agree (22%) with the statement that the intervention improved their students' math scores. The high appreciation for the program also surfaced in the interviews where teachers often used words such as *"improve"*, *"change"*, and *"enjoy"* when talking about the intervention (see Table D.1 in the appendix).

To better understand under which circumstances the participatory teaching methods promoted through the training work best, it is informative to take a look at how effects vary by characteristics

Table 2: Program effect on the math score of pupils by implementation version

	Standardized		Scored A or B		Passed	
	(1)	(2)	(3)	(4)	(5)	(6)
T1: Pedagogy	0.127	0.147^{*}	0.056^{+}	0.059^{*}	0.024	0.033
	(0.081)	(0.071)	(0.029)	(0.026)	(0.028)	(0.026)
T2: Pedagogy & Content	0.086	0.142^{+}	0.034	0.052^{+}	0.022	0.039
	(0.072)	(0.073)	(0.026)	(0.027)	(0.029)	(0.028)
Pupil baseline math score	0.466^{**}	0.327^{**}	0.121^{**}	0.082^{**}	0.210^{**}	0.155^{**}
	(0.017)	(0.021)	(0.008)	(0.008)	(0.008)	(0.010)
T2 − T1	-0.041	-0.005	-0.022	-0.008	-0.002	0.006
	(0.090)	(0.075)	(0.033)	(0.028)	(0.033)	(0.030)
Mean of dep. variable	-0.008	-0.008	0.155	0.155	0.592	0.592
Observations	10101	10101	10101	10101	10101	10101
Adjusted R^2	0.252	0.295	0.147	0.180	0.201	0.224
Controls	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is pupils' standardized math score for columns (1) and (2), a binary variable indicating whether a pupil scored A or B (highest grades) in math for columns (3) and (4), and a binary variable indicating whether a pupil passed the math exam for columns (5) and (6). *Pupil baseline math score* is a pupil's score in the SFNA exam administered in grade 4. Controls include *(i) pupil-level controls* for average SFNA baseline score across all subjects and sex, *(ii) school-level controls* for average PSLE baseline score (all subjects), class size, and number of pupils in grade 4, and *(iii) teacher-level controls* for sex, age, and math performance at baseline. Huber-White robust standard errors, clustered at the school level, in parentheses. + p<0.10, * p<0.05, ** p<0.01.

of classes, teachers and pupils. A key challenge for productive student engagement is posed by the typically very large classes in Tanzania. According to Table A.7, the impact of the interventions decreased with larger class sizes, but these effects are not statistically significant (columns 7 and 8). A further concern might be that the use of participatory teaching methods demands a high level of skills on the part of the teachers. We do not observe teachers' pedagogical skills, but their performance in the math test can serve as a proxy. Indeed, treatment effects appear to be larger for students who are taught by better-performing teachers (columns 5 and 6). Additional analyses by pupils' gender and initial performance levels do not point towards relevant effect heterogeneity along these dimensions.

4.2 Did the computer-based content training yield additional benefits?

We also estimate the effects of each program version separately, using

$$Y_{isk}^{PSLE} = \beta_1 T1_s + \beta_2 T2_s + X_i^{'}\gamma + V_s^{'}\lambda + \phi_k + \epsilon_{isk}, \qquad (2)$$

where $T1_s$ is a binary indicator for the Pedagogy intervention, and $T2_s$ indicates whether a treated teacher's school was additionally assigned to the content training component, i.e. to Pedagogy & Content.

As Table 2 shows, we do not find that providing laptops for content revision in addition to the pedagogical training yielded further learning gains for students. If anything, the point estimate for the extended intervention is slightly lower, but this difference is not significant.

One possible interpretation is that teachers did not use or appreciate the laptops for the intended purpose. Our complementary data suggests otherwise. Teachers report spending an average of 5 to 6 hours per week with the learning software, and provide very positive evaluations of the computer-assisted learning component with 68 percent rating it as excellent and 20 percent as good. The same affirmative feedback surfaced in interviews, where teachers unanimously expressed strong appreciation for the laptops and reported using them frequently for content revision or to prepare their lessons.

Another possibility is that teachers did use the laptops, but failed to meaningfully improve their content knowledge with the software. Figure 2 and Table A.8 present estimates for the causal impact of each intervention on teachers' content knowledge in math. Although teachers in the laptop group markedly improved their understanding of concepts related to NSEA by 0.22σ (columns 5 and 6 in Table A.8), the effect on an overall score of math proficiency is smaller (0.15σ) and misses conventional levels of statistical significance (columns 1 and 2).

A plausible interpretation for these modest effects is that most teachers already possessed good mastery of the primary school curriculum to begin with. As indicated in Figure A.2, the average teacher was able to answer 78 percent of the questions on materials covered in grades 2 to 7 correctly. While targeted teachers scored an average of 81 percent, peer teachers scored only 74 percent, suggesting that schools selected particularly well-performing teachers for program participation. Overall, 50 percent of the teachers pass the threshold for subject proficiency – at least 80 percent correct answers – advocated by the World Bank (Bold et al., 2017a). Only 2 percent of all teachers answered less than 50 percent of the questions correctly. A comparison with results from an almost identical assessment conducted with teachers in El Salvador suggests that the Tanzanian teachers perform considerably better than their counterparts in El Salvador (see Brunetti et al., 2020).[7] Hence, it appears plausible that many Tanzanian teachers are already sufficiently proficient in math for effective teaching at the primary school level. In line with this argument, Table A.9 in the appendix points to considerable effect heterogeneity by teachers' initial ability level. Low-performing teachers markedly improved their content knowledge (0.51σ, p = 0.004, for teachers below the median) due to the intervention, but these effects decline significantly as teachers' baseline scores improve, and are close to zero for high-performing teachers (not shown).

Hence, from an impact evaluation perspective, the additional investment in the IT equipment for content revision clearly did not pay off. Although we provide suggestive evidence that low-performing teachers used the software to catch up with their better-prepared colleagues, we do not find that such gains were transferred to students.

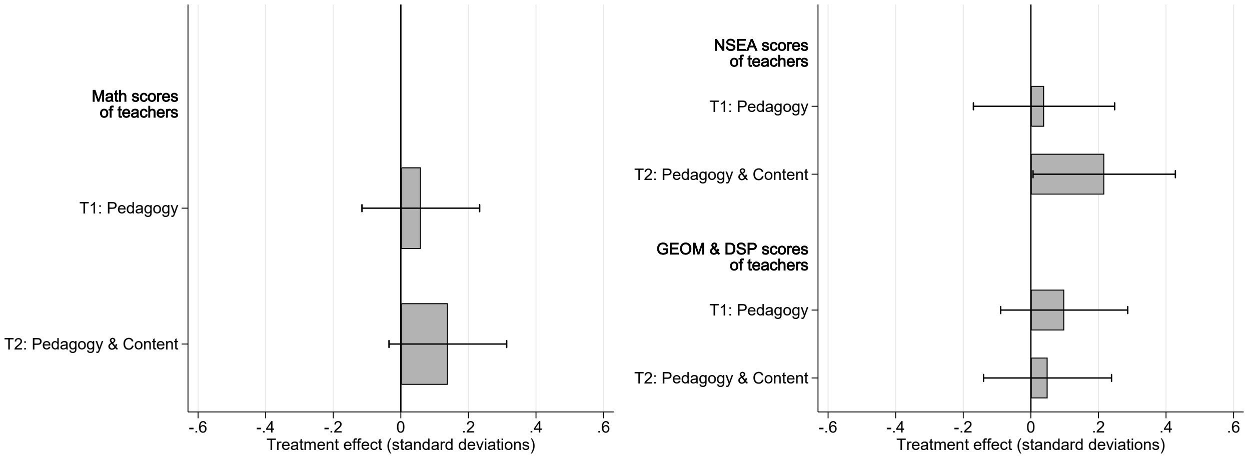
4.3 Did the interventions produce externalities for indirectly exposed students and teachers?

To estimate spillovers on indirectly exposed fourth-graders rather than directly exposed seventh graders, we use the following slightly adapted version of equation (1)

$$Y_{isk}^{SFNA} = \beta Treatment_s + X_i^{'}\gamma + V_s^{'}\lambda + \phi_k + \epsilon_{isk}, \tag{3}$$

where Y_{isk}^{SFNA} is the standardized math SFNA score of student i in school s and stratum k at

[7] The average teacher in the El Salvador study scored 47 percent on a math test covering materials from grades 2–6 and only 14 percent of teachers achieved at least 80 percent correct answers.



(a) Effects on overall test score

(b) Effect on test scores by domain

Figure 2: Treatment effects on teachers' overall and domain-specific math scores
Estimates for the effect of the two intervention versions on *targeted teachers* are shown. *Controls* include baseline score, sex, age, and years since graduation at baseline. 90 percent confidence intervals shown. For more information on the sample size and the estimation strategy, see Table A.8.

endline. As no nationally standardized assessment results are published for students below grade four, we include the school-level SFNA score as a baseline performance measure.

Table 3 examines spillover effects on students whose teachers were indirectly exposed to the treatment through peer learning activities in their school. In all specifications, estimates are close to zero and insignificant. In line with the moderate direct effects of the additional content training, we also find no indication of content knowledge spillovers at the teacher level, as Table A.8 shows.

A possible explanation for the absence of meaningful treatment externalities is that the observation period of our study was not long enough to capture effects on students of indirectly exposed teachers. Due to the time lag between the initial teacher training and the cascading activities, peer teachers may not have had sufficient time to put the new techniques into practice. To assess the plausibility of this hypothesis, we can draw on non-experimental data from the implementation phase 2013 to 2019, i.e. the period prior to the execution of the field experiment. Using both the PSLE and the SFNA scores for these years, we conduct a difference-in-difference analysis to assess the impact of the program over a longer time horizon (see Appendix B). As only one out of many teachers in each intervention school participated in the teacher training and all other teachers were indirectly exposed through cascading activities, our estimates correspond to an upper bound for spillover effects at the school level. As Table B.1 in the appendix shows, we find no indication for such effects.

Another possibility is that the knowledge sharing activities were not conducted. Again, our complementary data suggests otherwise. Almost all targeted teachers report organizing the model lessons (95%) and the peer learning groups (96%), and most peer teachers report participating in these activities (88% for both model lessons and peer learning groups), with the average peer teacher claiming to have attended 3.8 model lessons. Moreover, the knowledge sharing activities are rated very positively by both targeted and peer teachers.[8]

[8] This should not be seen as conclusive evidence for the successful implementation of the cascading elements as teachers may have succumbed to a common tendency of giving socially desirable, but dishonest answers. Indeed, in the in-depth interviews, teachers provided slightly more critical feedback on the cascading elements, with some interviewees mentioning challenges regarding their implementation due to the lack of interest of some of their colleagues.

Table 3: Cascading effect on the math score of pupils

	Standardized		Scored A or B		Passed	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.028	0.037	0.011	0.012	0.011	0.016
	(0.048)	(0.044)	(0.009)	(0.009)	(0.021)	(0.019)
School PSLE avg. score (std)	0.081^{*}	0.083^{**}	0.016^{**}	0.018^{**}	0.034^{**}	0.033^{**}
	(0.031)	(0.031)	(0.006)	(0.006)	(0.013)	(0.012)
School SFNA avg. score (std)	0.134^{**}	0.129^{**}	0.022^{**}	0.022^{**}	0.048^{**}	0.046^{**}
	(0.031)	(0.031)	(0.006)	(0.006)	(0.013)	(0.013)
Mean of dep. variable	-0.000	-0.000	0.075	0.075	0.368	0.368
Observations	15023	15023	15023	15023	15023	15023
Adjusted R^2	0.072	0.080	0.035	0.040	0.053	0.060
Controls	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is pupils' standardized SFNA math score for columns (1) and (2), a binary variable indicating whether a pupil scored A or B (highest grades) in math for columns (3) and (4), and a binary variable indicating whether a pupil passed the math exam for columns (5) and (6). *School-level baseline scores* are the school's average scores in the SFNA exam administered in grade 4 and the PSLE exam administered in grade 7. Controls include *(i) pupil-level controls* for sex, *(ii) school-level controls* for the number of pupils in grade 4 and *(iii) teacher-level controls* for sex, age, and math performance at baseline. Huber-White robust standard errors, clustered at the school level, in parentheses. + p<0.10, * p<0.05, ** p<0.01.

Hence, a more likely explanation is that although the cascading activities were conducted, they did not provide sufficient exposure to the new pedagogical techniques for peer teachers to effectively restructure their classes.

4.4 How informative are participants' self-reports about the impact of different program aspects?

An ongoing debate in the development community concerns the merits of two distinct evaluation traditions: a quantitative paradigm emphasizing causal inference methods and a qualitative tradition focusing on the experiences of project stakeholders (e.g., Banerjee and Duflo, 2009; Garbarino and Holland, 2009). The main contribution of this paper is quantitative, but we can also combine and compare our experimental findings with insights from qualitative surveys and interviews with project beneficiaries. In particular, we asked all participating teachers to assess the effect of the intervention on different outcomes, allowing us to contrast these self-reports with the actual causal effects we identified through the experiment (see Table 4).

Across all the outcomes and groups we study, participants are very confident about the impact of the intervention. While this is in line with the positive causal impact we report, response patterns appear to be unrelated to the success and failure of different project components. Most notably, directly participating teachers and peer teachers are equally optimistic about the impact of the intervention on their math skills and those of their students, even though we find no indication for spillover effects in our data. Similarly, we report no effect of the PEDAGOGY intervention on teachers' math skills, but 87 percent of teachers in this group strongly agree with the claim that they improved these skills

Table 4: Comparison between observed causal effects and participants' reported beliefs

	RCT: Observed impact	**Survey: Participants' beliefs about impact**
Impact of intervention on student learning	Significant effect of 0.15 SD*	Did the project improve the math skills of your pupils? Strongly agree: 74%, rather agree: 22%
Spillovers of intervention on students of peer teachers	Effect insignificant and close to zero	Did the project improve the math skills of your pupils? Strongly agree: 78%, rather agree: 19%
Impact of PEDAGOGY *intervention on teachers' math skills*	Effect insignificant and close to zero	Did the project improve your math skills? Strongly agree: 87%, rather agree: 5%
Impact of PEDAGOGY & CONTENT *intervention on teachers' math skills*	Effect of 0.15 SD, but insignificant	Did the project improve your math skills? Strongly agree: 85%, rather agree: 11%
Spillovers of intervention on peer teachers' math skills	Effect insignificant and close to zero	Did the project improve your math skills? Strongly agree: 81%, rather agree: 15%

thanks to the intervention. Finally, teachers rated the self-studying with the laptops very positively, but we find only limited evidence for its effects at the teacher level and no evidence for an impact on students.

These findings tie into a nascent literature studying biases in evaluations (e.g., Camfield et al., 2014). Two broad explanations accounting for participants' overoptimistic impact assessments can be distinguished. First, people's capacity for counterfactual thinking is limited, leading them to misattribute outcomes or changes in their lives to the programs they participated in (e.g., McKenzie, 2018). Comparing actual and self-reported effects in three labor market interventions, Smith et al. (2021) conclude that participants act as "lay scientists". Their assessments are largely unrelated to the actual causal impact estimated for their group, but tend to follow coarse heuristics for this impact such as unconditional outcomes or before-after comparisons. A second well-documented bias in social science research, known as courtesy bias, social desirability bias or experimenter demand effects, is a general tendency of subjects to provide answers they perceive as aligning with the researcher's expectations (Camfield et al., 2014; Krumpal, 2013; Zizzo, 2010). In project evaluation, the resulting pro-project bias is likely to be exacerbated if people believe that the evaluation will determine whether the project is continued. Our findings are in line with these biases and suggest that while qualitative evidence from participant surveys and interviews can provide a valuable complement to experimental evidence, it is ill-equipped for the assessment of causal impacts.

5 Conclusion

Addressing the learning global learning crisis calls for innovative strategies to track and improve education (e.g. Patrinos and Angrist, 2018; World Bank, 2018; Jakob and Heinrich, 2023). In this paper we turn our attention to the teachers, who are the key actors in the educational system. While previous research has strongly focused on the misaligned economic incentives teachers often face, this study is premised on the assumption that they could be using ineffective pedagogy. Through a randomized controlled trial with 440 teachers and about 25,000 students in Tanzania, we show that promoting participatory teaching strategies significantly improves students' learning outcomes by 0.15σ. Our findings are based on standardized national assessments conducted by the National Examinations Council of Tanzania and corroborated by evidence from our classroom observations and participant surveys affirming that teachers indeed implemented and appreciated the new participatory methods.

Our study also explores the potential of computer-assisted learning to improve teachers' content knowledge and, thereby, student learning. We find suggestive evidence that providing computers with a learning software helps low-performing teachers improve their math skills. However, this does not translate into measurable learning gains for their students. Previous research suggests that a 0.1σ gain student learning would require a 1σ improvement in teachers' content knowledge (Bau and Das, 2020; Metzler and Woessmann, 2012) – an unrealistically large effect for educational interventions. Our findings underscore that addressing shortfalls in teachers' content knowledge is not a low-hanging fruit for promoting student learning.

We report similarly discouraging results for spillovers on other teachers and their students through cascading activities. Cascading schemes are favored in the development community for their potential to increase the number of beneficiaries and extend a project's reach. However, our results suggest that producing measurable learning spillovers is not straightforward. More research is thus needed to explore if and how the promise of cascading can be realized in educational initiatives.

Nevertheless, even without relying on spillovers, building teacher competencies can be a very cost-effective approach to improve student learning in the long run. Teachers often remain in their profession for many years, influencing dozens of student generations. If they continue to apply the new teaching methods throughout their professional lives, pedagogical teacher training becomes a highly sustainable and cost-effective means to foster student learning. Hence, promoting participatory teaching could be a key ingredient to a comprehensive strategy to ensure that children in developing countries are not only going to school, but are actually learning.

References

Banerjee, Abhijit V and Esther Duflo. 2009. The experimental approach to development economics. *Annual Review of Economics* 1 (1):151–178.

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

Berlinski, Samuel and Matias Busso. 2017. Challenges in educational reform: An experiment on active learning in mathematics. *Economic Letters* 156:172–175.

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.

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.

Brunetti, Aymo, Konstantin Büchel, Martina Jakob, Ben Jann, and Daniel Steffen. 2023. Inadequate teacher content knowledge and what could be done about it: Evidence from El Salvador. *Journal of Development Effectiveness* :1–24.

Büchel, Konstantin, Martina Jakob, Kühnhanss Christoph, Daniel Steffen, and Aymo Brunetti. 2022. The relative effectiveness of teachers and learning software. Evidence from a field experiment in El Salvador. *Journal of Labor Economics,* 40 (3):737–777.

Callaway, Brantly and Pedro HC Sant'Anna. 2021. Difference-in-differences with multiple time periods. *Journal of Econometrics* 225 (2):200–230.

Camfield, Laura, Maren Duvendack, and Richard Palmer-Jones. 2014. Things you wanted to know about bias in evaluations but never dared to think. *IDS Bulletin* 45 (6):49–64.

Cornelius-White, Jeffrey. 2007. Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research* 77 (1):113–143.

De Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers. 2018. Double for nothing? Experimental evidence on an unconditional teacher salary increase in Indonesia. *The Quarterly Journal of Economics* 133 (2):993–1039.

Duflo, Esther, Rema Hanna, and Stephen P Ryan. 2012. Incentives work: Getting teachers to come to school. *American Economic Review* 102 (4):1241–78.

Garbarino, Sabine and Jeremy Holland. 2009. Quantitative and qualitative methods in impact evaluation and measuring results. Discussion Paper. University of Birmingham.

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*, eds. Eric Hanushek, Stephen Machin, and Ludger Woessmann. Amsterdam: Elsevier, 653–743.

Harbour, Kristin E, Lauren L Evanovich, Chris A Sweigart, and Lindsay E Hughes. 2015. A brief review of effective teaching practices that maximize student engagement. *Preventing School Failure* 59 (1):5–13.

Jakob, Martina Saskia and Sebastian Heinrich. 2023. Measuring human capital with social media data and machine learning. University of Bern Social Sciences Working Papers 46, University of Bern.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. The challenge of education and learning in the developing world. *Science* 340 (6130):297–300.

Krumpal, Ivar. 2013. Determinants of social desirability bias in sensitive surveys: A literature review. *Quality and Quantity* 47 (4):2025–2047.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania. *The Quarterly Journal of Economics* 134 (3):1627–1673.

Mbiti, Isaac, Mauricio Romero, and Youdi Schipper. 2023. Designing effective teacher performance pay programs: Experimental evidence from Tanzania. *The Economic Journal* 133 (653):1968–2000.

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.

McKenzie, David. 2018. Can business owners form accurate counterfactuals? eliciting treatment and control beliefs about their outcomes in the alternative treatment status. *Journal of Business & Economic Statistics* 36 (4):714–722.

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.

Muralidharan, Karthik and Venkatesh Sundararaman. 2011. Teacher performance pay: Experimental evidence from india. *Journal of Political Economy* 119 (1):39–77.

NECTA. 2018. Format for standard four national assessment. Tech. rep., National Examinations Council of Tanzania.

———. 2020. Format for primary school leaving examinations. Tech. rep., National Examinations Council of Tanzania.

Patrinos, Harry A and Noam Angrist. 2018. Global dataset on education quality: A review and update (2000-2017). *World Bank Policy Research Working Paper* (8592).

Seidel, Tina and Richard J Shavelson. 2007. Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research* 77 (4):454–499.

Sinha, Shabnam, Rukmini Banerji, and Wilima Wadhwa. 2016. *Teacher performance in Bihar, India: Implications for education.* Washington D.C.: The World Bank.

Smith, Jeffrey, Alexander Whalley, and Nathaniel Wilcox. 2021. *Are participants good evaluators?* WE Upjohn Institute.

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.

Sumra, Suleman, Sara Ruto, and Rakesh Rajani. 2015. Assessing literacy and numeracy in Tanzania's primary schools: The Uwezo approach. In *Preparing the next generation in Tanzania.*

UN, United Nations. 2015. The 2030 agenda for sustainable development. New York: United Nations.

UNESCO. 2022. Pupil-teacher ratio in primary school in 2018, Tanzania. Published online: `https://data.worldbank.org`.

World Bank. 2018. *World Development Report 2018: Learning to realize education's promise.* Washington D.C.: World Bank.

———. 2019. Teach. Our vision is to revolutionize how education systems track and improve teaching quality. World Bank Brief, published online: `www.https://www.worldbank.org/`.

Zizzo, Daniel John. 2010. Experimenter demand effects in economic experiments. *Experimental Economics* 13:75–98.

A Appendix: Additional results from experimental analysis

A.1 Baseline characteristics

Table A.1: Baseline characteristics

	Control	T1	T2	p-value
Panel 1: Teacher variables (N = 434)	(1)	(2)	(3)	(4)
Math score (percent correct)	77.390	78.523	77.606	0.644
	(0.874)	(0.914)	(0.994)	
Female	0.299	0.277	0.315	0.797
	(0.035)	(0.039)	(0.041)	
Age	38.203	38.654	36.984	0.285
	(0.676)	(0.816)	(0.757)	
Years since graduation	12.040	12.308	11.118	0.514
	(0.701)	(0.861)	(0.730)	
Panel 2: School variables (N = 219)				
Nr. of pupils that took SFNA	58.461	52.815	49.754	0.098
	(3.136)	(2.813)	(2.512)	
School PSLE avg. score (std)	-0.008	0.170	-0.096	0.323
	(0.103)	(0.119)	(0.146)	
Driving distance to district headquarters (h)	0.579	0.551	0.612	0.727
	(0.036)	(0.047)	(0.061)	
Nr. of pupils per class	43.574	39.755	40.377	0.309
	(2.022)	(1.540)	(2.037)	
Panel 3: Pupil variables (N = 10,101)				
Pupil math score (std)	-0.034	0.023	0.031	0.730
	(0.060)	(0.064)	(0.071)	
Pupil avg. score (std)	-0.007	0.028	-0.017	0.912
	(0.077)	(0.073)	(0.088)	
Pupil passed math exam	0.656	0.671	0.679	0.812
	(0.023)	(0.027)	(0.028)	
Pupil passed exam	0.764	0.788	0.742	0.544
	(0.026)	(0.027)	(0.033)	
Pupil scored A or B in math	0.390	0.422	0.421	0.603
	(0.025)	(0.027)	(0.029)	
Pupil scored A or B on avg.	0.359	0.378	0.371	0.901
	(0.031)	(0.031)	(0.035)	
Female pupil	0.523	0.512	0.504	0.293
	(0.008)	(0.009)	(0.009)	

Notes: Columns (1) - (3) report the mean for different covariates by experimental group (standard errors in parentheses). Column (4) reports the p-value of the F-test for differences in means across groups. Pupil baseline tests scores are taken from the *Standard Four National Examination* (SFNA), administered to all pupils in grade 4. School-level test scores from the *Primary School Leaving Examination* (PSLE), administered in grade 7, are used to assess the initial quality of the school.

A.2 Attrition at endline

Table A.2: Attrition of teachers at endline by experimental group

	All teachers		Targeted teachers		Peer teachers	
	(1)	(2)	(3)	(4)	(5)	(6)
T1: Pedagogy	0.006	0.004	0.032	0.039	-0.025	-0.036
	(0.038)	(0.037)	(0.056)	(0.055)	(0.054)	(0.054)
T2: Pedagogy & Content	0.057	0.046	0.044	0.041	0.064	0.042
	(0.046)	(0.047)	(0.059)	(0.061)	(0.065)	(0.065)
Baseline score	-0.014	0.004	-0.029	-0.017	-0.001	0.015
	(0.018)	(0.019)	(0.033)	(0.033)	(0.025)	(0.027)
Avg. attrition rate	0.151	0.151	0.146	0.146	0.156	0.156
Observations	434	434	219	219	215	215
Adjusted R^2	0.010	0.020	0.012	0.016	0.008	0.018
Controls	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Linear probability model estimating the impact of the treatments on attrition probability. Estimates reported for *all teachers* in columns (1) and (2), for *targeted teachers* in columns (3) and (4), and for *peer teachers* in columns (5) and (6). *Teacher level controls* include sex, age, and years since graduation. Huber-White robust standard errors in parentheses.
+ p<0.10, * p<0.05, ** p<0.01.

Table A.3: Attrition of pupils between SFNA 2018 and PSLE 2021 by experimental group

	Attrition	
	(1)	(2)
T1: Pedagogy	-0.011	-0.004
	(0.015)	(0.011)
T2: Pedagogy & Content	-0.011	-0.008
	(0.017)	(0.013)
Pupil baseline math score		-0.055**
		(0.007)
Observations	12657	11991
Adjusted R^2	0.018	0.044
Controls	No	Yes
Stratum fixed effects	Yes	Yes

Notes: Linear probability model estimating the impact of the treatments on attrition rates. Controls include *(i) pupil-level controls* for average SFNA baseline score across all subjects and sex, *(ii) school-level controls* for average PSLE baseline score (all subjects) and number of pupils, and *(iii) teacher-level controls* for sex, age, and math performance at baseline. Huber-White robust standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01.

A.3 Robustness checks for main effects at the student level

Table A.4: Robustness checks for effects on students' math scores

	Standardized				Scored A or B			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.11^{+}	0.11^{+}	0.13^{*}	0.14^{*}	0.05^{*}	0.05^{*}	0.05^{*}	0.06^{*}
	(0.06)	(0.06)	(0.06)	(0.06)	(0.02)	(0.02)	(0.02)	(0.02)
Pupil baseline math score	0.47^{**}	0.32^{**}	0.33^{**}	0.33^{**}	0.12^{**}	0.08^{**}	0.08^{**}	0.08^{**}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	10101	10101	10101	10101	10101	10101	10101	10101
Adjusted R^2	0.25	0.27	0.29	0.30	0.15	0.16	0.17	0.18
Pupil Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
School Controls	No	No	Yes	Yes	No	No	Yes	Yes
Teacher Controls	No	No	No	Yes	No	No	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is pupils' standardized math scores in all models. Controls include *(i) pupil-level controls* for average SFNA baseline score across all subjects and sex, *(ii) school-level controls* for average PSLE baseline score (all subjects), class size, and number of pupils, and *(iii) teacher-level controls* for sex, age, and math performance at baseline. Huber-White robust standard errors, clustered at the school level, in parentheses. + p<0.10, * p<0.05, ** p<0.01.

Table A.5: Program effect on the average score of pupils across subjects

	Standardized		Scored A or B		Passed	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.082	0.121^{*}	0.053^{*}	0.069^{**}	0.000	0.009
	(0.059)	(0.054)	(0.025)	(0.023)	(0.019)	(0.018)
Pupil baseline avg. score	0.499^{**}	0.306^{**}	0.175^{**}	0.105^{**}	0.152^{**}	0.095^{**}
	(0.021)	(0.024)	(0.010)	(0.011)	(0.009)	(0.010)
Mean of dep. variable	-0.015	-0.015	0.230	0.230	0.798	0.798
Observations	10101	10101	10101	10101	10101	10101
Adjusted R^2	0.272	0.325	0.200	0.250	0.169	0.193
Controls	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

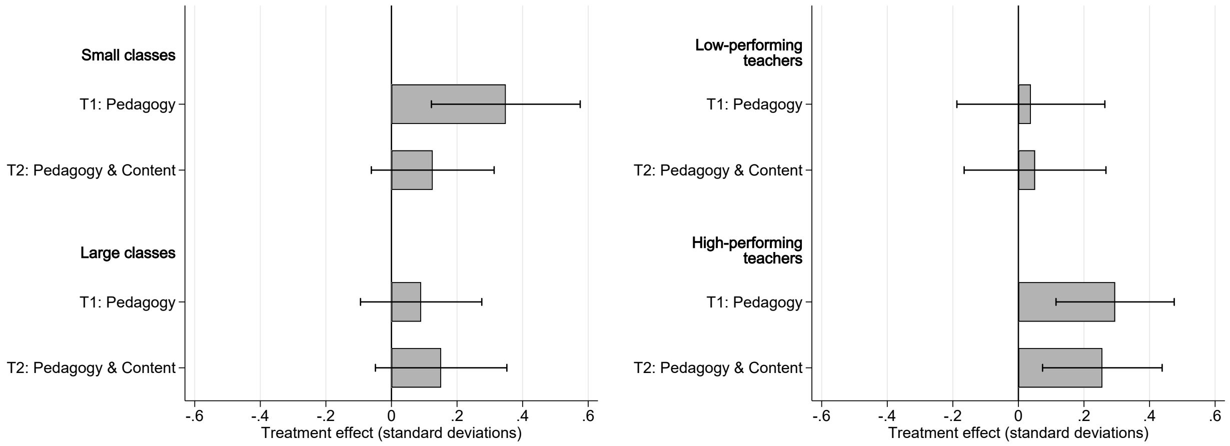
Notes: The *dependent variable* is pupils' standardized average score (across all subjects) for columns (1) and (2), a binary variable indicating whether a student's average score was A or B (3) and (4), and a binary variable indicating whether a pupil passed the exam for columns (5) and (6). *Pupil baseline math score* is a pupil's score in the SFNA exam administered in grade 4. Controls include *(i) pupil-level controls* for average SFNA baseline score across all subjects and sex, *(ii) school-level controls* for average PSLE baseline score (all subjects), class size, and number of pupils in grade 4 and *(iii) teacher-level controls* for sex, age, and math performance at baseline. Huber-White robust standard errors, clustered at the school level, in parentheses. + p<0.10, * p<0.05, ** p<0.01.

A.4 Effect heterogeneity and spillovers at the student level

Table A.6: Estimates for cascading effects on the math score of pupils

	Standardized		Scored A or B		Passed	
	(1)	(2)	(3)	(4)	(5)	(6)
T1: Pedagogy	0.066	0.087	0.013	0.016	0.017	0.027
	(0.059)	(0.055)	(0.011)	(0.011)	(0.025)	(0.024)
T2: Pedagogy & Content	-0.011	-0.012	0.009	0.009	0.006	0.006
	(0.056)	(0.055)	(0.011)	(0.012)	(0.024)	(0.023)
School PSLE avg. score (std)	0.077*	0.080*	0.016**	0.017**	0.033**	0.033**
	(0.032)	(0.031)	(0.006)	(0.006)	(0.013)	(0.012)
School SFNA avg. score (std)	0.132**	0.125**	0.022**	0.022**	0.047**	0.045**
	(0.032)	(0.032)	(0.006)	(0.006)	(0.013)	(0.014)
T2 − T1	-0.077	-0.099	-0.004	-0.007	-0.012	-0.021
	(0.064)	(0.066)	(0.013)	(0.013)	(0.027)	(0.028)
Mean of dep. variable	-0.000	-0.000	0.075	0.075	0.368	0.368
Observations	15023	15023	15023	15023	15023	15023
Adjusted R^2	0.073	0.081	0.035	0.040	0.053	0.060
Controls	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is pupils' standardized SFNA math score for columns (1) and (2), a binary variable indicating whether a pupil scored A or B (highest grades) in math for columns (3) and (4), and a binary variable indicating whether a pupil passed the math exam for columns (5) and (6). Controls include *pupil-level controls* for sex, *teacher-level controls* for sex, age, and math performance at baseline, and *school-level controls* for the number of pupils in grade 4 as well as each school's average SFNA and PSLE score in 2018. Huber-White robust standard errors, clustered at the school level, in parentheses. + p<0.10, * p<0.05, ** p<0.01.



(a) By class size (proxy)

(b) By teacher performance

Figure A.1: Heterogeneity in treatment effects on students' math scores by class size and teacher mathematical content knowledge at baseline.

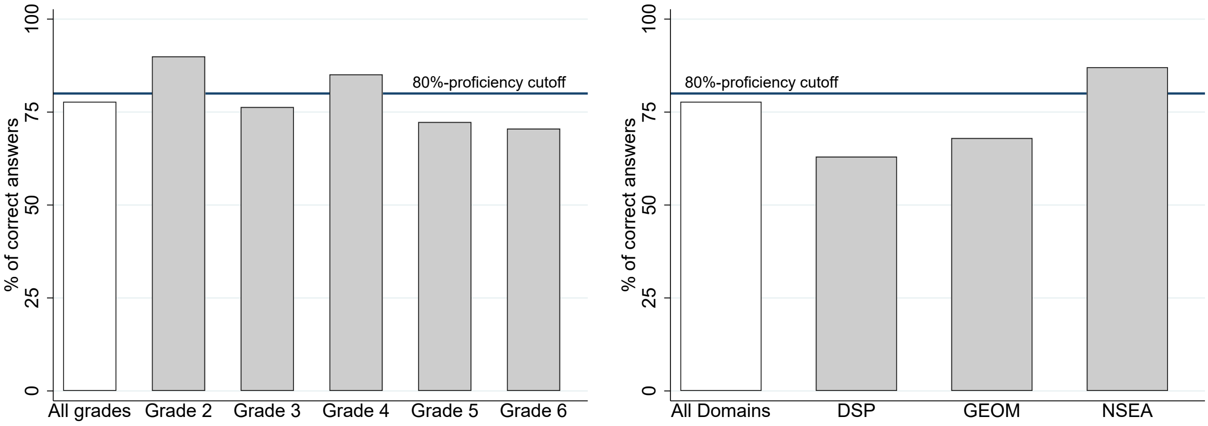
Groups are split at the median of class size and teacher performance. 90 percent confidence intervals shown.

Table A.7: Effect heterogeneity along attributes of pupils and teachers

Covariate:	Pupils' score		Female pupil		Teacher score		Class size	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.11^{+}	0.14^{*}	0.11^{+}	0.15^{*}	0.11^{+}	0.14^{*}	0.11^{+}	0.14^{*}
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Covariate	0.45^{**}	0.33^{**}	-0.03	-0.03	-0.06	-0.05	0.15	0.11
	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	(0.06)	(0.13)	(0.15)
Treatment × Covariate	0.02	-0.00	-0.05	-0.07	0.14^{+}	0.12	-0.27	-0.17
	(0.04)	(0.04)	(0.06)	(0.06)	(0.07)	(0.07)	(0.17)	(0.16)
Observations	10101	10101	10101	10101	10101	10101	10101	10101
Adjusted R^2	0.25	0.30	0.25	0.30	0.25	0.30	0.25	0.30
Teacher controls	No	Yes	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is pupils' standardized math scores in all models. Controls include *(i) pupil-level controls* for average SFNA baseline score across all subjects and sex, *(ii) school-level controls* for average PSLE baseline score (all subjects), class size, and number of pupils, and *(iii) teacher-level controls* for sex, age, and math performance at baseline. Huber-White robust standard errors, clustered at the school level, in parentheses. + p<0.10, * p<0.05, ** p<0.01.

A.5 Descriptive statistics on teacher content knowledge



(a) Average scores by grade level of question

(b) Average scores by domain

Figure A.2: Math proficiency of teachers prior to the project
The assessment featured 50 items covering the math curriculum of Tanzanian primary schools (grades 2–6) and was administered in November 2019. Participants are either *targeted teachers* (N=219) or *peer teachers* (N=215) nominated for the evaluation study by public primary schools in Siha, Karatu, Mbulu DC, and Mbulu TC. Note that the sample is neither representative for Tanzanian teachers nor for teachers in the study regions.

A.6 Additional results for program effects at the teacher level

We use the following equation to estimate intermediate effects on teachers:

$$Y_{isk} = \beta_1 T1_s + \beta_2 T2_s + \beta_3 Peer_i + \beta_4 T1_s \times Peer_i + \beta_5 T2_s \times Peer_i + X_i^{'}\gamma + \phi_k + \epsilon_{isk}, \qquad \text{(A.1)}$$

where Y_{isk} is a teacher's math score after the intervention, $T1_s$ indicates if the teacher's school was assigned to the PEDAGOGY intervention, $T2_s$ represents if a teacher's school was in the PEDAGOGY & CONTENT group, $Peer_i$ indicates if the teacher was only a peer teacher rather than being directly targeted, and $T1_s \times Peer_i$ and $T2_s \times Peer_i$ are interaction terms capturing if treatment effects for peer teachers are different from those on directly targeted teachers. Finally, $X_i^{'}\gamma$ is a vector of teacher-level controls for sex, age and baseline score, ϕ_k are strata fixed effects, and ϵ_{isk} captures the error term.

Table A.8: Main estimation results for program effects on the math score of teachers

Dependent	Overall				NSEA		GEOM + DSP	
variable:	%		Standardized		Standardized		Standardized	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
T1: Pedagogy	0.40	0.47	0.03	0.04	0.02	0.04	0.12	0.10
	(1.37)	(1.40)	(0.10)	(0.11)	(0.12)	(0.12)	(0.11)	(0.11)
T2: Pedagogy & Content	1.95	2.00	0.15	0.15	0.22^{+}	0.22^{+}	0.04	0.05
	(1.31)	(1.33)	(0.10)	(0.10)	(0.11)	(0.11)	(0.11)	(0.11)
Peer teacher	-2.92^{*}	-2.43^{+}	-0.22^{*}	-0.19^{+}	-0.19^{+}	-0.15	-0.09	-0.07
	(1.33)	(1.34)	(0.10)	(0.10)	(0.12)	(0.12)	(0.10)	(0.10)
T1 × Peer teacher	2.56	2.39	0.19	0.18	0.26	0.23	-0.06	-0.05
	(2.10)	(2.07)	(0.16)	(0.16)	(0.19)	(0.18)	(0.16)	(0.16)
T2 × Peer teacher	-0.59	-0.35	-0.05	-0.03	0.01	0.02	-0.15	-0.13
	(1.99)	(2.00)	(0.15)	(0.15)	(0.18)	(0.18)	(0.17)	(0.17)
Baseline score	10.01^{**}	9.54^{**}	0.76^{**}	0.73^{**}	0.68^{**}	0.64^{**}	0.74^{**}	0.72^{**}
	(0.56)	(0.59)	(0.04)	(0.04)	(0.06)	(0.06)	(0.03)	(0.04)
T2 − T1	1.55	1.53	0.12	0.12	0.20	0.18	-0.07	-0.05
	(1.48)	(1.51)	(0.11)	(0.11)	(0.13)	(0.13)	(0.11)	(0.11)
Observations	368	368	368	368	368	368	368	368
Adjusted R^2	0.62	0.63	0.62	0.63	0.48	0.49	0.58	0.59
Teacher controls	No	Yes	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is the share of correct answers for columns (1) and (2), standardized test scores for columns (3) and (4), standardized test scores on NSEA (numbers sense and elementary arithmetic) items for columns (5) and (6), and standardized test scores on GEOM (geometry and measurement) and DSP (data, statistics and probability) items for columns (7) and (8). Main treatment effects are reported for *targeted teachers*, i.e. teachers directly exposed to the treatments. *Teacher level controls* include sex, age, and years since graduation at baseline. Huber-White robust standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01.

Table A.9: Heterogeneity in program effects on teachers' mathematics performance

Covariate:	Baseline score		Age		Female	
	(1)	(2)	(3)	(4)	(5)	(6)
T1: Pedagogy	0.031	0.032	0.053	0.046	0.015	0.020
	(0.109)	(0.112)	(0.104)	(0.106)	(0.119)	(0.119)
T2: Pedagogy & Content	0.128	0.131	0.136	0.147	0.104	0.099
	(0.092)	(0.094)	(0.098)	(0.101)	(0.114)	(0.114)
Covariate	0.730**	0.712**	0.002	0.001	-0.131	-0.169
	(0.059)	(0.060)	(0.008)	(0.017)	(0.145)	(0.149)
T1 × Covariate	-0.115	-0.112	-0.013	-0.011	0.049	0.121
	(0.115)	(0.119)	(0.011)	(0.012)	(0.266)	(0.286)
T2 × Covariate	-0.341**	-0.349**	-0.016	-0.020^{+}	0.160	0.177
	(0.117)	(0.121)	(0.011)	(0.012)	(0.235)	(0.247)
Observations	368	368	368	368	368	368
Adjusted R^2	0.625	0.637	0.617	0.625	0.629	0.634
Teacher controls	No	Yes	No	Yes	No	Yes
Stratum fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The *dependent variable* is teachers' standardized test scores in all models. *Heterogeneity* is estimated along teachers' baseline score in columns (1) and (2), teachers' age in columns (3) and (4), and teachers' sex in columns (5) and (6). Main effects are reported for *targeted teachers*, i.e. teachers directly exposed to the treatments. Age is centered to have a mean of 0 for targeted teachers, and baseline scores are standardized to have a mean of 0 and a standard deviation of 1 for targeted teachers. Estimates for *Peer teacher*, *Peer teacher × Treatment*, *Peer teacher × Covariate* and *Peer teacher × Treatment × Covariate* not shown. *Teacher level controls* include sex, age, and years since graduation at baseline. Huber-White robust standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01.

B Appendix: Difference-in-differences analysis

To observe potential spillover effects over a long time horizon, we conduct a multi-year ex-post analysis based on school-level data for both grade 7 and grade 4 students. The Primary School Leaving Examination (PSLE) for seventh graders has been conducted on a yearly basis since 2013, while the Standard Four National Assessment (SFNA) assessment for fourth graders was launched in 2015. Combining the publicly available national examination data with the NGO documentation on the program implementation allows us to trace how tests scores in program schools evolve relative to test scores in schools that did not participate in the teacher training program. As only one teacher (or a very small group of teachers) per school was invited to participate in the program, and selected teachers were then instructed to organize knowledge sharing activities with their colleagues, this comes close to an estimation of cascading effects. To be precise, it provides an upper bound for these effects, given that a small share of students should have been taught by directly targeted teachers.

With these considerations in mind, we estimate cascading effects associated with the program using

$$Y_{st}^{Std} = \beta_1 \, Treatment_{st} + \lambda_s + \phi_t + \epsilon_{st} \quad \text{for } Grade \in \{4, 7\}, \tag{B.1}$$

where Y_{st}^{Std} represents the average test score in math of school s in year t for either grade 4 (SFNA) or grade 7 (PSLE), *Treatment* indicates whether one or several teachers from a school participated in the training on the new teaching methods and is set to 1 for a given year t and later years if school s was part of the program in year t (and to 0 otherwise), λ_s are school level fixed effects, ϕ_t are year fixed effects, and ϵ_{st} is the error term.

This corresponds to a standard two-way fixed effects estimator (TWFE). To assess the robustness of the difference-in-differences analysis, the standard TWFE-estimates are compared to results obtained from an alternative difference-in-differences estimator proposed by Callaway and Sant'Anna (2021). As a control group, we use both never and not yet treated units. In all models, the comparison group consists of all schools from the three Tanzanian regions – Arusha, Kilimanjaro, and Manyara – where the project was implemented.

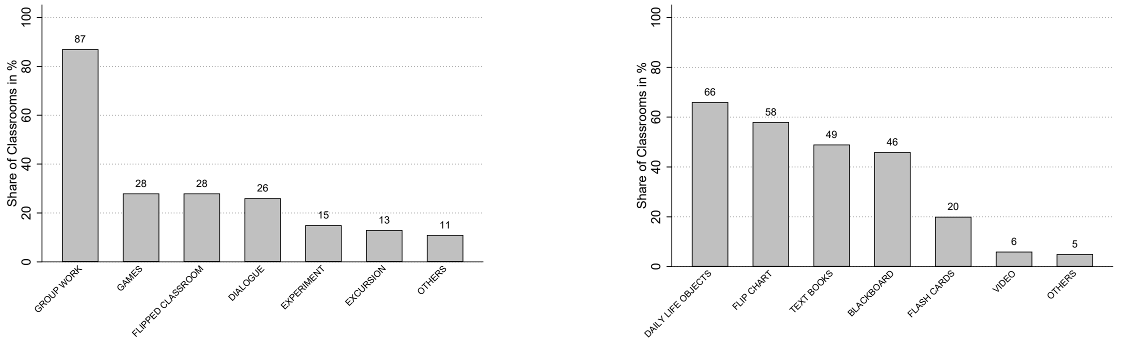
Results are presented in Table B.1. Across all models, effects are close to zero and insignificant.

Table B.1: School level difference-in-differences estimates for cascading effects, 2013–2019

	SFNA (grade 4)			PSLE (grade 7)		
	TWFE	CS		TWFE	CS	
		NE	NY		NE	NY
	(1)	(2)	(3)	(4)	(5)	(6)
ATT	0.032	-0.039	-0.033	-0.028	0.008	0.008
	(0.032)	(0.038)	(0.038)	(0.022)	(0.028)	(0.027)
Observations	11379	7168	7168	14954	11470	11470
Adjusted R^2	0.176			0.253		

Notes: The *dependent variable* are standardized test scores at the school level in all models. Effects in the standard TWFE model are compared with estimates obtained through the approach proposed by Callaway and Sant'Anna (2021), labeled as "CS". The presented CS coefficients stem from a comparison with *never treated* units (NE) or *not yet* treated units (NY). As the CS panel estimator does not take into account schools with incomplete data and always treated schools, they are based on a more restricted sample. To account for schools with incomplete data, results were also compared with a cross-sectional CS estimator and remain very similar (not shown). Standard errors in parentheses. + p<0.10, * p<0.05, ** p<0.01.

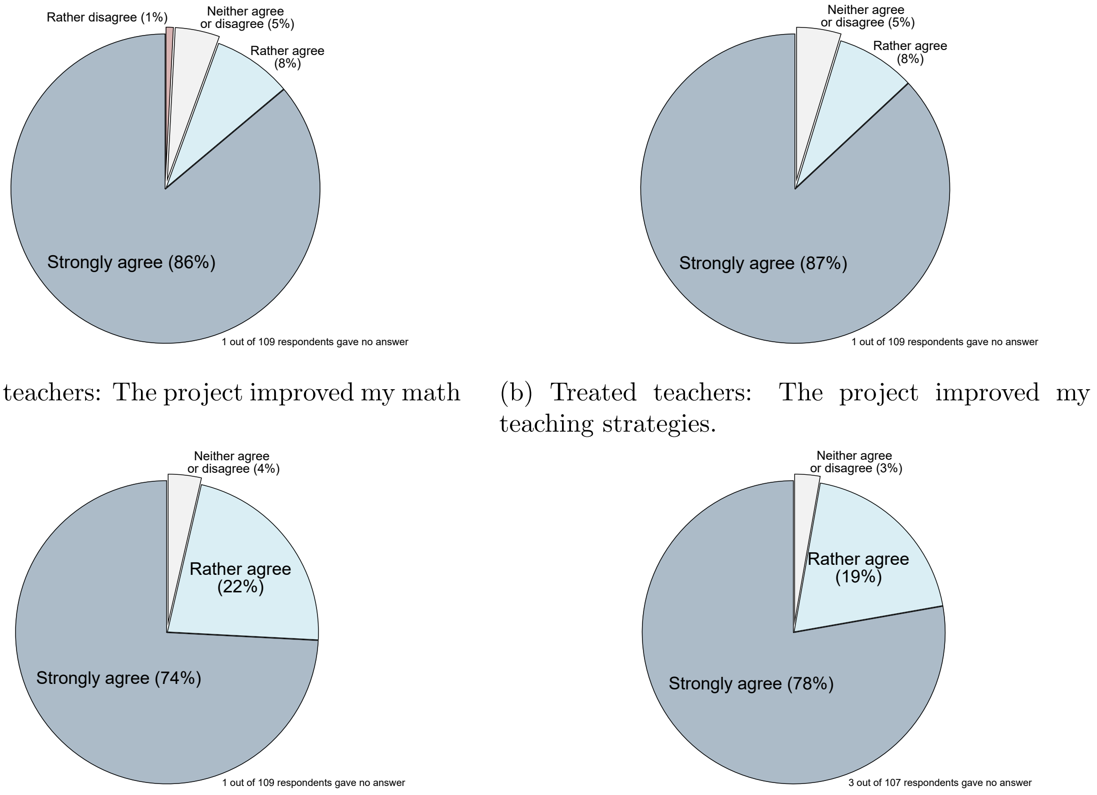
C Appendix: Classroom observations and opinion survey



(a) Involvement of pupils

(b) Use of teaching aids

Figure C.1: Observed teaching techniques in treatment schools.
The data was collected by government employed Quality Assurance Officers in 112 out of 130 program schools.



(a) Treated teachers: The project improved my math knowledge.

(b) Treated teachers: The project improved my teaching strategies.

(c) Treated teachers: The project improved the math skills of my pupils.

(d) Peer teachers: The project improved the math skills of my pupils.

Figure C.2: Perceived impact on teachers' *content knowledge in math*, their *teaching strategies* and their *students' math skills* as reported by the participants.
The treatment group includes 130 teachers, whereof 109 attended data collection, while the peer group includes 130 teachers, whereof 107 attended data collection.

D Appendix: Exemplary quotes from semi-structured interviews

Table D.1: Exemplary quotes from the semi-structured interviews conducted with Sitt participants, Sitt-d participants, peer teachers, and officials, part 1.

Group	General impression of SITT	Impact on math skills	Impact on teaching	Impact on pupils	Comparison with other educational programs
SITT	*"I really appreciate the SITT program, because it changed the way I deliver material to the classroom. [...] Thanks to SITT, I can use participatory methods that encourage pupils to contribute more actively."*	*"I understand mathematics very well. My main problem is how to teach it to the pupils. SITT showed me new ways in how to teach in the classroom. Concerning math skills, I gained some new ideas from the facilitators during the workshops."*	*"SITT helped me to involve kids in preparing teaching aids, and this helps the kids to remember the material better. [...] Another thing is that teachers are no longer working individually but together as a team. Pupils and teachers also came closer, you now find kids asking for the help of teachers."*	*"My knowledge increased and the way of teaching mathematics to my students improved so that my students learn better."*	Not discussed with Sitt participants.
SITT-D	*"SITT is really good. It helped me so much. Before SITT, I was afraid to teach math. After participating in this program, I feel comfortable teaching math."*	*"There is a change in my math proficiency, because I use the computer with the 'Kolibri' learning software."*	*"SITT changed me quite a lot. Now I engage children more actively in my lessons. Instead of narrating like a radio, I teach practically."*	*"The program probably helps the students. When I use SITT methods they like it and they learn better."*	Not discussed with Sitt-d participants.
Peers	*"SITT is useful to us, because it helps our pupils to prepare teaching aids [...] and it makes teaching more learner-centered. SITT will change our school, everybody loves it."*	Not discussed with peer teachers.	*"The SITT program has improved my teaching much, because it remembered me to use teaching aids and participatory methods."*	*"Pupils enjoy when we teach them according to SITT. That makes them understand more easily."*	Not discussed with peer teachers.
Officials	*"SITT is nice and very good for the teachers. Not only for the teaching aids and teaching materials but also for the technology. The teachers are learning through the computer and software."* *"I agree with my colleague. On WhatsApp, I observe what the teachers are sharing. It is really impressive and the teachers are enjoying it."*	Not discussed with officials.	*"During my school visits, I observed that SITT teachers have a different teaching approach. For instance, they try to use teaching aids and participatory methods."*	*"For now, it is difficult to say how large the effect of SITT is, because the pupils have been taught by several teachers between standard 1 and standard 7. So, I am not sure by how much SITT helps the performance of kids."*	*"I remember a program phasing out in 2012 that offered an in-service training. It was introduced and supported by UNICEF. [...] It was considered too burdensome by the teachers so they didn't work on it properly. [...] The program ended and the results were disappointing. For the case of SITT, the peer-sharing within school works better. Also the idea of model lessons helps. And SITT's unique participatory approach motivates pupils and makes them like mathematics more."*

Sources of quoted statements: Interviewees in Mbulu DC (×4), interviewees in Mbulu TC (×4), interviewees in Karatu (×3), interviewees in Siha (×4).
SITT refers to the group receving only the Pedagogy intervention, and SITT-D to the group that additionally recevied the laptops for content revisions, i.e. PEDAGOGY + CONTENT.

Table D.2: Exemplary quotes from the semi-structured interviews conducted with Sitt participants, Sitt-d participants, peer teachers, and officials, part 2.

Group	Feedback: Workshops	Feedback: Laptop/Kolibri	Feedback: Cascading	Relevance of evaluation	Additional remarks
SITT	*"I liked the training as it made me a better teacher. I also appreciated the change in environment from Mbulu to Arusha and the good service."*	Not discussed with Sitt participants.	*"The perspective of my colleagues was a problem. I called a meeting, and they agreed to my proposal. But once I asked them to join team teaching, most of them said 'Now, I have no time'. At other schools it is similar."*	Not discussed with Sitt participants.	**About Covid-19 and the future:** *"We temporarily closed schools due to Covid in 2020. Still, we used SITT to improve our teaching and that is why we achieve a good performance in our school. I ensure that we will keep it and improve even more."*
SITT-D	*"I liked the workshop very much, but I was disappointed that the additional meetings for SITT-D in 2019 were canceled [because of Covid-19]."*	*"The laptop and learning software are very useful. Kolibri helps mathematics teachers to be up to date. We use it to refresh our knowledge before teaching a certain topic. It makes us comfortable."*	*"We created a timetable to plan the model lessons and team teaching. Now, I see my colleagues using teaching aids. They like it and cooperate."*	Not discussed with Sitt-d participants.	**Training intensity:** *"It would be good to have more than a 5-day workshop to have additional time to learn and share with teachers from other districts."*
Peers	Not discussed with peer teachers.	Not discussed with peer teachers.	*"Once our colleague shared their SITT-knowledge, we agreed together to have team teaching. [...] Around ninety percent appreciate it. [...] We will continue to use the techniques that the SITT project introduced."*	Not discussed with peer teachers.	**The cascading approach:** *"We assessed each other on how we conduct model lessons and discussed it during meetings. But there are some challenges: Not all teachers were eager to participate in the knowledge sharing activities."*
Officials	Not discussed with officials.	Not discussed with officials.	Not discussed with officials.	*"It is important to conduct an evaluation so that the implementers get feedback on what they are doing and to see whether it is useful or not. Spending money on an evaluation is necessary."*	**The relevance of evaluations:** *"It is very important to do the evaluation and to understand whether the program delivers or not."*

Sources of quoted statements: Interviewee in Mbulu DC (×1), interviewee in Mbulu TC (×2), interviewee in Karatu (×4), interviewee in Siha (×4).
SITT refers to the group receving only the Pedagogy intervention, and SITT-D to the group that additionally recevied the laptops for content revisions, i.e. PEDAGOGY + CONTENT.

Chapter 4

Mobilizing for the Public Good: A Field Experiment on Community-Driven Development and Waste Management

Mobilizing for the Public Good: A Field Experiment on Community-Driven Development and Waste Management

Martina Jakob*

University of Bern
martina.jakob@unibe.ch

Carla Coccia*

University of Bern
carla.coccia@unibe.ch

October 31, 2023

Community-driven development has become a popular bottom-up alternative to the traditional top-down provision of local public goods. This is the first study to experimentally compare the effectiveness of these two approaches. Based on a randomized controlled trial with 120 communities in rural El Salvador, we assess the impact of two interventions addressing solid waste pollution: (i) a traditional top-down intervention where streets were cleaned by an external actor, and (ii) a community-driven intervention where a facilitator raised awareness and mobilized for collective action. We derive an objective measure of pollution using geotagged photos and deep learning. We find large immediate effects for both interventions, with reductions in waste pollution by 0.7–0.8σ for the traditional intervention and 0.5–0.6σ for the community-driven intervention. Four months after the end of the project, these effects depreciated by 80 percent for the top-down and 60 percent for the bottom-up treatment. Our complementary data from 2,421 surveys and 883 activity records is consistent with a theoretical model where many individuals are willing to contribute to public goods as long as others do so too, but fail to coordinate in the absence of a committed leader.

[*] Both authors contributed equally to this study. This project was financed through a *doc.CH* grant by the Swiss National Science Foundation (SNSF) awarded to Martina Jakob for the completion of her dissertation projects. A randomized controlled trial registry is available at: `https://www.socialscienceregistry.org/trials/10913`.

1 Introduction

Many of the world's most pressing challenges, like curbing emissions, maintaining global peace, or establishing a functioning health and education infrastructure in low-income countries, are public goods problems. As public goods benefit everyone irrespective of their personal contribution to them, individuals have an incentive to free-ride. To avoid the resulting underprovision, the standard solution calls for a *top-down* intervention by a powerful actor such as the state to provide the public good or enforce rules for its protection (Olson, 1971). Yet, ample empirical evidence documents that groups are often able to act collectively and overcome the social dilemma tied to public goods (e.g., Ostrom, 1990, 1999). This has inspired an alternative line of thinking advocating for *bottom-up* solutions through so-called community-driven development (CDD). In this study, we compare the effectiveness of these two approaches in the context of solid waste management.

Our paper is based on a randomized controlled trial with 120 communities in rural El Salvador. We study the impact of two programs designed to reduce local solid waste pollution. The first intervention pursued a traditional top-down approach with monthly community visits by an external cleaning team to collect litter from the streets. In the second intervention, a local facilitator was appointed for each community to raise awareness and mobilize for collective action to address the problem in a bottom-up process. Typical activities in this community-driven initiative were educational sessions about waste management, collective monthly cleanups, and community meetings to define common strategies. The two interventions had a duration of four months, were similar in cost, and implemented by the local NGO Consciente. We randomly assigned communities to three experimental groups: 40 communities received the traditional top-down intervention, 39 participated in the community-driven bottom-up initiative, and 41 were assigned to a control group. To track contamination levels in all communities, we took about 200,000 geo-tagged photos along all streets, and evaluated them using a deep learning model. Our model achieves state-of-the-art performance in trash detection, allowing us to establish a reliable and objective measure of contamination. To understand the mechanisms behind potential impacts, these contamination assessments were complemented with survey data from a sample of 2,421 villagers and a detailed registry of all the 883 activities conducted in the context of the interventions.

We find large *immediate impacts* for both interventions. The traditional intervention reduced solid waste pollution by 0.7–0.8σ or 36 percent ($p < 0.01$). Effects are significantly smaller ($p < 0.05$), but still substantial for the community-driven intervention, with a reduction by 0.5–0.6σ or 29 percent ($p < 0.01$). Our survey results further show that these improvements did not go unnoticed, as both interventions had significant immediate effects on people's cleanliness perceptions ($\sim$0.15σ for both interventions) and self-reported recycling practices ($\sim$10 percentage points for both interventions). For the community-driven intervention, we also observe a 13 percentage point increase in the share of respondents indicating that they dispose of their waste appropriately, rather than burning, burying or dumping it. *Long-term results* show that four months after the end of the intervention, the impact on observed pollution decreased by 80 percent for the traditional intervention and by 60 percent for the community-driven intervention. This yields a long-term effect of 0.1σ ($p = 0.11$) for the top-down treatment and of 0.2σ ($p < 0.05$) for the bottom-up treatment. While this is suggestive evidence for a higher persistence in the community-driven intervention, the difference between depletion rates is not statistically significant ($p = 0.2$). We observe no depletion in people's cleanliness perceptions, but the immediate changes in self-reported waste management behavior strongly depreciate or disappear for both treatments.

Our rich complementary data offers insights into the mechanisms driving the success and limitations of community-driven development. We find limited evidence for *information effects* through increased awareness of the problem or knowledge of others' concern for it. Although the CDD initiative had an immediate impact on people's beliefs about the prevalence of littering behavior in their community, this social norms effect was short-lived and not significantly more pronounced than in the traditional intervention. Our results are more consistent with the hypothesis that CDD can alleviate *organizational constraints* to collective action. However, much of the success along this dimension appears to be tied to the presence of the facilitator. While the number of cleanup events and participants remained consistently high during the intervention period, we observe a sharp decline in collective efforts – from 0.9 to 0.4 monthly cleanups – after the withdrawal of the NGO, and we find limited evidence for a sustained increase in social capital. Our results are most consistent with a theoretical model where many individuals are willing to contribute to public goods as long as others do so too, but struggle to coordinate in the absence of a dedicated leader.

This study makes three distinct contributions. First, we add to the debate on the effectiveness of community-driven development. The rise of CDD initiatives represents a major trend in international development cooperation (Mansuri and Rao, 2012;

Casey, 2018). Based on an analysis of 250,000 World Bank project reports, we find that the share of documents mentioning keywords connected with community-driven development increased rapidly from the early 1990s. By 2003, over 40 percent of all documents contained at least one related term. Our pre-survey further shows that practitioners and academics alike tend to be optimistic about community-driven initiatives, with roughly 80 percent of respondents in both groups believing they would outperform traditional top-down solutions in the long run. Despite the vast importance of the approach, rigorous evaluations of CDD initiatives remain scarce (Table A13). Most notably, the effectiveness of community-driven solutions has not yet been compared to that of the more traditional alternatives they seek to replace. This study contributes to filling this critical gap in empirical research. Our findings highlight that while CDD initiatives can indeed successfully promote the provision of local public goods, they are not always more effective in doing so than top-down interventions.

Second, our study also contributes to the discussion on how to tackle problems related to solid waste management in developing countries. While 96 percent of waste in high-income countries is collected and properly disposed of, only 39 percent of waste in low-income countries is. At the same time, solid waste generation in low- and middle-income countries is expected to triple by 2050 (Kaza et al., 2018). Finding effective ways to address the problem and limit the environmental and health repercussions it causes, is thus a critical and timely priority. Our study ties into the nascent literature evaluating different interventions to improve solid waste management (Table A14). We find that raising awareness and empowering communities to address the waste problem can be an important part of the solution, but may not be successful on its own without continued investment. In addition, our results suggest that interventions that focus on changing littering norms alone, without complementary efforts to collect waste that continues to accumulate on the streets, are unlikely to be sustainable.

Finally, our paper advances the burgeoning field of research using machine learning methods to track and understand global development. A rapidly expanding economic literature has shown that important socio-economic outcomes can be accurately predicted from alternative data sources such as satellite imagery (Jean et al., 2016; Yeh et al., 2020), phone records (Blumenstock et al., 2015), or tweets (Jakob and Heinrich, 2023). However, this literature is largely focused on providing proofs of concept, and scientific or practical applications remain scarce. In this study, we use deep learning to derive an objective and reliable measure for our main experimental outcome. By fine-tuning a YOLOv8 object detection model using publicly available trash data and a sample of images from our experiment, we achieve state-of-the-art performance in trash detection, with an AP50 of 59.5 percent on the popular TACO dataset and of 59.0 percent for our own images. The resulting contamination measure produced more robust results than an alternative approach based on subjective contamination assessments by enumerators. This highlights the potential of deep learning methods in settings where large amounts of data must be processed or human measurements are prone to subjectivity.

2 The Public Goods Problem and the Rise of Community Driven Development

The public goods problem models a situation where the benefits of a cooperative outcome accrue to everyone irrespective of people's individual contributions towards it. The dominant strategy for a self-motivated and rational agent is to free-ride by contributing nothing. Standard economic theory considers this a market failure, as it results in a single, Pareto-inefficient equilibrium where the public good is not provided, and individuals fail to realize a mutually beneficial outcome (Olson, 1971; Hardin, 1971, 1982). The conventional approach to addressing market failures associated with public goods calls for a top-down intervention by a powerful entity, such as the state, to either supply the public good or enforce protective regulations.

However, ample research documents that most people do not behave as the standard model of self-interested actors predicts. Zero contributions to public goods are neither the norm in laboratory experiments (e.g., Fischbacher et al., 2001; Willer, 2009; Chaudhuri, 2011) nor in real-world situations. For example, many people volunteer in associations, donate blood, contribute to charities, make environmentally friendly consumption choices, or take part in political protest. Rather than maximizing personal gains, the majority of individuals appear to follow norms of reciprocity and contribute as long as a sufficient number of others do so too (e.g., Keser and Van Winden, 2000; Gächter, 2006; Thöni and Volk, 2018). Under preferences for conditional cooperation, the provision of public goods becomes a coordination problem with multiple possible equilibria. This is consistent with numerous examples showing that groups sometimes succeed and sometimes fail in providing public goods or protecting common resources (e.g., Ostrom, 1990, 1999).

In this context, the idea has gained traction that groups can be empowered to coordinate and guarantee the provision of public goods in a bottom-up process. This approach is variously known as community-driven development (CDD), community-based development (CBD), community and local development (CLD), or participatory development (Mansuri and Rao, 2012; Casey, 2018).

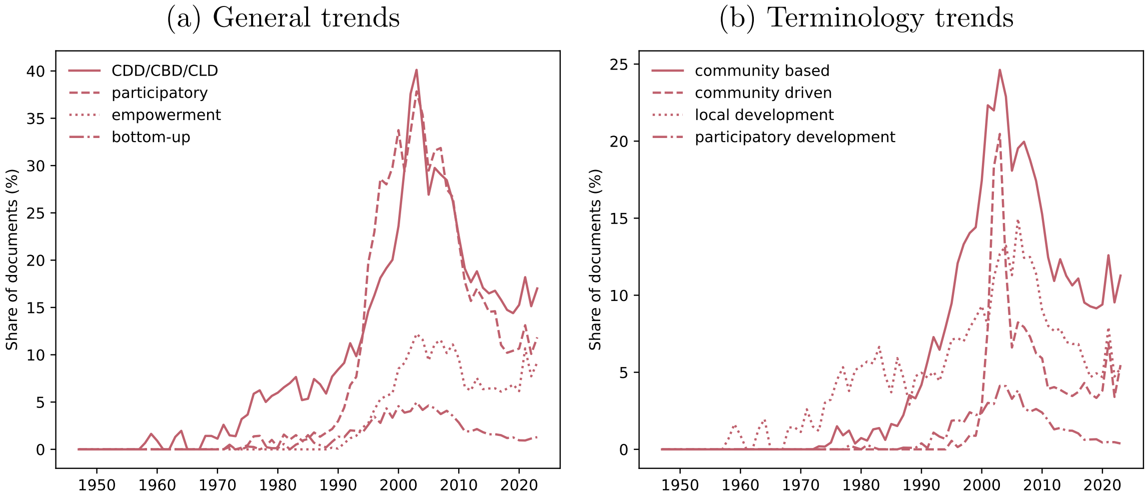


Figure 1: The Rise of Community-Driven Development

Illustration based on 259,668 project documents obtained through the World Bank API. Document types include, among others, procurement plans (23%), implementation reports (18%), project information documents (5%), or environmental assessments (5%). We exclude documents with less than 500 correct English words (10% of all documents), and documents that do not contain the word "development" (15% of the remaining documents). "CDD/CBD/CLD" refers to any of the keywords "community-driven", "community-based", "participatory development", or "local development" (different spellings accounted for).

The rise of CDD initiatives represents a major strategic shift in international development cooperation. In response to concerns about poorly maintained infrastructure following traditional top-down interventions, governments, NGOs and international organizations have increasingly turned to community-based solutions for public goods provision. This bottom-up approach is often hailed as "more responsive to demands, more inclusive, more sustainable, and more cost-effective than traditional centrally led programs"(Dongier et al., 2003), and believed to sustainably transform and strengthen local institutions. To illustrate this trend, we scraped over 250,000 World Bank project documents, published between 1947 and 2023. We find that the proportion of documents containing keywords directly related to community-driven development, along with more loosely connected keywords such as "participatory", "empowerment", or "bottom-up", began to increase rapidly in the early 1990s (Figure 1). At its peak in 2003, more than 40 percent of the documents mentioned at least one CDD keyword. Over the past two decades, this share has declined, but remains high, stabilizing at around 17 percent for the past three years. This strong focus on participatory, bottom-up initiatives is also reflected in funding priorities. In 2022, the World Bank alone had 373 ongoing community-based initiatives with more than $40 billion in total lending (World Bank, 2022).

Our pre-survey with 100 scientists and local practitioners further substantiates this sense of optimism regarding the potential of the CDD approach (Figure A1). Over 90 percent of the practitioners and scientists in our sample expressed confidence that adopting a community-based approach to waste management would lead to a reduction in community pollution in both the short and long term. In addition, about 80 percent of respondents from both groups agreed that a community-based approach would outperform a more traditional intervention in the long run. While the majority of academics believed that the relative advantage of CDD unfolds only in the long term, most practitioners also predicted better short-term outcomes.

The rise of community-driven development has also sparked interest in the academic community, leading to a number of rigorous evaluations to assess the effectiveness of the approach. Table A13 provides a comprehensive overview of this literature. Although the reviewed studies vary in the types of interventions and outcomes they examine, we can draw four general conclusions from this research. First, CDD initiatives are indeed often successful in delivering and maintaining public goods and improving the livelihoods of the poor (Avdeenko and Gilligan, 2015; Björkman and Svensson, 2009; Desai and Olofsgård, 2019; Duflo et al., 2015). Second, the evidence is inconclusive on the proposed transformative impact on local institutions. Many evaluations report no lasting effects on collective action capacity (Casey et al., 2012; Casey, 2018; Mansuri and Rao, 2012) or the empowerment of minority groups (Casey et al., 2012; Van der Windt and Mvukiyehe, 2020). Third, existing studies compare CDD initiatives with a status quo where no infusion of funds occurs. While this allows to assess whether such initiatives work, it does not tell us if they outperform alternative ways of service delivery, a key limitation noted by several recent studies in the field (e.g., Casey, 2018). Fourth, there appears to be little clarity about the precise mechanisms through which CDD interventions should affect the provision of public goods, limiting our understanding of where such initiatives may fail and how they can be improved. Our study addresses the limitations raised in the last two points by (i) offering a comparison between two modes of providing the same public good, and (ii) discussing the results within a more general theoretical framework.

3 A Theoretical Model of Collective Action

We propose a simple theoretical model to explain through what channels CDD potentially facilitates collective action and the provision of public goods.[1] We assume that individuals are willing to contribute to a public good as long as a certain fraction of the group does, and that they differ in these *thresholds for conditional cooperation.* A threshold of 0 corresponds to people who always cooperate, while a threshold of 1 indicates that someone never cooperates even if everyone else in the group does. Evidence from laboratory studies shows that these extreme types are in the minority, and that most people exhibit behavior consistent with varying degrees of conditional cooperation (Fischbacher et al., 2001). Individual thresholds may be determined by numerous factors, such as the importance the person places on the public good (i.e., preferences), the individual's pro-sociality, or his or her resources. As people usually cannot observe the actual number of contributors, they act based on their beliefs about it. This means that individuals will start contributing as soon as they believe that the proportion of contributors is higher than their personal threshold, and stop doing so if they think that this is no longer the case. For a given distribution of thresholds, multiple equilibria may thus be possible.[2] In a repeated game, we would expect self-reinforcing positive or negative dynamics, as people continually adjust their contributions based on the observed contributions of others until a stable equilibrium is reached (Berger, 2021; Berger et al., 2023).

Even when a socially more desirable equilibrium exists, attaining it often requires coordinated action. Take the example of a group of workers deciding whether to go on strike. If most people are willing to participate as long as most others do so too, the strike can only take place if the group coordinates to act simultaneously. Ample research shows that allowing people to communicate with each other increases the chance of reaching a stable high-level equilibrium (Chaudhuri, 2011). Following Cowen (1992) and Dahlman (1979), we thus assume that coordinating collective action entails *transaction costs.* The magnitude of these costs depends on how well people know and trust each other, and on the institutions they set up to facilitate cooperation. This idea is reflected in the notion of social capital, commonly understood as "the norms

[1] For simplicity, we limit ourselves to the extensive contribution margin (i.e., whether people contribute). Yet, a very similar case can be made for the intensive contribution margin (i.e., how much people contribute).

[2] Consider a community where 40 percent of individuals will contribute as long as at least 30 percent of the population contributes, and 60 percent contribute as long as at least 80 percent contribute. In this case, three stable equilibria could be reached: one where no one contributes, one where 40 percent contribute, and one where everyone contributes.

and networks that facilitate collective action" (Woolcock et al., 2001, p. 9). Thus, at higher levels of social capital, members of a group are more likely to succeed in organizing to collectively provide or protect public goods (e.g., Anderson et al., 2004). A related idea concerns the concept of leadership. In most real-world scenarios, transaction costs are not perfectly divisible, meaning that a single individual (or a small group of individuals) must bear a large portion of these costs. The presence of a committed leader (or leadership team) should thus be critical for a group to overcome organizational constraints to collective action. This is in line with extensive empirical evidence documenting the importance of leadership for collective action and the provision of public goods (e.g., Glowacki and von Rueden, 2015; Sahin et al., 2015).[3]

Finally, the provision of certain public goods requires a significant monetary investment. In a low-income setting, where time is not easily translated into money and people lack access to affordable loans, a group may fail to realize a collectively beneficial outcome due to *credit constraints*. For example, consider a poor community trying to build a paved road that is expected to yield high returns for everyone. Even if individuals are willing to contribute and able to organize themselves, the project will not be realized if the community does not have access to funding. This aligns with numerous studies documenting how financial markets often fail the poor (Banerjee and Duflo, 2007). Therefore, we conclude that collective action succeeds if the distribution of contribution thresholds allows for a high-level equilibrium, if the community is sufficiently organized to coordinate collective action so that this equilibrium can be reached, and if its members have access to sufficient funding to cover potential monetary investments.

Based on this framework, we distinguish three basic mechanisms through which community-driven development interventions could facilitate collective action. First, it can help to alleviate *informational constraints*. If individuals underestimate the share of others contributing to a public good, getting people to talk about the problem can eliminate these misconceptions. As interventions often convey information on the topics related to specific public goods and on effective solutions, they may also directly alter the distribution of thresholds, as people begin to care more about the problem

[3] Note that in some cases, the institutions groups set up to facilitate cooperation can also be seen as modifying the thresholds themselves (rather than lowering transaction costs). For example, if groups devise means of punishing defector, this would change people's thresholds for conditional cooperation. The same can be said if individuals get inspired by a charismatic leader (Jack and Recalde, 2015). For simplicity, we abstract from this alternative conceptualization, and view the organization of the collective action as a second-order public goods problem related to the bearing of transaction costs.

or become more confident about their ability to address it. Under certain threshold distributions, this would enable the community to reach a higher equilibrium. A second possible mechanism is related to *organizational constraints* and thus to the transaction costs that effective coordination entails. By bringing people together and encouraging them to set up organizational structures, CDD could build social capital and leadership, and thereby facilitate coordination. A final potential channel through which CDD could improve public goods is by mitigating *credit constraints*. People in poor communities may be sufficiently informed and organized to address local public goods problems, but simply lack access to funding to do so. This is the premise of the archetypal CDD intervention, which provides block grants to communities to invest in the provision of local public goods.

While the main focus of our study is on comparing the effectiveness of CDD with a more traditional top-down intervention, we will use this theoretical model to make sense of patterns in our data, thereby contributing to a better understanding of why CDD may work and where it may fail.

4 Context and Interventions

We conducted our study in the context of solid waste management in rural communities in El Salvador, a lower-middle-income country in Central America. Inadequate waste management is ubiquitous in developing countries and causes numerous detrimental health and environmental impacts. Over time, waste can spread over large areas, contaminating rivers, oceans, groundwater and soil. In addition to its environmental consequences, contaminated water can pose serious health risks by spreading infectious diseases such as diarrhea or hepatitis (Mohan and Joseph, 2021). While 96 percent of solid waste in high-income countries is collected and properly disposed of, the corresponding figures are only 51 percent for lower-middle-income countries and 39 percent for low-income countries (Kaza et al., 2018). At the same time, solid waste generation in low- and middle-income countries is expected to triple by 2050. Finding out what works to address the problem is thus an important and timely priority (see Table A14 for a review of the emerging literature in this area).

Solid waste contamination poses a typical public goods problem for local communities, as a clean environment benefits everyone regardless of their personal contribution towards it. This is in line with insights from our baseline survey, showing that even though most people in our sample are bothered by the waste pollution in their communities, the problem remains widespread. Nearly 80 percent of respondents indicate that the waste in their community bothers them much (39%) or very much (38%). Meanwhile, about half the people admit that they dispose of their waste improperly by burning, burying, or dumping it. Similarly, only 32 percent of the communities are visited by a municipal garbage truck collecting household waste at least every two weeks, and 36 percent have no such truck service at all. Almost all communities have considerable amounts of waste in public spaces, and contamination levels are not significantly correlated with the frequency of the garbage truck service.

Communities can adopt two distinct strategies to tackle the problem. People can either stop dumping waste in public spaces, or they can coordinate to collectively remove it and ensure proper disposal. Since both actions involve costs, and gains are shared between all residents, effective solutions are needed to overcome the free-riding problem inherent to the provision of public goods. We partnered with the local NGO Consciente to develop and implement two interventions to address the problem: (i) a traditional top-down intervention and (ii) a community-driven bottom-up intervention. Both initiatives had a duration of four months and were similar in costs.

In the *traditional intervention*, an external team of cleaners employed by the NGO made monthly visits to all communities to collect litter from public areas and gather household waste from residents. The team comprised two cleaners and a garbage truck driver, and each community visit typically lasted half a day. While the intervention was conducted by a non-governmental rather than a governmental institution, it mirrors what a top-down state intervention would look like in this context and corresponds to the typical approach pursued by governments worldwide to tackle solid waste pollution.

In the *community-driven intervention*, a team of 24 part-time facilitators was hired and trained in topics related to waste pollution and management and community organization strategies. Facilitators were typically young university graduates from the area, but not necessarily from the community, and responsible for one or two communities. Their job consisted in raising awareness for the problem, mobilizing for collective action, and encouraging the creation of local organizational structures to facilitate sustainable solutions. For this purpose, they could draw on extensive teaching materials developed by the NGO, but were instructed to adapt the proposed activities based on local needs. The typical community intervention consisted of an initial meeting, a series of educational sessions and community activities on waste management (such as input and discussion sessions, hands-on workshops, poster campaigns, or community movie nights), and monthly collective cleanups. The facilitator also assisted the community in organizing the disposal of the waste collected from the cleanups and households. This was typically done by using the private vehicle of a community member or by appealing to the municipality government. At the end of the intervention, each community presented a waste management plan indicating how the problem would be addressed after the withdrawal of the NGO.

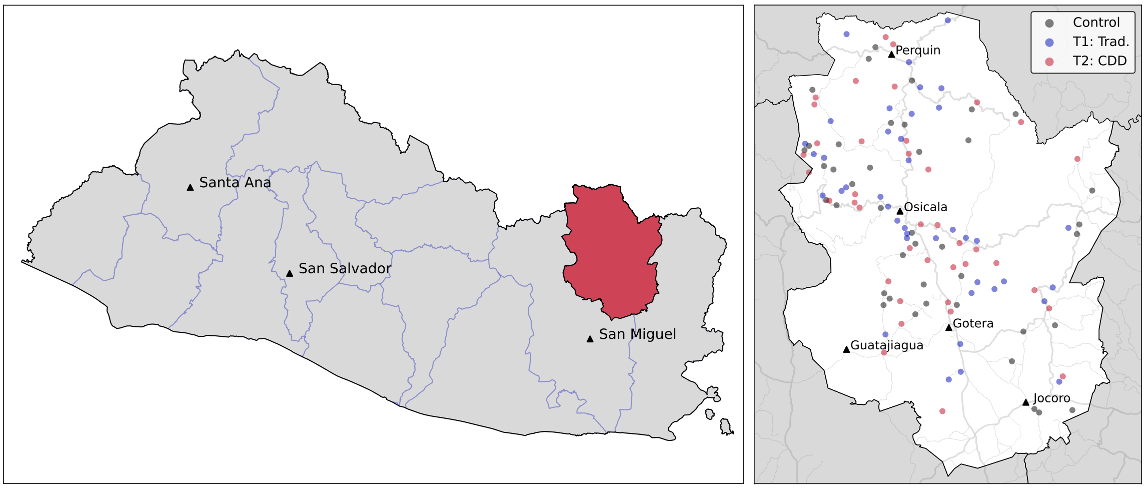


Figure 2: Study Area

5 Research Design

To study the impact of these two interventions, we conducted a randomized controlled trial with 120 communities in the rural department of Morazán in El Salvador (see Figure 2). The selection of these communities was undertaken in two steps. First, we compiled a list of medium-sized, non-urban communities (30–300 households) facing waste management problems with the help of municipal governments. In a second step, we conducted a baseline survey in 140 communities and selected 120 communities based on two criteria: contamination levels using our own measurements, and spatial distance to limit spillover effects. The resulting sample is not representative of our study area, but contains a diverse set of communities that have not solved the waste management problem through an endogenous bottom-up process or with the help of local government institutions. We randomly assigned these communities to three experimental conditions: (1) the traditional intervention (40 communities), (2) the community-driven intervention (39 communities), or (3) a control group that received no intervention (41 communities).[4] Randomization was stratified by base-

[4] The number of communities differs between experimental groups because remainders per stratum were assigned with probability 1/3 to each group or group combination (in the case of two

line contamination (three bins) and geographic zones (four bins). In all experimental groups, we conducted a measurement wave before the intervention (baseline), toward the end of the intervention (midline, after 3–4 intervention months), and four months after the intervention (endline). This allows us to study both the immediate impact of the two interventions and whether potential effects are sustained after the end of the program.

5.1 Data

To track different waste-related outcomes, we collected three types of data: (i) contamination assessments based on images taken along all streets, (ii) survey data on people's perceptions and self-reported behavior, and (iii) monitoring data on all the activities that were conducted in the context of the interventions.

5.1.1 Image Data on Contamination

For the main outcome of our experiment, we took *geocoded pictures* along all the streets and public spaces in the 120 communities. For this purpose, enumerators worked in pairs and simultaneously took geotagged photos on both sides of the street every five steps. Enumerators were carefully trained and received a detailed manual explaining how to take the photos. Photos typically show a portion of the street, the roadside, and the background. To ensure spatial consistency across the three measurement waves, we used an application that enabled us to outline the geographic boundaries of each community and display them on an interactive map. Enumerators were instructed to cover all roads, paths and public spaces within this designated area. To account for minor deviations in the covered area, we only include photos with spatial support across all three waves (92% percent of all images).[5] This procedure results in approximately 500 images per community and wave, ranging from 118 photos in the smallest community to 1,926 photos in the largest community, and a total of 181,393 images across all waves. We then used a deep learning model to predict the amount of trash on each image (see Section 5.2).

remainders). A common alternative is to group remainders over all strata and reassign them randomly. Assigning remainders with probabilities instead of grouping them means equal group sizes cannot be ensured, but assignment balance within the strata is preserved (McKenzie and Bruhn, 2011).

[5] A photo in a given wave is defined as having no spatial support in another wave if the closest photo is more than 8 meters away and the fifth closest photo is more than 25 meters away. This decision rule was found to produce good results, by excluding road segments that were not covered in all waves, but keeping photos in all other segments. We only include photos with spatial support in all waves, e.g., only baseline photos with nearby midline and endline photos.

A key challenge is to link midline and endline contamination levels in different areas of each community to their baseline contamination values. We use three different approaches of *spatial aggregation*: (i) a kernel approach, (ii) a raster approach, and (iii) raw community averages. The *kernel approach* consists in drawing a circle with a radius of 12.5 meters around each midline or endline image (see Figure 3).[6] We then use a triangular kernel to compute a weighted average of all baseline contamination values within the circle. The average circle contains 7.6 baseline images, and 99 percent of all circles contain at least one baseline image. Our final sample consists of 60,709 observations for the estimation of immediate effects (midline) and 65,673 observations for assessing long-term effects (endline). For the *raster approach*, we lay a fixed 16.5 x 16.5 meter grid over each community and compute the wave-specific average across all photos in each cell (see Figure 4).[7] The average cell contains 4.6 images, and 81 percent of all cells with baseline images also contain midline and endline images. The raster approach results in a final sample of 10,740 cells, with an average of 90 cells per community, ranging from 21 cells in the smallest community to 278 cells in the largest community. Finally, we also compute *raw averages* across all images for each community and wave, resulting in 120 (unclustered) observations.

As a robustness check, enumerators were also told to make a *subjective assessment* of the general cleanliness of the environment every 25 steps or 5 photos. Based on representative example images, they had to classify their environment into four categories, ranging from “very clean” to “very dirty”. Our final sample consists of about 100 ratings per community and wave, with 23 assessments in the smallest community and 408 in the largest community.[8] We use a triangular kernel with a radius of 25 meters and a raster of 33 x 33 meters for spatial aggregation of the enumerator assessment data (see Figures A4 and A5).

[6] Note that this approach results in different baseline circles for midline and endline measurements respectively. To determine an appropriate radius, we created and examined community maps showing all included observations (with baseline values) and excluded observations (without baseline values) for different circle sizes. With a radius of 12.5m, almost all dropped observations were at the community boundaries (which were interpreted slightly differently across waves) rather than within communities.

[7] The ideal raster produces enough observations (cells) per community while maintaining a good support across waves, so that few of these observations need to be dropped. A 16.5 x 16.5 meter grid was found to strike a good balance between these competing criteria.

[8] To obtain geocoded ratings, we used a simple low-tech strategy. Enumerators had to take a picture of a placard with the number corresponding to the level of contamination. We then used the weights of a Github model pretrained on the popular Street View House Numbers (SVHN) dataset (Netzer et al., 2011) to predict the number corresponding to each image. To make sure that all predictions were correct, we manually reviewed the few cases where the model predicted low certainties. The number images were integrated with all other photos to determine the spatial support across waves.

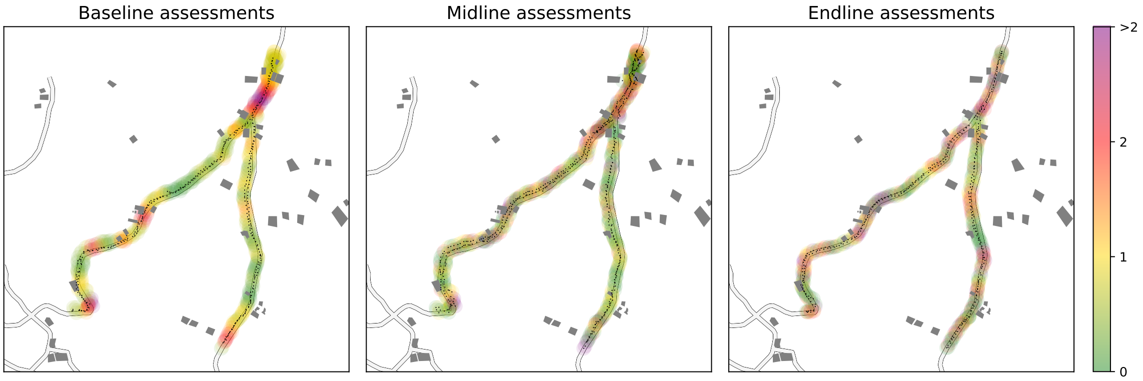


Figure 3: Illustration of Kernel Approach in Example Community

Black dots represent image locations. Circle color corresponds to the number of trash pieces identified on each image. Baseline values are imputed based on circles around each midline and endline assessment respectively. Circle radius is 12.5 m. A triangular kernel is used to give higher weights to closer assessments. Baseline map is shown with respect to the midline assessment. We use OpenStreetMap for all base maps.

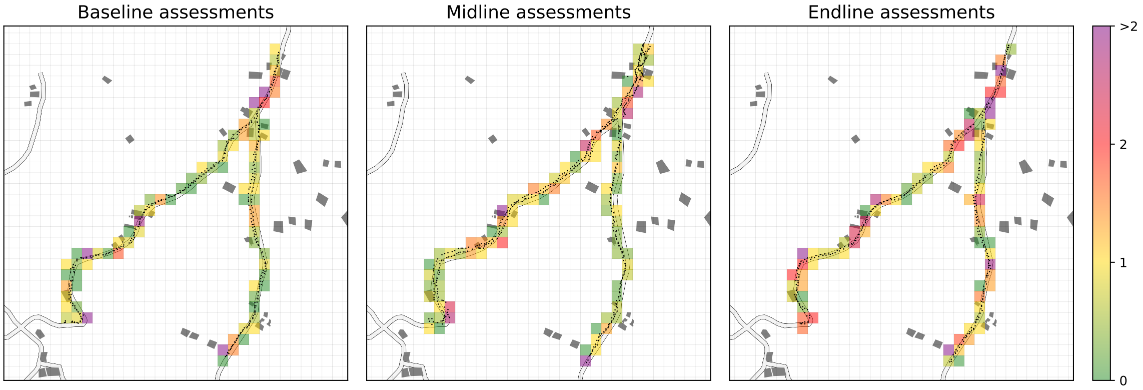


Figure 4: Illustration of Raster Approach in Example Community

Black dots represent image locations. Cell color corresponds to the average number of trash pieces identified on an image in the cell. Resolution of the raster is 0.00015 degrees (approx. 16.5 m).

5.1.2 Survey Data

To better understand the mechanisms behind potential effects, we administered short surveys to 20 residents per community. Our survey includes questions about waste-related activities respondents observed or participated in, the perceived cleanliness of the community, waste disposal and recycling behaviors, littering norms and self-reported littering behaviors, and various measures of social capital. Table A12 provides an overview on all included survey questions. Participants were selected by enumerators during the community visit for the baseline assessments. Enumerators were instructed to recruit survey participants by randomly knocking on doors until the target of 20 interviews was reached. While the resulting sample is not representative (mainly due to different propensities to be home during the day), it is very diverse and comparable across experimental groups. Our final sample consists of 2,421 individuals.[9] Attrition was 15 percent in the midline and 24 percent in the endline assessment, resulting in 2,066 observations to estimate immediate effects and 1,832 observations for long-term effects. We find no indication of differential attrition by treatment status (see Table A9). Missing values were imputed using the mean of the respective experimental group.[10]

5.1.3 Activity Registry

To gain insights into how the program was implemented, a detailed registry of all activities performed under each intervention was compiled. For the traditional intervention, we collected data on every cleaning visit, including the amount of garbage collected and the number of working hours devoted to the task. The activity registry for the community-driven intervention contained information about the type and duration of each activity, the number of participants, facilitator preparation time, and subjective ratings regarding activity success and participant interest. For cleanup campaigns, the log additionally recorded how much litter was collected in how many working hours, and how its removal was organized. All intervention activities were registered by the NGO staff responsible for conducting each activity (i.e., cleaners or facilitators). People were instructed to report honestly on all activities, and neither pay nor promotion was contingent on the successful execution of these activities. In addition, facilitators were required to submit photos of each activity to the project coordination team of the NGO. For the community-driven intervention, we also recorded all activities during the post-intervention period through phone calls to community leaders, allowing us to study to which extent collective action efforts continued after the withdrawal of the NGO. To better understand the challenges communities faced in the post-intervention period, we also conducted interviews with the person who remained in charge in each community after the end of the intervention.

[9] This sample is larger than 2,400 because we grouped 8 communities into 4 community clusters at baseline due to geographic proximity and in an effort to avoid spillovers (meaning that our analyses include 124 communities and 120 community clusters). The community clusters received the same treatment, but 40 interviews were conducted instead of 20. Throughout the study, these clusters are treated like communities.

[10] Missing values were rare, with fewer than 1 percent missings in all our main survey variables presented in Table 5.

5.2 Deep Learning for Waste Detection

We employ a novel approach that uses deep learning to create an objective measure of contamination based on the approximately 200,000 images included in our analysis. This is achieved by fine-tuning a YOLOv8 object detection model using publicly available trash datasets and manually labeled images from our own study. The YOLOv8 model is the latest addition to the YOLO (You Only Look Once) family, which comprises state-of-the-art object detection systems employed in real-time tasks for robotics, self-driving cars, and video surveillance applications (Terven and Cordova-Esparza, 2023). In contrast to other object detection models, as implied by their name, YOLO models have the ability to simultaneously identify all objects within an image. This is achieved by dividing the image into a grid and making predictions for multiple bounding boxes for each grid section, accompanied by confidence scores and a vector of class probabilities (Redmon et al., 2016). This feature marks a significant improvement in terms of speed while maintaining a high accuracy and is therefore a key factor behind the popularity of the YOLO family. YOLOv8 was released by Ultralytics, the company behind one of the older model versions (YOLOv5), in January 2023. Ultralytics offers five different model sizes, varying in features such as their mean average precision on the the popular COCO dataset (330,000 images and 200,000 annotations) and the number of parameters the model has to estimate (ranging from 3.2 million for the smallest and 68.2 million parameters for the largest model). To balance speed, accuracy and necessary computational power, we opted for the median model, YOLOv8m, with an mAP50-95 of 50.2 percent for the COCO dataset and 25.9 million estimated parameters.[11]

To fine-tune the model, we use the publicly available TACO (Trash Annotations in Context) dataset, consisting of 1,500 official images with 4,784 annotated trash bounding boxes (Proença and Simoes, 2020). The TACO data is often used as the benchmark dataset to compare the performance of different trash detection algorithms. In addition to the official images, TACO contains a set of photos with crowd-sourced annotations, which have not yet been subjected to a quality check. We manually reviewed all these unofficial images to exclude instances with incorrect bounding boxes, resulting in 3,432 additional images with 7,511 additional annotations, and a total of roughly 5,000 and 12,000 annotations for the extended TACO dataset (official + unofficial TACO). We also test if the performance is improved by adding a second

[11] The mAP (mean average precision) corresponds to the mean of the average precision (AP) over all classes and IoU (Intersection over Union) thresholds from 0.5 to 0.95 (see below for an explanation).

popular trash detection dataset, the PlastOPol data containing 2,418 images with 5,300 annotations, to the fine-tuning procedure (see Córdova et al., 2022). As the images in this dataset usually center on a single piece of trash in the foreground, they differ markedly from our own images, which depict natural settings potentially containing multiple small pieces of trash, meaning that it is a priori unclear whether adding PlastOPol to our training data would improve or degrade model performance for our task. Finally, we also include 600 manually labeled images with 3,024 annotations from our own images (200 images per wave) and 216 of our own images without any trash.

We trained our model using 70 percent of the data for training and 30 percent for testing, and computed separate performance statistics for each data source. For training, we use 200 epochs and a batch size of 8, mainly determined by computational power limitations. For prediction and evaluation, we set the detection threshold to 50 percent, meaning that objects are only detected if the model is at least 50 percent confident of its prediction. Our principal performance statistic is the AP (Average Precision), a measure that is widely used in the deep learning literature to compare results across different models. This metric is based on the area under the precision-recall curve and thus captures how well the model performs averaging over different certainty thresholds. In line with previous research, we will use AP50, meaning that a predicted bounding box is considered as accurate if the intersection between the true and the predicted box corresponds to at least 50 percent of the union of the two boxes. As additional more intuitive measures, we will also report the precision (the proportion of detected instances that are correct), the recall (the proportion of true instances that are detected), and the F1 score (a combination of precision and recall).

For the TACO dataset, the AP50 reaches 59.5 to 61.2 percent depending on whether we include the PlastOPol dataset for training or not. Table 1 illustrates that these results are similar to the best-performing models reported in the literature, ranging from an AP50 of 57.4 percent (Das et al., 2023) to an AP50 of 63.3 percent (Córdova et al., 2022). Our best model specification performs almost equally well on our own data as on the TACO dataset, achieving an AP50 of 57–59 percent. As including PlastOPol slightly decreases the AP50 for our images (Table 1), we do not use it for the training of our final model. We thus attain an AP50 of 59.0 percent, a precision of 78.6 percent, and a recall of 39.6 percent, suggesting that our model produces few incorrect detections, but misses many true instances. As many pieces of garbage are small, partially hidden, or in the background and thus difficult to detect even for human coders, this is a remarkable performance.

Table 1: Model Performance

	Our photos				*TACO*			
	AP50	Precision	Recall	F1	AP50	Precision	Recall	F1
Our model								
With PlastOPol	56.7	75.5	38.9	51.4	61.2	83.5	37.0	51.3
Without PlastOPol	59.0	78.6	39.6	52.7	59.5	82.9	34.1	48.3
Other models								
Córdova et al. (2022)	-	-	-	-	63.3	48.4	66.4	56.0
Das et al. (2023)	-	-	-	-	57.4	82.8	49.1	61.6
Majchrowska et al. (2022)	-	-	-	-	62.4	-	-	-

Majchrowska et. al (2022) included the extended TACO dataset in their performance evaluation.

The fact that our model is not perfectly accurate at detecting trash has a predictable impact on treatment effect estimates. First, we know that 21.4 percent of all detections are *false positives* due to a tendency of our model to identify other objects, typically stones or leaves, as trash. In our test set, we observe an average of 0.187 false positives per image (46 false positives for 246 images in the test set). Assuming that the number of false positives is unrelated to the treatment status, this implies that the average trash count in all experimental groups is biased upward by 0.187 pieces of trash. This does not, however, affect treatment effects, as the bias cancels out when comparing different experimental groups. A second bias is related to *false negatives*. The recall of 0.4 suggests that our model misses a bit more than half of trash on our images (i.e., the false negative rate is 0.6). Assuming that the capacity of the model to detect a given trash piece is unrelated to the treatment status, the average reported trash count for each experimental group thus corresponds to only 40 percent of the true trash count. Consequently, the treatment effect, reported in pieces of trash, is underestimated by the same factor. As the reduced differences between treatment groups are accompanied by a lower variance, this bias disappears when effects are reported in standard deviations. In summary, under plausible assumptions, raw group means and treatment effects can be biased due to the occurrence of false positives and false negatives, while standardized effects are not. When reporting on group means or effects in pieces of trash (or percent), we will thus also present results

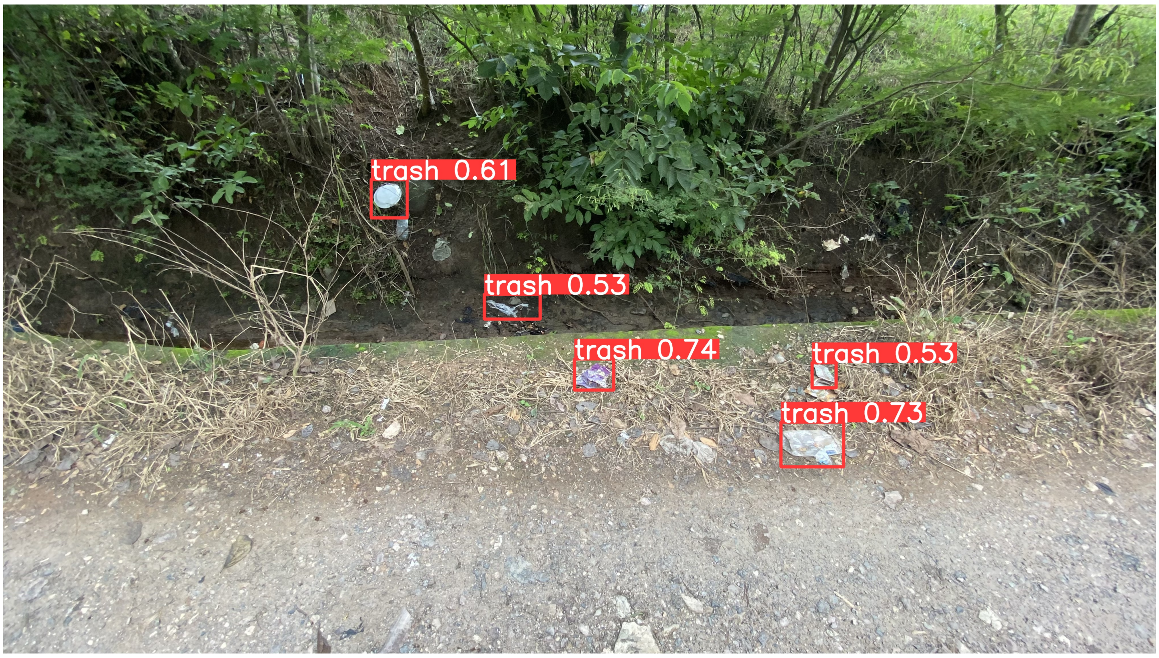


Figure 5: Illustration of Deep Learning Model Performance

The image shows model predictions for an example image. The decimal number represents the confidence of the model.

accounting for these two biases. This is done using the following simple correction:

$$Y_g = (\hat{Y}_g - FP) \cdot \frac{1}{recall} \tag{1}$$

where Y_g is the true average trash count in treatment group g after applying the bias correction to the predicted trash count $\hat{Y}_g$, FP is the average number of false positives per image in our test set and thus 0.187, and *recall* is the overall share of true trash pieces that are correctly detected in our test set and thus 0.396.[12]

[12] To correct the bias in (non-standardized) treatment effects, we only need to multiply the raw treatment effect by $\frac{1}{recall}$, since the first part of the equation cancels out. Note further that the assumptions that the probability of false positives and false negatives is unrelated to the treatment status is likely to be only approximately true. In the case of false positives, one could argue that false detections are more likely in cleaner images (where less space is covered by trash). This would introduce an additional downward bias in treatment effects, as contamination in the (cleaner) treatment group is overstated more strongly compared to the (dirtier) control group. In this case, our corrected treatment effect estimate would represent a lower bound for the true effect. A similar argument holds for false positives. If trash is harder to detect in dirtier environments (where the model may struggle to tell many different trash pieces apart), our corrected estimates would still be too conservative.

Table 2: Balance at Baseline

	Control	T1: Trad.	T2: CDD	P-value	N
Photo trash count: Contamination					
Kernel approach wrt. midline (count)	0.911	0.988	0.952	0.804	60709
Kernel approach wrt. endline (count)	0.899	0.958	0.957	0.831	65673
Raster approach (count)	0.917	1.002	1.007	0.510	13342
Raw averages (count)	0.933	0.951	0.975	0.957	120
Enumerator assessments: Contamination					
Kernel approach wrt. midline (1-4)	1.990	1.981	1.976	0.245	12216
Kernel approach wrt. endline (1-4)	1.988	1.960	1.998	0.268	13163
Raster approach (1-4)	2.016	1.976	2.014	0.454	5217
Raw averages (1-4)	2.011	1.962	2.021	0.394	120
Survey: Sociodemographics					
Female	0.740	0.756	0.722	0.392	2421
Age	42.881	42.186	43.417	0.433	2418
Education	2.526	2.468	2.394	0.362	2421
Poverty (1-5)	3.162	3.046	2.979	0.178	2354
Community size	305.349	300.439	294.964	0.784	2420
Survey: Contamination and waste disposal					
Perceived cleanliness (1-5)	3.146	3.103	3.118	0.587	2421
Appropriate disposal (%)	0.452	0.532	0.489	0.530	2421
Survey: Social norms					
Littering (%)	0.599	0.603	0.571	0.131	2418
Littering is bad (%)	0.705	0.687	0.665	0.133	2413
Punish littering (%)	0.565	0.562	0.532	0.108	2417
Survey: Social capital					
Strong ties (%)	0.339	0.404	0.312	0.259	2419
Weak ties (%)	0.695	0.712	0.645	0.276	2375
Trust (1-5)	3.485	3.499	3.629	0.249	2421
Organizations (%)	0.174	0.191	0.207	0.445	2421
Voluntary work (%)	0.252	0.296	0.338	0.036	2415
Altruism (%)	0.439	0.454	0.406	0.038	2421

The last row indicates the p-value of a joint F-test that each treatment dummy coefficient is equal to 0. Education refers to highest completed degree: None = 1, incomplete primary = 2, complete primary = 3, high school degree = 4), technical = 5, and university degree = 6. Standard errors are clustered at the community level.

5.3 Baseline Characteristics

Table 2 shows that contamination levels and survey responses at baseline are well-balanced across experimental groups. Only for two of our main variables, the share of people engaging in voluntary work and the percentage of an endowment people choose to donate in a framed dictator game (altruism), we report significant differences between groups.

Our deep learning model detects approximately one piece of garbage on the average image. Based on Equation 1 in Section 5.2, this implies that an average image contains about 2 real pieces of trash. Considering that the photos were taken randomly along all streets and not specifically in places with garbage, this indicates substantial solid waste contamination. For the average community, this corresponds to roughly 1,000 visible pieces of trash on our images alone. There is considerable variance between communities with 0.19 detections (hardly any real pieces) on the average image in the least polluted community and 3.12 detections ($\approx$ 7.38 real pieces) in the most polluted community. Similarly, enumerators rated the average site across all communities as a 2 ("a bit polluted") on a scale from 1 to 4. Community averages based on these subjective enumerator assessments range from 1.37 in the cleanest community to 2.78 in the dirtiest community.

The average survey respondent is 43 years old, and 75 percent of respondents are female. About two-thirds of the individuals in our sample have not completed any educational degree (no schooling: 20%, incomplete primary: 45%), 11 percent have a primary degree, 19 percent have completed high school, and 5 percent possess a tertiary degree. On average, respondents believe that roughly 60 percent of people in their community litter, that 70 percent of people in their community disprove of littering, and that 55 percent of people in their community would punish litterers with a disapproving gesture. The average community has about 300 residents, corresponding to roughly 90 households. People tend to know each other, with the average person reporting that 70 percent of community members are known and 40 percent are friends or family. Approximately 20 percent of respondents belong to a community organization and 30 percent report having done voluntary work for the community in the last month.

6 Empirical Results

This chapter discusses the main findings of our study. We will (i) take a look at how the program was implemented, (ii) present our main findings, and finally (iii) use insights from the survey and the activity registry to discuss potential mechanisms based on the theoretical framework developed in Section 2.

6.1 Program Implementation

Our data suggests that both interventions were successfully implemented. In the traditional intervention, an average of 3.9 cleanups were conducted in each community (i.e., roughly one per month), and no community received fewer than 3 cleanups. The community-driven intervention was implemented in 95 percent of the communities assigned to this condition, with an average of 0.9 meetings, 14.3 educational activi-

Table 3: Community Activities Summary Statistics

Activity	*During intervention*			*After intervention*	
	Completed activities	% with one or more	Participants	Completed activities	% with one or more
Sessions	7.20	0.95	18.90	0.08	0.08
Workshops	3.50	0.95	18.40	0.00	0.00
Community activities	3.60	0.95	19.10	0.03	0.03
Cleanup campaigns	3.40	0.92	18.70	2.05	0.67
Meetings	0.90	0.90	21.70	1.38	0.33

The period for both during as well as after the intervention spans a total of 4 months. Intervention activities were recorded by facilitators. The post-intervention data was obtained through phone calls to the responsible person at each community. The number of participants is conditional on the activity taking place.

ties (two-hour sessions, practical workshops, or community activities such as movie nights), and 3.5 collective cleanups per community (Table 3). This corresponds to a total of 18.6 activities and 4.7 monthly activities per community. About 20 community members, corresponding to 7 percent of the population, participated in a typical activity in this intervention arm.

For the community-driven intervention, we also collected data during the four months following the intervention to observe whether efforts to keep the community clean continued. In the post-intervention period, around two-thirds of the communities report conducting at least one cleanup campaign, with 2.05 campaigns (0.51 per month) in the average community. Similarly, about one third of all communities realized at least one meeting about solid waste contamination. In line with expectations, educational activities were largely discontinued in the post-intervention period.

Our survey data shows that the sudden increase in activities related to solid waste management activities associated with the interventions did not go unnoticed (Table A1). The community-driven intervention had a large and significant immediate impact on the percentage of respondents who reported being aware of various waste-related activities – namely community meetings, education sessions, or collection campaigns – in their community within the past four months. In addition, it raised the number of people claiming to have participated in each of these activities. The traditional intervention also increased the number of individuals observing or participating in cleaning efforts, though to a lesser extent than the community-driven intervention.

For the CCD treatment, with the exception of educational sessions, substantial effects on all activities persist into the post-intervention period, while no lasting impacts are observed for the traditional intervention. Overall, our activity registries and survey data consistently indicate proper implementation of both interventions according to the specifications of each experimental group.

6.2 Program Effects

To assess the causal effect of the two treatments on contamination levels for each post-treatment $wave \in \{midline, endline\}$, we use

$$Y_{iv}^{wave} = \alpha + \beta_1 T_1 + \beta_2 T_2 + \delta Y_{iv}^{baseline} + \mu_s + \epsilon_{iv} \tag{2}$$

where Y_{iv}^{wave} are midline or endline outcomes for kernel or raster cell i in village v; T_1 and T_2 are treatment indicators for treatment 1 (traditional intervention) and treatment 2 (community-driven development); $Y_{iv}^{baseline}$ is the baseline kernel or cell contamination level; and μ_s are strata fixed effects. With the exception of the models analyzing effects on raw community averages, standard errors are clustered at the community level. For survey outcomes, we extend Equation 2 by adding individual-level controls for sex, age, and education.

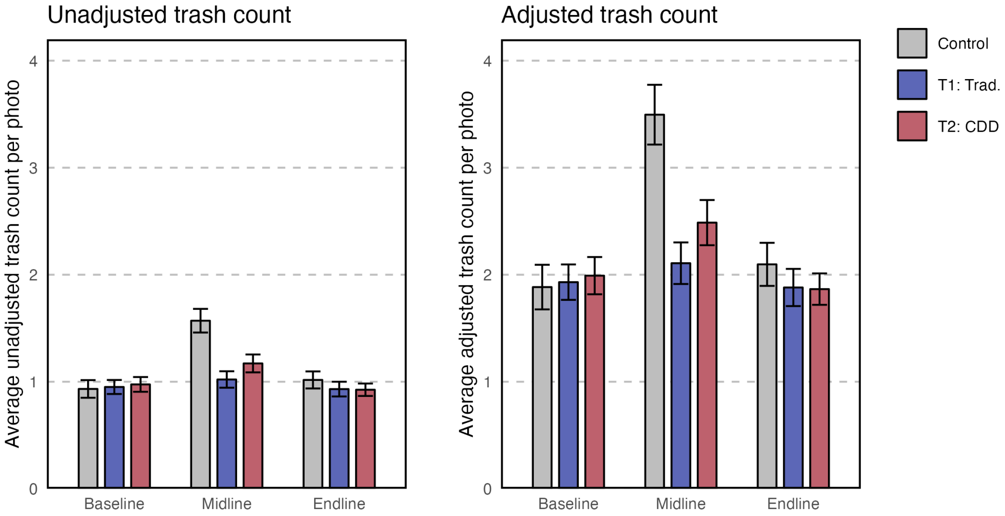


Figure 6: Average Trash Count per Image by Wave and Treatment

The baseline measurement was conducted in September and October 2022, the midline in March 2023, and the endline in July 2023. The increase in the amount of litter during the midline assessment is likely to be a seasonal effect, as this was the only measurement conducted during the dry season, when waste is less likely to be washed away or covered by vegetation.

Table 4: Main Results Based on Trash Detection and Enumerator Assessments

	Immediate effects				*Long-term effects*			
	T1: Trad.	T2: CDD	T2 - T1	N	T1: Trad.	T2: CDD	T2 - T1	N
Photo trash detection								
Kernel approach	-0.755***	-0.540***	0.215**	60709	-0.129	-0.199**	-0.070	65673
	(0.129)	(0.130)	(0.107)		(0.107)	(0.100)	(0.096)	
Raster approach	-0.727***	-0.471***	0.256**	10740	-0.134	-0.200*	-0.066	10740
	(0.142)	(0.158)	(0.128)		(0.116)	(0.109)	(0.106)	
Raw averages	-0.792***	-0.604***	0.188	120	-0.175	-0.248**	-0.073	120
	(0.119)	(0.120)	(0.121)		(0.113)	(0.114)	(0.115)	
Enumerator assessments								
Kernel approach	-0.932***	-0.771***	0.161	12216	-0.047	-0.057	-0.009	13163
	(0.184)	(0.178)	(0.193)		(0.210)	(0.173)	(0.230)	
Raster approach	-0.803***	-0.616***	0.187	4272	0.007	-0.006	-0.012	4272
	(0.185)	(0.176)	(0.185)		(0.218)	(0.171)	(0.234)	
Raw averages	-0.999***	-0.853***	0.146	120	-0.089	-0.088	0.000	120
	(0.204)	(0.204)	(0.206)		(0.195)	(0.195)	(0.197)	

Results reported in standard deviations at the community level. Controls include contamination at baseline and strata fixed effects. Standard errors are clustered at the community level for the kernel and the raster approach. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$

For our main outcome based on *trash counts*, we find large immediate effects for both interventions (Figure 6, Table 4, and Table A2). The traditional intervention reduced solid waste contamination by 0.7–0.8σ or roughly 0.5 detected trash pieces on the average image ($p < 0.01$). Applying our bias correction, this corresponds to a decrease of about 1.25 trash pieces or 36 percent. The community-driven intervention had a significantly smaller ($p < 0.05$), but still substantial impact of 0.5–0.6σ or approximately 0.4 trash detections ($p < 0.01$). This translates into an effect of 1 piece of garbage or 29 percent. Estimates for long-term impacts reveal a stark depletion of effects for both treatments four months after the end of the intervention. Communities in the traditional intervention outperform the control group by only 0.1σ, an effect that is statistically indistinguishable from zero at conventional levels (p $\approx$ 0.2). This corresponds to a depreciation by about 0.6σ or 80 percent compared to immediate effects. For the community-driven intervention, we document a slightly larger and statistically significant long-term effect of 0.2σ or 0.25 trash pieces ($p < 0.05$). The depletion of immediate impacts corresponds to about 0.3σ or 60 percent. While the absolute depreciation (i.e., in standard deviations) is significantly lower in the CDD intervention than in the traditional intervention ($p < 0.01$), the difference in relative depletion rates is not statistically significant ($p = 0.2$, see Table A4).[13]

As a robustness check, we compare these results with effects based on *subjective enumerator assessments* (lower panel in Table 4 and Table A2). In line with our main outcome based on trash detections, we observe large immediate effects for both interventions. However, the difference between the two treatments is no longer significant and the long-term effects disappear. A likely explanation for these deviations is that the subjectivity of the ratings introduced considerable noise into the assessment measure. While the resulting measurement errors should be uncorrelated with the treatment, they are clustered at the community level (because enumerators always covered an entire community), which considerably reduces the precision of the estimates. Indeed, if we include enumerator fixed effects, estimates for long-term effects based on enumerator assessments change markedly, indicating significant long-term effects for both treatments (Table A3). Our main results based on trash detection are less sensitive to the inclusion of these fixed effects. This underscores the advantages of the objective contamination measure that we derive using deep learning.

Our *survey results* show that the changes in solid waste pollution did not go unnoticed (Table 5, panel "Contamination and waste disposal"). In line with our findings from trash detections and contamination ratings, perceived cleanliness improved significantly by about 0.16σ immediately after both treatments. In addition, both interventions had a significant short-term impact on people's recycling practices, with a 10 percentage point increase in the share of people recycling at least one type of solid waste. For the community-driven intervention, we further report a significant immediate improvement in self-reported waste disposal practices. The share of people indicating that they use an official deposit or a garbage truck to dispose of their waste, as opposed to burning, burying, or dumping it, increased by about 10 percentage points. Estimates for long-term effects show that impacts on perceived cleanliness persist, with effects of 0.17–0.18σ for both interventions. For the CDD intervention, we further report a sustained increase in the share of people indicating appropriate waste disposal by roughly 7 percentage points (50% depreciation compared to immediate effects). Recycling effects disappear in the long run for both interventions.

[13] Whether absolute or relative depreciation is more appropriate depends on the assumptions about counterfactual trends in the two groups. Under a parallel trends assumption, absolute depreciation would be the correct measure. On the other hand, if we assume convergence back to the level of the control group, we should use a relative measure. Since the second scenario seems more plausible, we use relative depreciation as our main measure.

Table 5: Survey Regression Results

	Immediate effects			*Long-term effects*		
	T1: Trad.	T2: CDD	T2 - T1	T1: Trad.	T2: CDD	T2 - T1
Contamination and waste disposal						
Perceived cleanliness (sd)	0.158*	0.163**	0.005	0.173**	0.182**	0.009
	(0.085)	(0.083)	(0.076)	(0.079)	(0.082)	(0.078)
Appropriate disposal (%)	0.040	0.137***	0.097**	-0.005	0.067*	0.072*
	(0.044)	(0.041)	(0.046)	(0.042)	(0.039)	(0.040)
Recycling (%)	0.083**	0.112***	0.029	-0.026	-0.001	0.025
	(0.041)	(0.037)	(0.029)	(0.040)	(0.040)	(0.040)
Social norms						
Littering (%)	-0.065**	-0.104***	-0.039	0.011	-0.019	-0.030
	(0.028)	(0.026)	(0.026)	(0.019)	(0.019)	(0.021)
Littering is bad (%)	-0.015	-0.037**	-0.022	-0.001	0.004	0.005
	(0.020)	(0.017)	(0.020)	(0.018)	(0.018)	(0.016)
Punish littering (%)	0.027	0.003	-0.024	0.017	0.036*	0.019
	(0.028)	(0.025)	(0.027)	(0.022)	(0.020)	(0.022)
Social capital						
Strong ties (%)	0.059	0.059*	0.001	0.024	0.006	-0.018
	(0.038)	(0.034)	(0.038)	(0.031)	(0.029)	(0.032)
Weak ties (%)	-0.000	0.007	0.007	0.006	0.023**	0.017**
	(0.017)	(0.014)	(0.016)	(0.010)	(0.009)	(0.008)
Trust (sd)	0.113	0.031	-0.082	0.013	0.054	0.041
	(0.095)	(0.090)	(0.096)	(0.086)	(0.080)	(0.081)
Organizations (%)	-0.007	0.042	0.049*	-0.009	0.011	0.020
	(0.023)	(0.026)	(0.028)	(0.023)	(0.025)	(0.027)
Voluntary work (%)	0.011	0.159***	0.149***	0.035	0.129***	0.094***
	(0.031)	(0.034)	(0.037)	(0.029)	(0.031)	(0.031)
Altruism (%)	-0.014	-0.001	0.013	0.012	0.017	0.005
	(0.019)	(0.020)	(0.020)	(0.017)	(0.015)	(0.015)

Sample sizes are n = 2066 for the estimation of immediate effects and n = 1832 for long-term effects. Social norm variables refer to beliefs about other people's behavior. Controls include strata fixed effects, sex, age, education (dummies), and the baseline value for the respective outcome. Standard errors are clustered at the community level. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$

6.3 Discussion

Our complementary data from surveys, activity records, and interviews allow us to explore the mechanisms behind the observed effect patterns using the theoretical framework we propose in Section 2. In this chapter, we shed light on two key policy questions arising from our project. We will (i) explore the extent to which the CDD intervention may have addressed informational, organizational, and credit constraints, and (ii) discuss if impacts were mainly driven by cleaning efforts or by changes in littering behavior.

6.3.1 How Can CDD Alleviate Constraints to Collective Action?

Community-based initiatives can mitigate *information constraints* in two ways. Residents could become more aware of the problem and of effective means to address it, inducing them to lower their thresholds for cooperative behavior, or they could correct their (potentially biased) beliefs about the number of others who are contributing. In either case, a successful intervention would induce a gradual shift toward a stable higher equilibrium, as more and more individuals join the camp of cooperators. Thus, we should observe increasing participation in cleanups over time, and a gradual and sustained reduction in littering. We find limited evidence for either of these patterns. The number of participants in the average cleanup was stable throughout the intervention and the post-intervention period, suggesting that when campaigns were organized, similar numbers of residents continued to participate.[14] Results for littering behavior are inconclusive as well. In line with potential information effects, survey respondents tend to be much more positive about their own littering behavior than that of their neighbors, and the community-driven intervention narrowed this gap: The CDD treatment reduced the proportion of residents who respondents believed to engage in littering by about 10 percentage points ($p < 0.01$, Table 5, panel “Social norms”).[15] However, a similar change in descriptive norms, namely a reduction by 7 percentage points, occurred in the traditional intervention, and both effects disappear in the long run. In addition, no clear effects are found for all other outcomes related to littering norms and behaviors (Table A7). This is consistent with

[14] We also inspected separate trends for all communities to see if averages mask diverging trends toward high levels of collective action in some communities and low levels in others, and find no support for this hypothesis.

[15] While only 15 percent of people say they have littered in the past month, the average person believes that 60 percent of others have done so (see Table A5). Note, however, that this does not necessarily indicate that people's perceptions are biased, as responses about self-reported behavior may be driven by a social desirability bias.

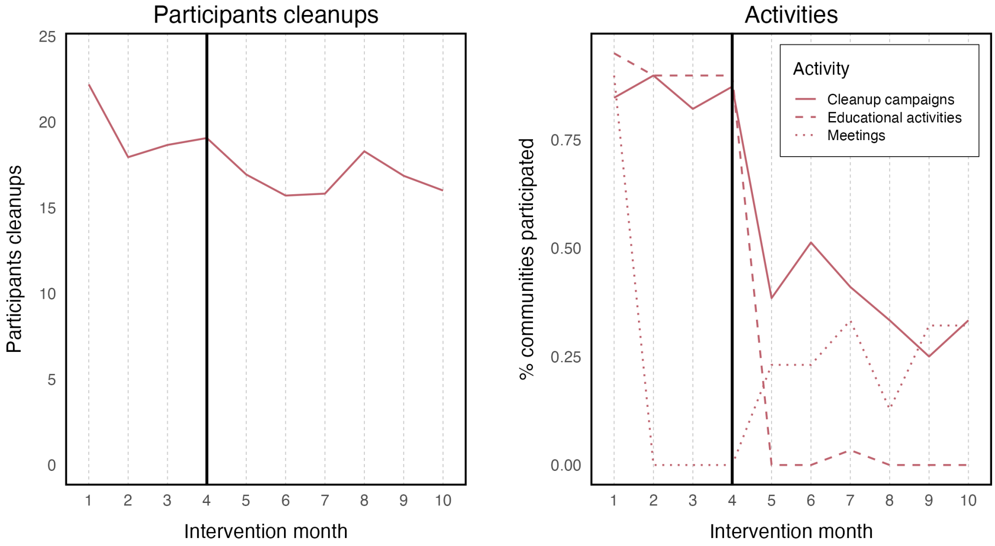


Figure 7: Collective Action in the CDD Intervention Over Time

The left figure shows the number of participants in the average cleanup per month. The right figure documents what share of communities realized different types of activities in each month. The black line corresponds to the end of the intervention.

our registry data on the amount of trash collected during cleanups (see Figure 8). While the average number of garbage bags collected in CDD cleanups decreases by about 50 percent from the first to the last intervention month – a potential indication for reduced littering – we observe a very similar and statistically indistinguishable decline for the traditional intervention. Similarly, we do not find a steeper reduction in the total number of working hours required for cleaning in the CDD intervention than in the traditional intervention.[16]

A key argument for CDD interventions is that they alleviate *organizational constraints* to collective action, thereby enabling communities to coordinate the provision of public goods. If groups manage to get organized and agree on joint actions, this would lead to an immediate shift towards a higher equilibrium. This is consistent with the observation that participation in cleanups was high from the first month and remained stable throughout the intervention. This suggests that a sufficiently large number of community members were willing to (conditionally) commit time to a cleaner environment from the outset, and that the intervention succeeded in bringing them together to do so. Organizational effects can either be limited to the

[16] Note that the reduction in collected waste (or the time used to do so) over the course of the intervention is not only driven by social norms, but also by the fact that waste in the early months may have accumulated over longer periods of time. For work hours, we may also observe changes in the efficiency of the group, as a large group may be more productive at collecting large amounts of waste compared to smaller amounts.

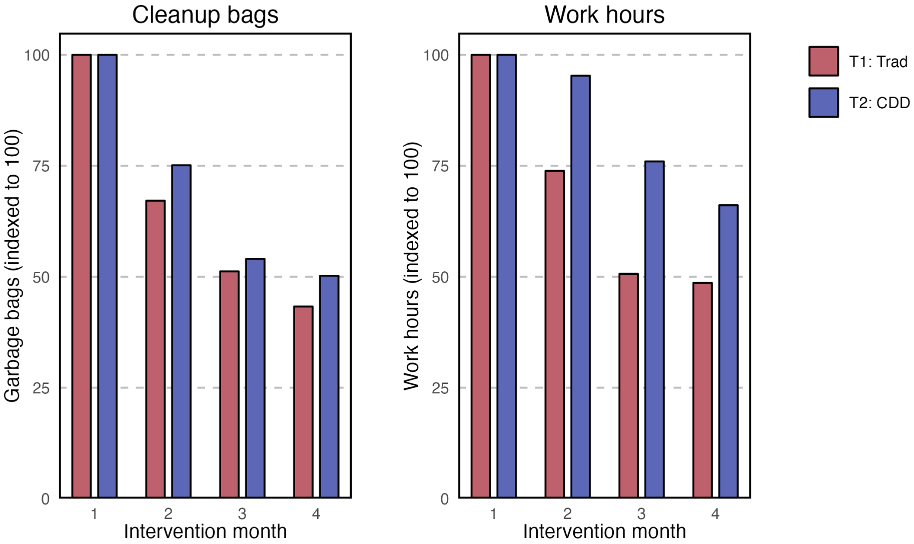


Figure 8: Cleanup Statistics Over Time by Treatment

The left figure shows the number of garbage bags filled in the average cleanup over time. The right figure documents the average number of working hours (added over all contributors) needed for the task. To make trends comparable, results are expressed as percentages of the treatment-specific average of bags (T1: 4.7, T2: 9.5) and work hours (T1: 11.6, T2: 15.1) in the first month.

intervention period, where paid facilitators take the lead in mobilizing for collective action, or, ideally, be enduring if communities succeed in strengthening local institutions. Our survey data provides only limited support for the latter type of effects (Table A12, panel "Social Capital"). We present clear evidence that the CDD intervention increased engagement in voluntary work (likely through participation in the cleanups) and suggestive evidence that it improved social ties (strong ties in the short run and weak ties in the long run), but we find no immediate or lasting effects on trust, membership in organizations, and altruism. Together with the steep decline in collective action immediately after the end of the intervention (Figure 7), these findings suggest that much of the success in the organizational dimension was tied to the presence of the facilitator. Mobilizing for collective action is time-consuming and demands a disproportionate contribution from the person (or persons) taking the lead in the endeavor. If people are willing to contribute about as much as others do, no such leader will emerge to take over from the facilitator. This aligns with qualitative evidence from interviews with the community members who assumed responsibility after the departure of the facilitator, where "time constraints to mobilize people and organize campaigns" emerged as the most frequently mentioned challenge to project continuation. It is also consistent with our heterogeneity analyses, which suggest that the CDD intervention had a higher short-term impact at lower initial levels of social capital ($p = 0.07$), where organizational constraints addressed by the temporary leadership of the facilitator might have been more binding (Figure A2).

A final channel through which CDD interventions could facilitate the provision of public goods is by easing *credit constraints*. This mechanism is less relevant for the particular public good we study, because a clean environment can be maintained with a minimal financial investment. Removing litter from the streets can be accomplished with voluntary work and a few plastic bags, and communities typically took advantage of the municipal garbage truck or a resident's journey to transport the collected waste to an official depot.[17] Accordingly, no financial transfer was made to communities in the context of the project. A notable exception is the provision of snacks to volunteers during the collective cleanups. However, few interviewees mentioned the lack of such provisions as a key constraint to the continuation of collective action activities in the post-intervention phase. Hence, it appears unlikely that the short-term and partial long-term success of the CDD intervention was driven by mechanism related to credit constraints.

Overall, our results are most consistent with a theoretical model where many individuals are willing to contribute to public goods as long as others do so too, but struggle to coordinate in the absence of a dedicated leader. Community-based interventions have the potential to build leadership and strengthen the local institutions needed to coordinate collective action. However, achieving transformations that outlive the presence of a paid facilitator may often be beyond the scope of a four-month intervention.

6.3.2 Should Solid Waste Interventions Aim for Cleanups or Changes in Littering Norms?

Communities can pursue two interrelated strategies to provide the public good of a clean environment. They can either mobilize for regular collective cleanups, or establish informal institutions to discourage littering in the first place. The complementary data discussed in the previous section can also be used to gauge the importance of each of these channels. The cleanups clearly played an important role, as the typical CDD

[17] Our post-intervention interviews with community leaders reveal that removing the collected trash from the community was an major challenge in a few communities where the municipality charged a (usually substantial) fee to send the garbage truck or a private vehicle had to be hired. However, no financial support was provided for the removal of the collected waste during the intervention, and facilitators successfully devised solutions in coordination with community members. This underscores that waste transportation was primarily an organizational challenge rather than a financial one.

community conducted about one monthly cleanup with roughly 20 participants (Table 3), and the intervention had a large effect on the proportion of survey respondents who reported observing or participating in such campaigns (Table A1). As discussed above, the results for littering behavior are more mixed. The shift in beliefs about other people's littering practices induced by the CDD intervention was mirrored by a similar change for the traditional intervention, and did not persist after the end of the program (Table 5). Similarly, while we find that the amount of trash collected in monthly campaigns decreased substantially over time, this decline was not greater for the bottom-up than for the top-down intervention (Figure 8). A plausible explanation is that individuals form their beliefs about other people's littering behavior based on the amount of waste they observe on the streets, and modify their own practices in response to this inferred social norm. This is in line with ample research documenting that people are substantially less likely to litter in clean than in dirty environments (Cialdini et al., 1990; Ramos and Torgler, 2012; Bateson et al., 2013; Sagebiel et al., 2020). As the two interventions lead to similar reductions in solid waste pollution due to the cleaning efforts, individuals in both treatment groups may have concluded that fewer people are littering and, potentially, adapted their own behaviors accordingly.

Overall, our data points to the cleanups as the main driver of the success of both interventions. While a shift in littering norms may also have played a role, our data does not provide much support for the hypothesis that the CDD intervention was more effective in inducing this change. Viewed through the theoretical framework developed in Section 2, our findings suggest that interventions focusing on changes in littering behavior alone are unlikely to be sustainable. Maintaining a clean environment without any cleaning requires perfect adherence to a non-littering norm by all community members and visitors. If a small minority litters regardless of what others do, waste will accumulate, inducing conditional cooperators to start littering as well. As a result, communities will revert to a low equilibrium where everyone litters except those who are willing to cooperate irrespective of what others do. In contrast, reaching a stable high equilibrium through collective cleanups requires the cooperation of only a small group of committed residents, which may be much easier to achieve. The positive dynamics induced by the cleanups may then be reinforced by changes in littering behavior, as people are less likely to dump waste into clean environments.

7 Conclusion

Community-driven development has become a popular alternative to the conventional top-down approach to the provision of public goods. While several recent studies have evaluated such programs, their effectiveness has not yet been compared to the more traditional strategy they often replace. In this study, we present the results of a randomized controlled trial comparing the effectiveness of bottom-up and top-down strategies to address local waste pollution in rural El Salvador. Immediate effects on contamination level are substantial for both interventions, but significantly larger in the traditional intervention. Four months after the end of the intervention, we observe a strong diminution of these effects, which is only slightly less pronounced in the community-driven intervention. Our complementary data suggests that the presence of the facilitator may have helped the communities overcome organizational constraints to collective action, but many communities were unable to sustain these efforts independently.

Our findings have important implications for the policy debates around *community-driven development*. We find that while CDD initiatives can indeed successfully promote the provision of local public goods, they are not always more effective in doing so than top-down interventions. More specifically, our findings highlight that many individuals are willing to voluntarily contribute to public goods, and involving them in the development of their communities may indeed produce more sustainable outcomes. However, sustaining the high levels of collective action needed to provide public goods at optimal levels requires strong informal institutions and local leadership. Building such capabilities may be beyond the scope of a short-term intervention, and entail considerable costs, including facilitation expenses for the implementing organization, and opportunity costs for participants. A combined approach that strengthens government institutions alongside communities may thus be a promising long-term strategy. How much and what kind of bottom-up participation produces the most sustainable and cost-effective solutions is an important question for future research. In this context, two important limitations of our study should not go unmentioned. First, our study is based on the provision of a specific public good in a particular context, meaning that more research is needed to draw confident conclusions about the relative effectiveness of bottom-up development initiatives. Second, the top-down intervention in our study was implemented by a committed NGO rather than a governmental institution and its effectiveness may thus be an upper bound for what a state-led arrangement in developing countries could achieve. Nevertheless, by providing a first rigorous comparison between a top-down and a bottom-up provision strategy, our study constitutes a critical starting point for the necessary discussion on the relative effectiveness of different approaches to local public good provision.

The findings presented in this study are also relevant to policy makers seeking to devise effective *solid waste management* strategies. Based on our findings and theoretical considerations, we draw two cautious conclusions. First, raising awareness and empowering communities to tackle the waste problem can be an important part of the solution, but the assumption that a one-time investment in facilitation will effectively solve the problem forever is clearly unrealistic. Second, picking up waste may be more critical to the success of waste management interventions than inducing changes in littering behavior. Although shifts in social norms may reinforce the positive trend induced by cleaning efforts, interventions focusing exclusively on littering behavior are unlikely to lead to a stable high-level equilibrium. In light of the rapid increase in solid waste production in developing countries and the scarcity of research on how best to address the problem, these are crucial and timely insights.

Finally, our study also advances the use of *deep learning methods* to understand, track, and improve outcomes related to global development. A rapidly growing body of research has shown that a variety of outcomes, including poverty, education or agricultural yields, can be predicted from alternative data sources such as satellite imagery (Kuwata and Shibasaki, 2015; Jean et al., 2016; Yeh et al., 2020), phone records (Blumenstock et al., 2015), social media posts (Jakob and Heinrich, 2023), or Google Street View images (Suel et al., 2019). However, the main focus of this literature is on proof-of-concept, and applications that bridge real gaps in data availability remain scarce. By using image data and deep learning to derive an objective measure of contamination, our study provides such an application. We illustrate how predicted measures can be used in an experimental setup, and how potential biases can be accounted for. As deep learning methods continue to penetrate the social sciences, such applications and discussions of the biases they may introduce, are likely to become increasingly important.

3 References

Anderson, L. R., Mellor, J. M., and Milyo, J. (2004). Social capital and contributions in a public-goods experiment. *American Economic Review*, 94(2):373–376.

Arcand, J.-L. (2008). Does community driven development work? Evidence from Senegal. Available at SSRN: `http://dx.doi.org/10.2139/ssrn.1265231`.

Avdeenko, A. and Gilligan, M. J. (2015). International interventions to build social capital: Evidence from a field experiment in Sudan. *American Political Science Review*, 109(3):427–449.

Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R., and Khemani, S. (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India. *American Economic Journal: Economic Policy*, 2(1):1–30.

Banerjee, A. V. and Duflo, E. (2007). The economic lives of the poor. *Journal of Economic Perspectives*, 21(1):141–167.

Bateson, M., Callow, L., Holmes, J. R., Redmond Roche, M. L., and Nettle, D. (2013). Do images of "watching eyes" induce behaviour that is more pro-social or more normative? A field experiment on littering. *PloS One*, 8(12):e82055.

Beath, A., Christia, F., and Enikolopov, R. (2013). Empowering women through development aid: Evidence from a field experiment in Afghanistan. *American Political Science Review*, 107(3):540–557.

Berger, J. (2021). Social tipping interventions can promote the diffusion or decay of sustainable consumption norms in the field. Evidence from a quasi-experimental intervention study. *Sustainability*, 13(6):3529.

Berger, J., Efferson, C., and Vogt, S. (2023). Tipping pro-environmental norm diffusion at scale: Opportunities and limitations. *Behavioural Public Policy*, 7(3):581–606.

Björkman, M., de Walque, D., and Svensson, J. (2017). Experimental evidence on the long-run impact of community-based monitoring. *American Economic Journal: Applied Economics*, 9(1):33–69.

Björkman, M. and Svensson, J. (2009). Power to the people: Evidence from a randomized field experiment on community-based monitoring in Uganda. *The Quarterly Journal of Economics*, 124(2):735–769.

Blumenstock, J., Cadamuro, G., and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076.

Casey, K. (2018). Radical decentralization: Does community-driven development work? *Annual Review of Economics*, 10:139–163.

Casey, K., Glennerster, R., and Miguel, E. (2012). Reshaping institutions: Evidence on aid impacts using a preanalysis plan. *The Quarterly Journal of Economics*, 127(4):1755–1812.

Castaldi, G., Cecere, G., and Zoli, M. (2021). Smoke on the beach: On the use of economic vs behavioral policies to reduce environmental pollution by cigarette littering. *Economia Politica*, 38:1025–1048.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics*, 14:47–83.

Chitotombe, J. W. (2014). Interrogating factors associated with littering along road servitudes on Zimbabwean highways. *Environmental Management and Sustainable Development*, 3(1):181.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6):1015.

Córdova, M., Pinto, A., Hellevik, C. C., Alaliyat, S. A.-A., Hameed, I. A., Pedrini, H., and Torres, R. d. S. (2022). Litter detection with deep learning: A comparative study. *Sensors*, 22(2):548.

Cowen, T. (1992). *Public goods and market failures: A critical examination.* Transaction Publishers.

Dahlman, C. J. (1979). The problem of externality. *The Journal of Law and Economics*, 22(1):141–162.

Das, D., Deb, K., Sayeed, T., Dhar, P. K., and Shimamura, T. (2023). Outdoor trash detection in natural environment using a deep learning model. *IEEE Access*.

Desai, R. M. and Olofsgård, A. (2019). Can the poor organize? Public goods and self-help groups in rural India. *World Development*, 121:33–52.

Dongier, P., Van Domelen, J., Ostrom, E., Ryan, A., Wakeman, W., Bebbington, A., Alkire, S., Esmail, T., and Polski, M. (2003). Community driven development. *World Bank Poverty Reduction Strategy Paper*, 1:303–327.

Duflo, E., Dupas, P., and Kremer, M. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics*, 123:92–110.

Dur, R. and Vollaard, B. (2015). The power of a bad example: A field experiment in household garbage disposal. *Environment and Behavior*, 47(9):970–1000.

Fearon, J. D., Humphreys, M., and Weinstein, J. M. (2009). Can development aid contribute to social cohesion after civil war? Evidence from a field experiment in post-conflict Liberia. *American Economic Review*, 99(2):287–91.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71(3):397–404.

Gächter, S. (2006). Conditional cooperation: Behavioral regularities from the lab and the field and their policy implications. In *CeDEx Discussion Paper No. 2006–03 Available at: http://hdl.handle.net/10419/67977.*

Glowacki, L. and von Rueden, C. (2015). Leadership solves collective action problems in small-scale societies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1683):20150010.

Hardin, R. (1971). Collective action as an agreeable n-prisoners' dilemma. *Behavioral Science*, 16(5):472–481.

Hardin, R. (1982). *Collective action.* Johns Hopkins University Press., Baltimore, MD.

Humphreys, M., Sánchez de la Sierra, R., and Van der Windt, P. (2019). Exporting democratic practices: Evidence from a village governance intervention in Eastern Congo. *Journal of Development Economics*, 140:279–301.

Jack, B. K. and Recalde, M. P. (2015). Leadership and the voluntary provision of public goods: Field evidence from Bolivia. *Journal of Public Economics*, 122:80–93.

Jakob, M. S. and Heinrich, S. (2023). Measuring human capital with social media data and machine learning. *University of Bern Social Sciences Working Papers No. 46.* Available at: https://ideas.repec.org/p/bss/wpaper/46.html.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Kaza, S., Yao, L., Bhada-Tata, P., and Van Woerden, F. (2018). *What a waste 2.0: a global snapshot of solid waste management to 2050.* World Bank Publications.

Keser, C. and Van Winden, F. (2000). Conditional cooperation and voluntary contributions to public goods. *Scandinavian Journal of Economics*, 102(1):23–39.

Kuwata, K. and Shibasaki, R. (2015). Estimating crop yields with deep learning and remotely sensed data. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 858–861.

Labonne, J. and Chase, R. S. (2011). Do community-driven development projects enhance social capital? Evidence from the Philippines. *Journal of Development Economics*, 96(2):348–358.

Lewis, A., Turton, P., and Sweetman, T. (2009). *Litterbugs: How to deal with the problem of littering.* Policy Exchange.

Liu, J. H. and Sibley, C. G. (2004). Attitudes and behavior in social space: Public good interventions based on shared representations and environmental influences. *Journal of Environmental Psychology*, 24(3):373–384.

Majchrowska, S., Mikołajczyk, A., Ferlin, M., Klawikowska, Z., Plantykow, M. A., Kwasigroch, A., and Majek, K. (2022). Deep learning-based waste detection in natural and urban environments. *Waste Management*, 138:274–284.

Mansuri, G. and Rao, V. (2012). *Localizing development: Does participation work?* World Bank Publications.

McKenzie, D. and Bruhn, M. (2011). Tools of the trade: Doing stratified randomization with uneven numbers in some strata. Available at: https://blogs.worldbank.org/impactevaluations/tools-of-the-trade-doing-stratified-randomization-with-uneven-numbers-in-some-strata. Last accessed: 2023-10-29.

Mohan, S. and Joseph, C. P. (2021). Potential hazards due to municipal solid waste open dumping in India. *Journal of the Indian Institute of Science*, 101(4):523–536.

Nepal, M., Karki Nepal, A., Khadayat, M. S., Rai, R. K., Shyamsundar, P., and Somanathan, E. (2023). Low-cost strategies to improve municipal solid waste management in developing countries: Experimental evidence from Nepal. *Environmental and Resource Economics*, 84(3):729–752.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.

Nguyen, T. C. and Rieger, M. (2017). Community-driven development and social capital: Evidence from Morocco. *World Development*, 91:28–52.

Nkwocha, E. E. and Okeoma, I. O. (2009). Street littering in Nigerian towns: Towards framework for sustainable urban cleanliness. *African Research Review*, 3(5).

Olken, B. A. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. *Journal of Political Economy*, 115(2):200–249.

Olson, M. (1971). *The logic of collective action: Public goods and the theory of groups, with a new preface and appendix*. Harvard University Press.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.

Ostrom, E. (1999). Coping with tragedies of the commons. *Annual Review of Political Science*, 2(1):493–535.

Proença, P. F. and Simoes, P. (2020). Taco: Trash annotations in context for litter detection. *arXiv preprint arXiv:2003.06975*.

Raffler, P., Posner, D. N., and Parkerson, D. (2019). The weakness of bottom-up accountability: Experimental evidence from the Ugandan health sector. *Innovations for Poverty Action Working Paper*.

Ramos, J. and Torgler, B. (2012). Are academics messy? Testing the broken windows theory with a field experiment in the work environment. *Review of Law and Economics*, 8(3):563–577.

Rangoni, R. and Jager, A. (2017). Social dynamics of littering and adaptive cleaning strategies explored using agent-based modelling. *The Journal of Artificial Societies and Social Simulation*, 20(2):1.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788.

Sagebiel, J., Karok, L., Grund, J., and Rommel, J. (2020). Clean environments as a social norm: A field experiment on cigarette littering. *Environmental Research Communications*, 2(9):091002.

Saguin, K. (2018). Why the poor do not benefit from community-driven development: Lessons from participatory budgeting. *World Development*, 112:220–232.

Sahin, S. G., Eckel, C., and Komai, M. (2015). An experimental study of leadership institutions in collective action games. *Journal of the Economic Science Association*, 1:100–113.

Schultz, P. W. (1999). Changing behavior with normative feedback interventions: A field experiment on curbside recycling. *Basic and Applied Social Psychology*, 21(1):25–36.

Sheely, R. (2013). Maintaining local public goods: Evidence from rural Kenya. In *CID Working Papers 273, Center for International Development at Harvard University.*

Suel, E., Polak, J. W., Bennett, J. E., and Ezzati, M. (2019). Measuring social, environmental and health inequalities using deep learning and street imagery. *Scientific Reports*, 9(1):6229.

Tanyanyiwa, V. I. (2015). Motivational factors influencing littering in Harare's Central Business District (CBD), Zimbabwe. *IOSR Journal of Human and Social Sciences*, 20(2):58–65.

Terven, J. and Cordova-Esparza, D. (2023). A comprehensive review of YOLO: From YOLOv1 and beyond. *arXiv preprint arXiv:2304.00501.*

Thöni, C. and Volk, S. (2018). Conditional cooperation: Review and refinement. *Economics Letters*, 171:37–40.

Torgler, B., Frey, B. S., and Wilson, C. (2009). Environmental and pro-social norms: Evidence on littering. *The BE Journal of Economic Analysis and Policy*, 9(1).

Van der Windt, P. and Mvukiyehe, E. (2020). Assessing the longer term impact of community-driven development programs: Evidence from a field experiment in the Democratic Republic of Congo. *World Bank Policy Research Working Paper*, (9140).

Willer, R. (2009). Groups reward individual sacrifice: The status solution to the collective action problem. *American Sociological Review*, 74(1):23–43.

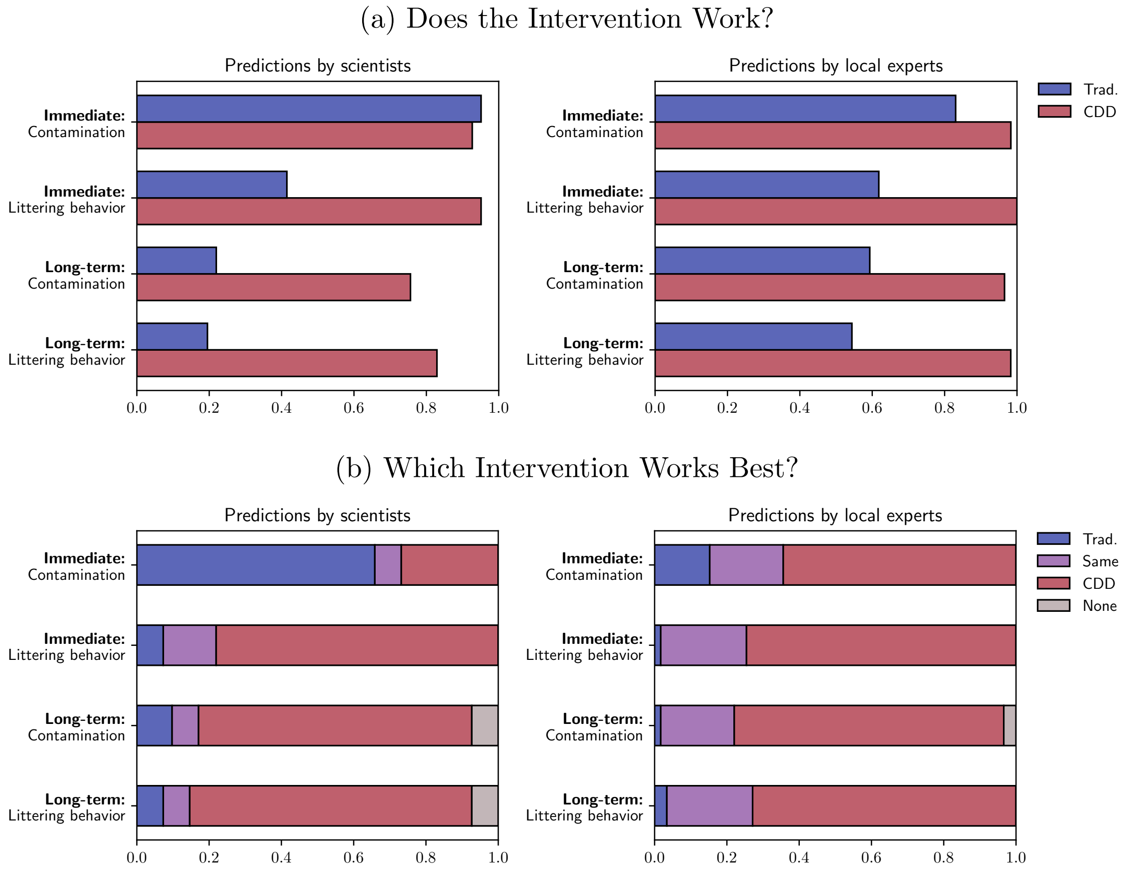
Woolcock, M. et al. (2001). The place of social capital in understanding social and economic outcomes. *Canadian Journal of Policy Research*, 2(1):11–17.

World Bank (2022). Community and local development. Available at: `https://www.worldbank.org/en/topic/communitydrivendevelopment`. Last accessed: 2023-10-26.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., and Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1):2583.

A Appendix

A1 Additional Results



(a) Does the Intervention Work?

(b) Which Intervention Works Best?

Figure A1: Pre-Survey Results

Illustration based on a prediction survey with 41 social scientists and 59 local experts. The upper figure shows the percentage of respondents who expect each intervention to have a positive effect. The lower figure shows the share of respondents indicating that a particular intervention worked best.

Table A1: Community Activities: Survey Answers

	Immediate Effects			*Long-term Effects*		
	Control	T1: Trad.	T2: CDD	Control	T1: Trad.	T2: CDD
Activities observed						
Community meeting	0.29	0.35	0.70***	0.24	0.21	0.40***
Session or workshop	0.06	0.13**	0.48***	0.01	0.02	0.05*
Cleaning	0.26	0.55***	0.74***	0.36	0.45	0.59***
None	0.55	0.38***	0.15***	0.57	0.50	0.31***
Activities participated						
Community meeting	0.16	0.15	0.42***	0.18	0.16	0.27**
Session or workshop	0.03	0.06	0.33***	0.01	0.01	0.04
Cleaning	0.21	0.34***	0.51***	0.30	0.36	0.47***
None	0.69	0.62	0.43***	0.64	0.61	0.47***
Perception						
Level of activities (sd)	0.00	0.16	0.81***	0.00	-0.06	0.20**
Waste management organization (sd)	0.00	0.25***	0.59***	0.00	0.00	0.22***
Frequency waste truck	1.86	2.44	2.26	1.79	2.19	2.20
Frequency waste truck usage	1.64	2.20	2.07	1.66	2.04	1.95
Frequency community cleaning	0.96	1.18	1.40	0.82	0.71	0.89

Sample sizes are n=2066 for the estimation of immediate effects and n=1832 for long-termm effects. Missings are imputed using the mean value per treatment group. The displayed values are sample means per group. The stars indicate the p-value of a joint F-test that each treatment dummy coefficient is equal to 0. Standad errors were clustered at the community level, controls are baseline education, sex, age and strata fixed effects. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$

Table A2: Raw Contamination Results

	Immediate effects				*Long-term effects*			
	T1: Trad.	T2: CDD	T2 - T1	N	T1: Trad.	T2: CDD	T2 - T1	N
Photo trash detection								
Kernel approach	-0.536*** (0.092)	-0.384*** (0.092)	0.153** (0.076)	60709	-0.067 (0.055)	-0.103** (0.052)	-0.037 (0.050)	65673
Raster approach	-0.511*** (0.100)	-0.331*** (0.112)	0.180** (0.090)	10740	-0.071 (0.062)	-0.106* (0.058)	-0.035 (0.056)	10740
Raw averages	-0.561*** (0.084)	-0.427*** (0.085)	0.133 (0.086)	120	-0.089 (0.058)	-0.126** (0.058)	-0.037 (0.059)	120
Enumerator assessments								
Kernel approach	-0.283*** (0.056)	-0.234*** (0.054)	0.049 (0.059)	12216	-0.016 (0.071)	-0.019 (0.058)	-0.003 (0.077)	13163
Raster approach	-0.263*** (0.060)	-0.202*** (0.058)	0.061 (0.061)	4272	0.002 (0.071)	-0.002 (0.055)	-0.004 (0.076)	4272
Raw averages	-0.311*** (0.063)	-0.266*** (0.063)	0.046 (0.064)	120	-0.030 (0.066)	-0.030 (0.066)	0.000 (0.066)	120

Outcomes refer to the number of detected trash items per image in the upper panel and to enumerator assessment scores (1-4) in the lower panel. Controls include contamination at baseline and strata fixed effects. Standard errors are clustered at the community level for the kernel and the raster approach. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$

Table A3: Contamination Results with Coder Fixed Effects

	Immediate effects				*Long-term effects*			
	T1: Trad.	T2: CDD	T2 - T1	N	T1: Trad.	T2: CDD	T2 - T1	N
Photo trash detection								
Kernel approach	-0.893***	-0.689***	0.204**	60709	-0.193**	-0.232**	-0.039	65673
	(0.135)	(0.138)	(0.103)		(0.096)	(0.099)	(0.097)	
Raster approach	-0.968***	-0.737***	0.231**	10740	-0.238**	-0.243**	-0.005	10740
	(0.154)	(0.140)	(0.115)		(0.117)	(0.113)	(0.114)	
Raw averages	-0.923***	-0.641***	0.282*	120	-0.250*	-0.327**	-0.076	120
	(0.157)	(0.158)	(0.159)		(0.147)	(0.151)	(0.152)	
Enumerator assessments								
Kernel approach	-1.220***	-0.949***	0.272	12216	-0.403***	-0.391***	0.013	13163
	(0.155)	(0.185)	(0.181)		(0.150)	(0.149)	(0.159)	
Raster approach	-1.124***	-0.884***	0.240	4272	-0.375**	-0.378**	-0.003	4272
	(0.160)	(0.189)	(0.177)		(0.154)	(0.165)	(0.182)	
Raw averages	-1.157***	-1.004***	0.153	120	-0.364**	-0.452***	-0.088	120
	(0.209)	(0.208)	(0.211)		(0.165)	(0.170)	(0.170)	

Results reported in standard deviations at the community level. Controls include contamination at baseline, coder fixed effects, and strata fixed effects. Standard errors are clustered at the community level for the kernel and the raster approach. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$

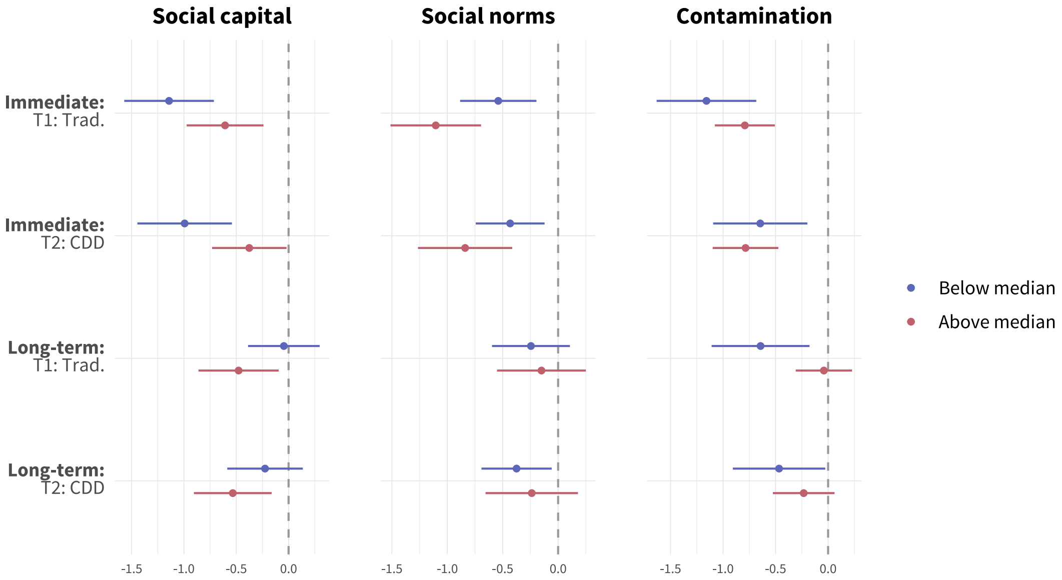


Figure A2: Effect Heterogeneity by Social Capital, Social Norms, and Contamination

The outcome variable is standardized trash counts per image. Social capital refers to an index of networks (strong ties), trust, organizations, and voluntary work (sum of standardized variables); social norms is an index of the share of villagers believed to egage in littering, believed to disapprove of littering, and believed to punish littering (sum of standardized variables); and contamination is the baseline contamination level, measured as standardized trash counts. Heterogeneity analyses are conducted at the community level.

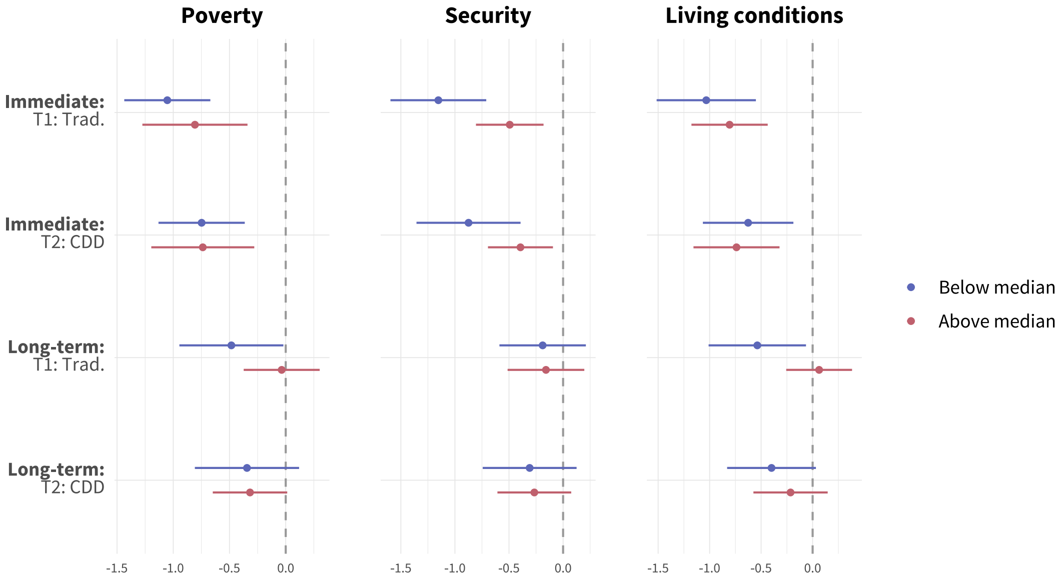


Figure A3: Effect Heterogeneity by Poverty, Security, and Living Conditions

The outcome variable is standardized trash counts per image. Hetereogeneity analyses are conducted at the community level.

Table A4: Coefficient Depletion Rates: Models without Fixed Effects

	Absolute depletion			*Relative depletion*		
	T1: Trad.	T2: CDD	T1 - T2	T1: Trad.	T2: CDD	T1 - T2
Photo trash detection						
Raster approach	0.593	0.271	0.322 ***	0.816	0.575	0.241
Kernel approach	0.627	0.341	0.285 ***	0.829	0.631	0.198
Raw averages	0.617	0.356	0.261 **	0.779	0.590	0.189
Enumerator assessments						
Raster approach	0.810	0.610	0.200	1.008	0.991	0.018
Kernel approach	0.885	0.714	0.171	0.949	0.926	0.023
Raw averages	0.911	0.765	0.146	0.911	0.896	0.015

Absolute depletion indicates the difference between short-term and long-term effects in standard deviations. Relative depletion indicates the difference between short-term and long-term effects as a percentage value of short-term effect. For linear differences, the p-values were obtained with a t-test. For nonlinear differences, the p-values were obtained with the delta method. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$

Table A5: Balance at Baseline for Additional Survey Variables

	Control	T1: Trad.	T2: CDD	P-value	N
Pay for cleaning, me (log)	0.352	0.374	0.315	0.889	2420
Pay for cleaning, others (log)	0.454	0.344	0.341	0.385	2417
Bothered by litter (1-5)	3.894	3.849	3.982	0.511	2420
Littering, me (%)	0.142	0.163	0.141	0.532	2421
Littering is bad, me (1-5)	4.574	4.548	4.635	0.307	2420
Punish littering, me (%)	0.394	0.343	0.324	0.035	2421
Living conditions (1-5)	3.267	3.215	3.165	0.081	2420
Security (1-5)	4.133	4.096	4.141	0.441	2421
Trust comm. leaders (1-5)	3.318	3.150	3.304	0.151	2416
Trust municipal gov. (1-5)	2.739	2.718	2.759	0.912	2415
Trust central gov. (1-5)	3.386	3.215	3.276	0.196	2413

The last row indicates the p-value of a joint F-test that each treatment dummy coefficient is equal to 0. Standard errors are clustered at the community level.

Table A6: Survey Results for Contamination and Waste Disposal

	Immediate effects			*Long-term effects*		
	T1: Trad.	T2: CDD	T2 - T1	T1: Trad.	T2: CDD	T2 - T1
Perceived cleanliness (sd)	0.158*	0.163**	0.005	0.173**	0.182**	0.009
	(0.085)	(0.083)	(0.076)	(0.079)	(0.082)	(0.078)
Bothered by litter (sd)	-0.069	-0.008	0.061	-0.023	0.073	0.096
	(0.093)	(0.092)	(0.087)	(0.086)	(0.078)	(0.082)
Appropriate disposal (%)	0.040	0.137***	0.097**	-0.005	0.067*	0.072*
	(0.044)	(0.041)	(0.046)	(0.042)	(0.039)	(0.040)
Recycling (%)	0.083**	0.112***	0.029	-0.026	-0.001	0.025
	(0.041)	(0.037)	(0.029)	(0.040)	(0.040)	(0.040)
Recycling items (nr)	0.093	0.268**	0.175	-0.162	0.055	0.216**
	(0.113)	(0.125)	(0.111)	(0.098)	(0.102)	(0.095)
Pay for cleaning, me (log)	-0.109**	-0.109*	0.001	0.060	0.028	-0.031
	(0.056)	(0.061)	(0.050)	(0.068)	(0.061)	(0.062)
Pay for cleaning, others (log)	-0.085	-0.081	0.004	0.024	-0.026	-0.050
	(0.062)	(0.069)	(0.053)	(0.058)	(0.053)	(0.047)

Sample sizes are n = 2066 for the estimation of immediate effects and n = 1832 for long-term effects. Controls include strata fixed effects, sex, age, education (dummies), and the baseline value for the respective outcome. Note that no baseline values are available for recycling outcomes. Standard errors are clustered at the community level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table A7: Survey Results for Self-Reported Behaviors and Social Norms

	Immediate effects			*Long-term effects*		
	T1: Trad.	T2: CDD	T2 - T1	T1: Trad.	T2: CDD	T2 - T1
Littering, me (%)	-0.030**	-0.016	0.014	-0.011	-0.006	0.005
	(0.015)	(0.015)	(0.013)	(0.016)	(0.014)	(0.016)
Littering is bad, me (sd)	0.041	0.043	0.002	-0.011	0.022	0.034
	(0.041)	(0.044)	(0.039)	(0.054)	(0.044)	(0.052)
Punish littering, me (%)	-0.027	-0.002	0.024	0.005	-0.014	-0.020
	(0.026)	(0.026)	(0.026)	(0.031)	(0.029)	(0.030)
Littering, others (%)	-0.065**	-0.104***	-0.039	0.011	-0.019	-0.030
	(0.028)	(0.026)	(0.026)	(0.019)	(0.019)	(0.021)
Littering is bad, others (%)	-0.015	-0.037**	-0.022	-0.001	0.004	0.005
	(0.020)	(0.017)	(0.020)	(0.018)	(0.018)	(0.016)
Punish littering, others (%)	0.027	0.003	-0.024	0.017	0.036*	0.019
	(0.028)	(0.025)	(0.027)	(0.022)	(0.020)	(0.022)

Sample sizes are n = 2066 for the estimation of immediate effects and n = 1832 for long-term effects. Controls include strata fixed effects, sex, age, education (dummies), and the baseline value for the respective outcome. Standard errors are clustered at the community level. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$

Table A8: Survey Results for Other Outcomes

	Immediate effects			*Long-term effects*		
	T1: Trad.	T2: CDD	T2 - T1	T1: Trad.	T2: CDD	T2 - T1
Living conditions (sd)	-0.141**	-0.094	0.046	-0.075	0.029	0.103
	(0.059)	(0.066)	(0.068)	(0.059)	(0.053)	(0.063)
Security (sd)	-0.014	0.053	0.068	-0.035	-0.074	-0.040
	(0.081)	(0.086)	(0.085)	(0.097)	(0.088)	(0.109)
Trust comm. leaders (sd)	0.035	0.104	0.069	-0.096	0.056	0.152**
	(0.103)	(0.093)	(0.106)	(0.076)	(0.065)	(0.064)
Trust municipal gov. (sd)	0.024	0.097	0.072	0.012	-0.061	-0.074
	(0.085)	(0.084)	(0.081)	(0.088)	(0.071)	(0.082)
Trust central gov. (sd)	0.045	0.130	0.085	-0.060	-0.044	0.016
	(0.114)	(0.110)	(0.107)	(0.073)	(0.064)	(0.070)

Sample sizes are n = 2066 for the estimation of immediate effects and n = 1832 for long-term effects. Controls include strata fixed effects, sex, age, education (dummies), and the baseline value for the respective outcome. Standard errors are clustered at the community level. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Table A9: Attrition by Treatment Group

	Control	T1: Trad.	T2: CDD	P-value	N
Attrition Midline	0.160	0.151	0.128	0.423	2421
Attrition Endline	0.228	0.275	0.228	0.131	2421

The last row indicates the p-value of a joint F-test that each treatment dummy coefficient is equal to 0. Standard errors are clustered at the community level.

Table A10: Attriter Characteristics at Midline by Treatment Group

	Control	T1: Trad.	T2: CDD	P-value	N
Sociodemographics					
Female	0.706	0.664	0.670	0.936	355
Age	47.360	40.899	41.810	0.002	355
Education	2.147	2.361	2.360	0.464	355
Poverty (1-5)	3.000	3.104	3.010	0.819	342
Community size	305.463	279.706	269.530	0.645	355
Contamination and waste disposal					
Perceived cleanliness (1-5)	3.154	3.008	3.290	0.102	355
Appropriate disposal (%)	0.353	0.445	0.460	0.164	355
Social norms					
Littering (%)	0.606	0.605	0.514	0.083	355
Littering is bad (%)	0.707	0.733	0.662	0.141	354
Punish littering (%)	0.571	0.582	0.493	0.016	355
Social capital					
Strong ties (%)	0.310	0.461	0.311	0.110	354
Weak ties (%)	0.687	0.756	0.654	0.088	345
Trust (1-5)	3.463	3.370	3.540	0.578	355
Organizations (%)	0.110	0.109	0.110	0.911	355
Voluntary work (%)	0.154	0.254	0.300	0.071	354
Altruism (%)	0.405	0.505	0.413	0.066	355

The first three columns represent attriter group means for each treatment. The last row indicates the p-value of a joint F-test that each treatment dummy coefficient is equal to 0. Education refers to highest completed degree: None = 1, incomplete primary = 2, complete primary = 3, high school degree = 4), technical = 5, and university degree = 6. Standard errors are clustered at the community level.

Table A11: Attriter Characteristics at Endline by Treatment Group

	Control	T1: Trad.	T2: CDD	P-value	N
Sociodemographics					
Female	0.660	0.724	0.607	0.065	589
Age	40.984	39.654	41.421	0.440	588
Education	2.546	2.599	2.522	0.687	589
Poverty (1-5)	3.247	3.118	3.034	0.469	572
Community size	300.289	288.507	282.854	0.620	589
Contamination and waste disposal					
Perceived cleanliness (1-5)	3.093	3.115	3.129	0.999	589
Appropriate disposal (%)	0.330	0.498	0.478	0.112	589
Social norms					
Littering (%)	0.618	0.612	0.548	0.013	589
Littering is bad (%)	0.730	0.715	0.678	0.163	589
Punish littering (%)	0.584	0.582	0.534	0.100	588
Social capital					
Strong ties (%)	0.333	0.408	0.307	0.423	589
Weak ties (%)	0.697	0.711	0.638	0.530	578
Trust (1-5)	3.392	3.461	3.567	0.601	589
Organizations (%)	0.144	0.143	0.135	0.923	589
Voluntary work (%)	0.201	0.194	0.316	0.060	587
Altruism (%)	0.467	0.456	0.430	0.628	589

The first three columns represent attriter group means for each treatment. The last row indicates the p-value of a joint F-test that each treatment dummy coefficient is equal to 0. Education refers to highest completed degree: None = 1, incomplete primary = 2, complete primary = 3, high school degree = 4), technical = 5, and university degree = 6. Standard errors are clustered at the community level.

A2 Supplementary Information on Data and Measurement Instruments

Table A12: Coding of the Survey Questions

Variable	*Survey question*	*Possible answers*	*Computation*
Contamination and waste disposal			
Perceived cleanliness	How do you evaluate the garbage contamination situation in your community?	Scale from 1 (Very clean) to 5 (Very dirty)	Standardized
Bothered by litter	Personally, how bothered are you by the trash in your community?	Scale from 1 (Not at all) to 5 (Very much)	Standardized
Appropriate disposal	In the past month, how has your household gotten rid of trash?	1: Trash truck; 2: Deposit; 3: Bury it; 4: Burn it; 5: Informal deposit, street	Percentage that used the trash truck or formal deposits
Recycling	In the past month, has your household separated any trash for recycling?	0: None of the below; 1: At least one of the below	Percentage
Recycling items	What types of garbage have been recycled?	0: None; 1: Plastic; 2: Glass; 3: Paper; 4: Organic waste	Number of different items
Pay for cleaning, me	Imagine if a service was hired in your community to clean the streets. How much would you be willing to contribute per month?	Decimal	Log
Pay for cleaning, others	On average, how much do you think a person in your community would be willing to contribute?	Decimal	Log
Self-Reported behaviors and social norms			

Table A12: Coding of the Survey Questions

Variable	*Survey question*	*Possible answers*	*Computation*
Littering, me	Being very, very honest, in the last month, have you ever thrown trash in the street?	0: No; 1: Yes	Percentage
Littering is bad, me	In your personal opinion, is it bad to litter on the street?	Scale from 1 (Not at all bad) to 5 (Very bad)	Standardized
Punish littering, me	If you observed someone in your community throwing trash in the street, what would you do?	0: Nothing, I do not want to get involved / it does not seem serious to me; 1: React with disapproval	Percentage of people reacting with disapproval
Littering, others	Out of every 10 people in your community, how many do you think have thrown trash in the street in the last month?	Integer	Percentage
Littering is bad, others	Out of every 10 people in your community, how many do you think believe it is wrong to litter in the street?	Integer	Percentage
Punish littering, others	Out of every 10 people in your community, how many do you think would react with a gesture of disapproval to a person throwing trash in the street?	Integer	Percentage
Social capital			
Strong ties	Q1: Approximately how many people live in your community?; Q2: Of these people, how many persons are close acquaintances (family, friends)?	Integers	Percentage (Q2/Q1)

Table A12: Coding of the Survey Questions

Variable	*Survey question*	*Possible answers*	*Computation*
Weak ties	Q1: Approximately how many people live in your community?; Q2: Of these people, how many persons are close acquaintances (family, friends)?; Q3: Of these people, how many persons are casual acquaintances?	Integers	Percentage ((Q2+Q3)/Q1)
Trust	Compared to other people, do you trust people in your community more or less?	Scale from 1 (Much less) to 5 (Much more)	Standardized
Organizations	Do you belong to one or more community organizations or groups (e.g. ADESCO, youth organization, collectives, etc.)?	0: No; 1: Yes	Percentage
Voluntary work	In the past month, have you participated in any type of volunteer work for the community?	0: No; 1: Yes	Percentage
Other outcomes			
Living conditions	Compared to other communities, how do you evaluate the living conditions in your community?	Scale from 1 (Very bad) to 5 (Very good)	Standardized
Security	Compared to other communities, how do you evaluate the security situation in your community?	Scale from 1 (Very safe) to 5 (Very dangerous)	Standardized
Trust community leaders	How much do you trust your community leaders?	Scale from 1 (Not at all) to 5 (Very much)	Standardized
Trust municipal government	How much do you trust municipal officials?	Scale from 1 (Not at all) to 5 (Very much)	Standardized

Table A12: Coding of the Survey Questions

Variable	*Survey question*	*Possible answers*	*Computation*
Trust central government	How much do you trust central government officials?	Scale from 1 (Not at all) to 5 (Very much)	Standardized
Altruism	There will be a lottery for 100 USD among the participants. The winner will have to decide how much of this money to keep and how much to donate to a family in need in the department (photos of the delivery would be sent).	Integers	Percentage of how much was donated
Socio-demographic variables			
Female	Gender	0: Male; 1: Female	
Age	Age	Integer	
Education	Highest level of education	1: None; 2: Incomplete primary; 3: Complete primary; 4: High school degree; 5: Technical; 6: University degree	Dummies for each level
Poverty	What is the family's economic situation like? The family's poverty level was recorded by the enumerators based on pictures of potential housing conditions.	Scale of 1 (Not poor) to 5 (Very poor)	Standardized
Community size	Approximately how many people live in your community?	Integer	
Waste management activities			

Table A12: Coding of the Survey Questions

Variable	*Survey question*	*Possible answers*	*Computation*
Activities observed	In the last 4 months, have you heard of any activity related to the issue of garbage in your community?	Community meeting; Session or workshop; Cleaning; None	Dummies for each activity
Activities participated	In the last 4 months, have you participated in any activity related to the issue of garbage in your community?	Community meeting; Session or workshop; Cleaning; None	Dummies for each activity
Level of activities	In the last 4 months, do you think there were more or fewer activities than before regarding the issue of garbage in your community?	Scale from 1 (Much less) to 5 (Much more)	Standardized
Waste management organization	In your opinion, how organized is your community in relation to garbage management?	Scale from 1 (Not at all organized) to 5 (Perfectly organized)	Standardized
Frequency waste truck	In the last 4 months, how often has a toilet train arrived in your community?	1: Never; 2: Every 2 months; 3: Every month; 4: Every 2 weeks; 5: Every week; 6: Twice a week; 7: Every day	Frequency per month
Frequency waste truck usage	In the last 4 months, how often have you used the garbage train to dispose of your garbage?	1: Never; 2: Every 2 months; 3: Every month; 4: Every 2 weeks; 5: Every week; 6: Twice a week; 7: Every day	Frequency per month
Frequency community cleaning	In the last 4 months, how often has your community been cleaned?	1: Never; 2: Every 2 months; 3: Every month; 4: Every 2 weeks; 5: Every week; 6: Twice a week; 7: Every day	Frequency per month

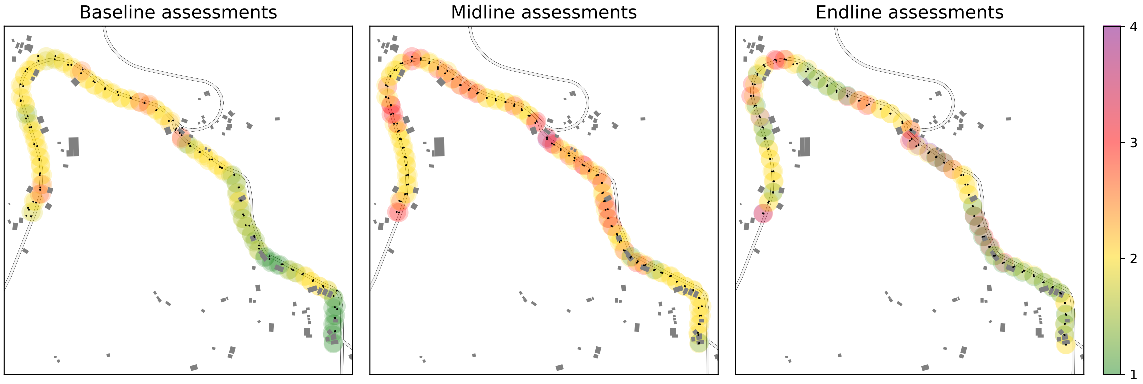


Figure A4: Illustration of Kernel Approach for Subjective Enumerator Assessments

Black dots represent assessment locations. Circle color corresponds to contamination level: 1 = very clean, 4 = very dirty. Baseline values are imputed based on circles around each midline and endline assessment respectively. A triangular kernel is used to give higher weights to closer assessments. Circle radius is 25m. Baseline map is shown with respect to the midline assessment and would be slightly different for the endline assessment.

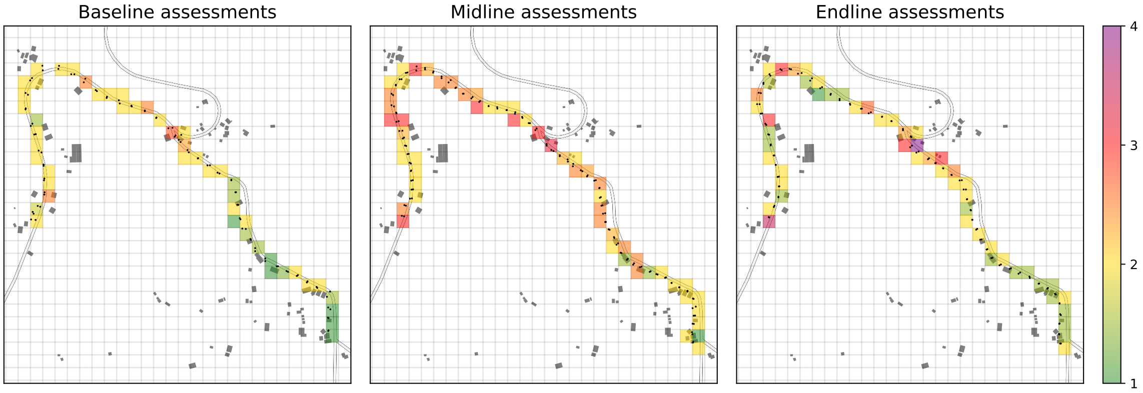


Figure A5: Illustration of Raster Approach for Subjective Enumerator Assessments

Black dots represent assessment locations. Cell color corresponds to contamination level: 1 = very clean, 4 = very dirty. Resolution of the raster is 0.0003 degrees (approx. 33m).

A3 Literature Review

Table A13: Literature Review on Community-Driven Development

Study	*Study type*	*Description and results*
Meta studies of community-driven development programs		
Casey 2018	Meta study on evolution of CDD and CDD RCTs.	A synthesis of seven CDD RCTs shows that CDD effectively delivers public goods and some economic benefits at a low cost in challenging environments. However, it does not seem to lead to lasting transformations in local decision-making or empowerment of the poor. This raises the question of how much participation is necessary to preserve the benefits of decentralization while minimizing the time costs imposed on impoverished communities.
Mansuri and Rao 2012	Meta study proposing general concept of CDD based on literature from different fields.	The report discusses the history of participatory development and presents a framework for understanding participatory development, emphasizing the concept of "civil society failure" and its interaction with government and market failures. It is based on literature from anthropology, economics, political science and sociology. Evidence on key development outcomes, public service delivery and quality, but also on issues related to CDD is reviewed. The report also discusses World Bank-funded projects, emphasizing the importance of local context, as well as effective monitoring and evaluation for successful outcomes.
Evaluations of community-driven development programs		
Arcand 2008	IV study with panel data on 71 villages with 756 households in Senegal.	This paper investigates the impact of a national CDD program on access to basic services, household expenditures, and child wellbeing. The program had a positive effect on villagers' access to clean water and health services, as well as on child malnutrition. Completed income-generating agricultural infrastructure projects and improved primary education significantly increased household expenditures per capita, while health and hydraulic projects did not.

Table A13: Literature Review on Community-Driven Development

Study	*Study type*	*Description and results*
Avdeenko and Gilligan 2015	RCT with 576 households in 24 communities, and 475 lab-in-the-field subjects.	The intervention had no impact on networks and social norms, but it increased people's involvement in civic activities and local governance. Therefore, the authors attribute the increase in citizen participation not to the growth of social capital, but to the greater openness of the local governments.
Beath et al. 2013	RCT with 500 villages in Afghanistan	The RCT examines the impact of a CDD program that requires female participation on several outcomes related to women's empowerment. Positive effects on women's participation in economic, social, and political activities are reported. However, no impacts on gender roles or family decision-making are found.
Casey et al. 2012	RCT with 2,832 households in 236 villages in the Republic of Sierra Leone.	The study evaluates a CDD program aiming to make local institutions more democratic and egalitarian by imposing participation requirements for marginalized groups. The program had positive short-term effects on local public services and economic outcomes. However, it did not result in sustained impacts on collective action, decision-making, or the involvement of marginalized groups, indicating that the intervention did not durably reshape local institutions.
Desai and Olofsgård 2019	RCT combined with behavioral experiment with 80 villages in India.	The "self-help" groups established in treatment villages significantly improved people's access to and the quality of certain public goods, especially water, due to better information through the groups, stronger community engagement and reduced coordination costs. The behavioral experiment 4 years after the RCT revealed that cooperative norms are stronger in villages that had self-help groups.
Fearon et al. 2009	RCT with 83 communities in Liberia	The study evaluates the impact of a community-driven (post-war) reconstruction project on social cohesion, as measured by an anonymous public goods game. Contributions were significantly higher in the treated communities, with a 9 percent increase in funds raised for a community-selected public good.

Table A13: Literature Review on Community-Driven Development

Study	*Study type*	*Description and results*
Humphreys et al. 2019	RCT with 1,250 communities in the Congo	The study evaluates the impact of a community-driven reconstruction program on democratic governance. Behavior in an unconditional cash transfer program is used to assess whether the intervention had an impact on elite capture. No effects are found.
Labonne and Chase 2011	DiD with 2,100 households in 135 communities in the Philippines.	Using difference-in-differences (DiD) and propensity score matching, the study evaluates a CDD program where communities competed for grants for infrastructure investments. The program increased the participation in village meetings and the frequency of interactions between local officials and village officials, but had a negative impact on collective action.
Nguyen and Rieger 2017	RDD with 1,300 communes in Morocco	The study assesses the impact of a CDD initiative on social capital, employing a regression discontinuity design (RDD) based on the program's poverty selection threshold. The program increased contributions in a public goods game, but had no effect on altruism and a negative effect on trust.
Saguin 2018	DiD based on surveys in 16 municipalities in the Philippines	The "KALAHI-CIDSS" CDD program was found to increase the incomes of poor households. However, it did not improve outcomes such as solidarity and trust. In addition, poor households are underrepresented in village assemblies, with declining participation over time.
Van der Windt and Mvukiyehe 2020	RCT with 1,250 villages in the Republic of Congo.	The study assesses the long-term impact of a CDD initiative 8 years after its launch. The program had a lasting impact on infrastructure quality (e.g., of schools or hospitals), but no effects on other dimensions of service delivery, on economic welfare, and on local institutions (e.g., governance, social cohesion, or female empowerment) were found.
Related studies		

Table A13: Literature Review on Community-Driven Development

Study	*Study type*	*Description and results*
Banerjee et al. 2010	RCT with three interventions in India.	This paper examines if citizen involvement can shape public service provision in education. Three interventions were evaluated: (i) providing information on public school organization, (ii) introducing citizens to a simple monitoring tool for their local school, and (iii) training volunteers to hold reading camps in order to improve literacy knowledge. Information and monitoring did not improve outcomes, but the volunteer-led reading camps did.
Björkman and Svensson 2009	RCT with 50 public dispensaries in Uganda.	The intervention aimed at encouraging community engagement in monitoring health services and holding local health providers accountable for their performance. To this end, community members developed village action plans together with the health care providers. One year after the intervention, treatment communities exhibited greater involvement in monitoring providers, resulting in increased effort from health workers to serve the community as well as significant improvements in healthcare utilization and health outcomes.
Björkman et al. 2017	Follow-up of RCT in Björkman and Svensson (2009).	The authors evaluate the long-run impact (4 years) of the experiment in Björkman and Svensson (2009). Even with minimal follow-up, short-term enhancements in healthcare delivery and health outcomes were sustained over the long run. The results indicate that a lower-cost version of the treatment, which primarily aimed to boost participation without information on staff performance, did not influence the quality of care or health outcomes both in the short and in the in the longer run.

Table A13: Literature Review on Community-Driven Development

Study	*Study type*	*Description and results*
Duflo et al. 2015	RCT with 70 schools in Kenya.	The study evaluates a program in Kenya where parents and the school committees of randomly selected schools received (i) funding or (ii) funding and a short School-Based Management (SBM) empowerment training to hire an additional teacher, outside normal Ministry of Education civil-service channels. Centrally hired civil-service teachers in schools receiving only funding endogenously reduced their effort and captured rents for their families by getting relatives the contract teacher positions. The SBM program cut by half both the reduction in the regular teacher effort in response to the program and the fraction of contract teachers who were relatives of regular teachers.
Olken 2007	RCT in 608 villages in Indonesia.	The paper evaluates different interventions aiming at reducing corruption, measured by missing expenditures, in village road projects in Indonesia. Results show that increased government audits significantly reduce missing expenditures. In contrast, enhancing grassroots monitoring had limited impacts on corruption.
Raffler et al. 2019	RCT with 376 health care centers and 14,609 households in rural Uganda.	The authors evaluate a large-scale information intervention aiming to improve bottom-up monitoring of health service delivery. The study finds only modest positive effects of citizen monitoring on service quality and patient satisfaction, and no effects on utilization and health outcomes such as child mortality.

Table A14: Literature Review on Waste Management

Study	*Study type*	*Description and results*
Evaluations of waste interventions		
Bateson et al. 2013	Field experiment with 620 bicycle riders on Newcastle university campus, UK.	This study tests if displaying images of "watching eyes" causes people to litter less and if a potential effect depends on the cleanliness of the environment. People were more likely to litter in dirty environments, but the watching eyes only had an effect when many people were around, and this effect does not depend on the amount of litter in the environment.
Castaldi et al. 2021	RCT in 8 beach resorts in Italy.	The resorts were randomly assigned to 3 groups: (i) free portable ashtrays, (ii) free portable ashtrays and anti-littering message, and (iii) control. Results show a reduction in daily litter (cigarette butts in sand on day/costumers): -10% to -12% for the ashtray group; -7% to -10% for the ashtray + message group.
Cialdini et al. 1990	5 field experiments in different public spaces with 127–484 observations.	The authors argue that injunctive and descriptive norms must be separated to understand littering behavior since behavior changes only in accordance with the more salient type. In their experiments, they find that littering increases in littered environments, and even more so when someone is observed littering. Conversely, littering decreases when someone is observed littering into a very clean environment. Men are more likely to litter than women across different settings.
Dur and Vollaard 2015	Field experiment with 4,000 households in the Netherlands.	This paper studies littering behavior and free-riding mechanisms related to public services. In a randomly assigned part of residential area, the frequency of cleaning around the garbage containers is drastically reduced from daily cleanups to 2-3 times a week during a 3-month period. Removing the morning cleanup increased the presence of litter in the early afternoon (11% to 27%). Litter accumulation around the garbage disposal increased (from 20% to 75%). Telephone appointments for retrieval of large trash increased, meaning that some people started to clean up more by themselves. The effects persisted at least one month after the treatment ended.

Table A14: Literature Review on Waste Management

Study	*Study type*	*Description and results*
Lewis et al. 2009	Nationwide survey on littering attitude and field study in cinemas in the UK.	The survey revealed personal differences in acceptance and justification of littering depending on the age group, rural/urban living environment, smoker/non-smoker, feeling connected/unconnected to community. Also, missing infrastructure was identified as a cause for littering. In the cinema field study, leaflets with a (i) control message unrelated to littering, (ii) polite anti-littering message, or (iii) direct anti-littering message were distributed, and it was observed how much litter was left behind. People in the control group littered more than people that were politely or directly asked not to.
Liu and Sibley 2004	Field study in a public space in New Zealand with over 3,000 observations.	In a first sub-study, littering attitudes were observed during 3 weeks and the people who disposed of waste (correctly and incorrectly) were interviewed. In the second week, a banner with an anti-littering message was added. People were found to litter less in crowded public spaces compared to less-crowded public places. The banner did not change littering behavior. In a second sub-study, bins and ashtrays were installed, and found to reduce littering by 64% without changing attitudes towards littering.
Nepal et al. 2023	RCT with 75 treatment and 75 control communities in Nepal.	The study evaluates a low-cost treatment to improve municipal solid waste management: Providing information to households and installing waste bins on the streets. Perceived cleanliness in treatment communities increased by 25% at midline (3 months after installation) and 43% at endline (9 months after installation). Giving household waste to collectors increased by 13% at midline and 9% at endline while there was no statistically significant change in at-source waste segregation.

Table A14: Literature Review on Waste Management

Study	*Study type*	*Description and results*
Ramos and Torgler 2012	Field study with 98 observations in Australia.	The authors test the broken-window-theory in a field study, which was conducted over 6 days in university common rooms, alternating between an orderly and a disorderly environment. 59% of participants littered in the disorderly room, compared to 18% in the clean room. Multivariate analysis shows that the disorder variable is always large and statistically significant. Older individuals and senior staff were more likely to litter.
Rangoni and Jager 2017	Simulation in an agent-based model with 100 simulated pedestrians.	The goal of the simulation is to evaluate how social influence may cause a transition from a clean to a littered environment in 3 situations: (i) no trash bins; (ii) trash bins which can get full, and (iii) adding cleaners who can pick up litter and empty bins. For the parameterization of the model, data from a field study is used. The simulations suggest that litter does not grow linearly. Furthermore, a dynamic cleaning regime is cheaper and more effective than pre-determined regimes.
Sagebiel et al. 2020	Field experiment with 200 observations on university benches in Germany.	To test the broken-window theory in the context of littering cigarette butts, two types of environment were prepared: (i) clean environment in which all cigarette butts were removed around the benches; (ii) dirty environment in which 25 cigarette butts were placed around each bench. The authors conclude that increased cleaning effort reduces littering a little, but the effect might be too small to justify additional cleaning costs.
Schultz 1999	Field experiment with 605 residents of single-family dwellings in the US.	The study aims to find out if a plea alone or accompanied by (i) information, (ii) neighbor feedback, or (iii) household feedback increases proper waste disposal. Results show that feedback targeting personal or social norms increased the proportion of people recycling and the amount of recycled materials while not changing the level of contamination through littering. The author argues that a link between norm activation and behavior change exists.

Table A14: Literature Review on Waste Management

Study	*Study type*	*Description and results*
Sheely 2013	RCT with 36 communities in Kenya (based on qualitative study).	This study aims to explain variation in maintaining a clean environment through interactions between government and community institutions. Communities are randomly assigned to 4 experimental groups: (i) collective action to organize cleanups promoted by a local NGO, (ii) collective action and punishment for littering by government chiefs, (iii) collective action and punishment for littering by traditional elders, and (iv) a control group. The author finds that communities with no formal punishment for littering experienced a sustained reduction in littering behavior and an increase in the frequency of public cleanups. Communities in which government administrators or traditional leaders punished littering experienced short-term reductions in littering that were not sustained.
Other waste studies		
Chitotombe 2014	Interviews in Zimbabwe and literature review.	The unavailability of bins, socio-cultural consumption styles in particular related to fast foods, illegal display of posters in the streets, and abandoned motor vehicles are mentioned as problems in Zimbabwe. Anti-littering campaigns have shown little success in the past. The interviews revealed that bins were not used even if available and that communities are reluctant to participate in cleanups. Also, the study shows that language barriers and political inefficiencies impede proper waste management.
Nkwocha and Okeoma 2009	Interviews of 6,000 individuals in 6 geo-political zones in Nigeria.	Littering is very common in Nigeria. Reasons given by respondents included lack of bins or long distances to dumpsites, inefficiency of local authorities in keeping public spaces clean, missing legislation against littering, convenience, and ignorance of the environmental and health consequences of littering. Low levels of education were highly correlated with littering.

Table A14: Literature Review on Waste Management

Study	*Study type*	*Description and results*
Tanyanyiwa 2015	Interviews with residents and street workers in Zimbabwe.	Reasons for the high littering levels are identified as: missing sense of ownership of public areas, the belief that someone else will clean up, and that littering is tolerated. Suggested ways to reduce litter include the provision of dedicated recycling bins, a volunteer environmental police force, and the establishment of a coordinated waste management system.
Torgler et al. 2009	Analysis of over 30,000 respondents of the European Value Survey (EVS).	Using EVS data on basic values and beliefs of people in Europe, the authors find a positive, albeit small, relationship between how people perceive environmental cooperation (public littering) and their voluntary environmental morale.

Bibliography

Abdullah, A., H. Doucouliagos, and E. Manning (2015). Does education reduce income inequality? A meta-regression analysis. *Journal of Economic Surveys 29*(2), 301–316.

Aiken, E., S. Bellue, D. Karlan, C. Udry, and J. E. Blumenstock (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature 603*(7903), 864–870.

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics 106*(4), 979–1014.

Barber, M. and M. Mourshed (2007). *How the world's best-performing schools systems come out on top.* McKinsey & Company.

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

Belfield, C. R., M. Nores, S. Barnett, and L. Schweinhart (2006). The high/scope perry preschool program. Cost–benefit analysis using data from the age-40 followup. *Journal of Human Resources 41*(1), 162–190.

Blumenstock, J., G. Cadamuro, and R. On (2015). Predicting poverty and wealth from mobile phone metadata. *Science 350*(6264), 1073–1076.

Bold, T., D. Filmer, G. Martin, E. Molina, B. Stacy, C. Rockmore, J. Svensson, and W. Wane (2017). Enrollment without learning: Teacher effort, knowledge, and skill in primary schools in Africa. *Journal of Economic Perspectives 31*(4), 185–204.

Brunetti, A., K. Büchel, M. Jakob, B. Jann, C. Kühnhanss, and D. Steffen (2020). Teacher content knowledge in developing countries: Evidence from a math assessment in El Salvador. *University of Bern Social Sciences Working Papers No. 34.* Available at: `https://ideas.repec.org/p/bss/wpaper/34.html`.

Brunetti, A., K. Büchel, M. Jakob, B. Jann, and D. Steffen (2023). Inadequate teacher content knowledge and what could be done about it: Evidence from El Salvador. *Journal of Development Effectiveness*, 1–24.

Büchel, K., M. Jakob, C. Kühnhanss, D. Steffen, and A. Brunetti (2022). The relative effectiveness of teachers and learning software: Evidence from a field experiment in El Salvador. *Journal of Labor Economics 40*(3), 737–777.

Burke, M., A. Driscoll, D. B. Lobell, and S. Ermon (2021). Using satellite imagery to understand and promote sustainable development. *Science 371*(6535), eabe8628.

Cutler, D. M. and A. Lleras-Muney (2006). Education and health: Evaluating theories and evidence. Technical report, National Bureau of Economic Research.

Duflo, E., R. Glennerster, and M. Kremer (2008). Using randomization in development economics research: A toolkit. *Handbook of Development Economics 4*, 3895–3962.

Ganimian, A. J. and R. J. Murnane (2016). Improving education in developing countries: Lessons from rigorous impact evaluations. *Review of Educational Research 86*(3), 719–755.

Gebru, T., J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei (2017). Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences 114*(50), 13108–13113.

Gertler, P. J., S. Martinez, P. Premand, L. B. Rawlings, and C. M. Vermeersch (2016). *Impact evaluation in practice.* Inter-American Development Bank and World Bank.

Glewwe, P. and K. Muralidharan (2016). Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In *Handbook of the Economics of Education*, Volume 5, pp. 653–743. Elsevier.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review 30*(3), 466–479.

Hanushek, E. A. and L. Woessmann (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth 17*(4), 267–321.

Harris, M., J. Marti, H. Watt, Y. Bhatti, J. Macinko, and A. W. Darzi (2017). Explicit bias toward high-income country research: A randomized, blinded, crossover experiment of English clinicians. *Health Affairs 36*(11), 1997–2004.

Head, A., M. Manguin, N. Tran, and J. E. Blumenstock (2017). Can human development be measured with satellite imagery? *ICTD 17*, 16–19.

Jakob, M., K. Büchel, D. Steffen, and A. Brunetti (2023). Participatory teaching improves learning outcomes: Evidence from a field experiment in Tanzania. *University of Bern Social Sciences Working Papers No. 48.* Available at: `https://econpapers.repec.org/paper/ubedpvwib/dp2310.htm`.

Jakob, M. and B. Combet (2020). Educational aspirations and decision-making in a context of poverty. A test of rational choice models in El Salvador. *Research in Social Stratification and Mobility 69*, 100545.

Jakob, M. S. and S. Heinrich (2023). Measuring human capital with social media data and machine learning.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016). Combining satellite imagery and machine learning to predict poverty. *Science 353*(6301), 790–794.

Kremer, M., C. Brannen, and R. Glennerster (2013). The challenge of education and learning in the developing world. *Science 340*(6130), 297–300.

Lochner, L. and E. Moretti (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review 94*(1), 155–189.

McEwan, P. J. (2015). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research 85*(3), 353–394.

Metzler, J. and L. 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.

Milligan, K., E. Moretti, and P. Oreopoulos (2004). Does education improve citizenship? Evidence from the United States and the United Kingdom. *Journal of public Economics 88*(9-10), 1667–1695.

Oreopoulos, P. (2007). Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling. *Journal of Public Economics 91*(11-12), 2213–2229.

Peet, E. D., G. Fink, and W. Fawzi (2015). Returns to education in developing countries: Evidence from the living standards and measurement study surveys. *Economics of Education Review 49*, 69–90.

Plancikova, D., P. Duric, and F. O'May (2021). High-income countries remain overrepresented in highly ranked public health journals: a descriptive analysis of research settings and authorship affiliations. *Critical Public Health 31*(4), 487–493.

Popova, A., D. K. Evans, M. E. Breeding, and V. Arancibia (2018). Teacher professional development around the world: The gap between evidence and practice. World Bank Policy Research Working Paper (8572).

Ratledge, N., G. Cadamuro, B. De la Cuesta, M. Stigler, and M. Burke (2021). Using satellite imagery and machine learning to estimate the livelihood impact of electricity access. Technical report, National Bureau of Economic Research.

Sheehan, E., C. Meng, M. Tan, B. Uzkent, N. Jean, M. Burke, D. Lobell, and S. Ermon (2019). Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2698–2706.

Sinha, S., R. Banerji, and W. Wadhwa (2016). *Teacher performance in Bihar, India: Implications for education.* The World Bank.

Snilstveit, B., J. Stevenson, D. Phillips, M. Vojtkova, E. Gallagher, T. Schmidt, H. Jobse, M. Geelen, M. G. Pastorello, and J. Eyers (2015). Interventions for improving learning outcomes and access to education in low-and middle-income countries: A systematic review. *3ie Final Review. International Initiative For Impact Evaluation, London.*

Treisman, D. (2000). The causes of corruption: A cross-national study. *Journal of Public Economics 76*(3), 399–457.

United Nations (2015). Transforming our world: The 2030 agenda for sustainable development.

Vogl, T. S. et al. (2012). Education and health in developing economies. *Encyclopedia of Health Economics 1453*, 246–249.

Wantchekon, L., M. Klašnja, and N. Novta (2015). Education and human capital externalities: Evidence from colonial Benin. *The Quarterly Journal of Economics 130*(2), 703–757.

World Bank (2018). *World development report 2018: Learning to realize education's promise.* World Bank.

Yeh, C., A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications 11*(1), 2583.

Selbstständigkeitserklärung

(Studienreglement WISO vom 1. September 2006 Art. 19 bzw. Art. 31)

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. 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.“

Ort/Datum: Bern, 2. November 2023

Name: