Essays in Development Economics

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

vorgelegt von

Simona Helmsmüller

im Sommersemester 2020

Abteilungssprecher:	Prof. Dr. Volker Nocke
Referent:	Prof. Dr. Markus Frölich
Korreferent:	Dr. Alexandra Avdeenko

Tag der Verteidigung: 07.10.2020

Für Kjell

Acknowledgements

I would like to thank Markus Frölich for his support as supervisor. His brilliance, research integrity, and dedication to evidence-based, pro-poor development continuously serve as inspiration to me. I am grateful for his confidence in me, for the great degree of flexibility shown, and for giving me the freedom to pursue my research interests.

Alexandra Avdeenko has also been an exemplary research supervisor as well as personal mentor and role model. I thank her as well as my co-author Andreas Landmann for their constructive criticism, patience, trust, and continuous motivation.

I would like to thank the staff of KfW's FZ-Evaluation Unit, in particular Eva Terberger, Anja Bentlage, Lena Hauk and Sarah Nohr, and the staff of the ILO PROMISE-IMPACT program, in particular Owais Parray and Yousra Hamed. Thank you for sharing data and information, for helpful comments, and friendly cooperation. The research in Pakistan would not have been possible without the commitment, dedication, and high quality of work of the staff of RSPN and NRSP, in particular Muhammad Tahir Waqar and Sharafat Hussain.

I thank my fellow researchers and colleagues at the Graduate School and at the Centre for Evaluation and Development for fruitful discussions about this work and academic life in general, including but not limited to (in alphabetical order): Adelina Garamow, Albrecht Bohne, Frederik Eidam, Jan Berkes, Justin Leduc, Katja Hofmann, Maria-Isabel Santana-Penczynski, Richard Senner, Torben Fischer, and Ulugbek Aminjonov. Tobias Etzel deserves a special thanks for coffee and discussions during walks and swims, for learning sessions, and for saving my plants. Thanks also go to Marion Lehnert for her patient and efficient assistance in many administrative matters, and for her warm smile.

I could not have completed this work without the support of my family and friends, near and far. Special thanks go to my parents, and Ulla and Jürgen Hauser for taking loving care of my son Kjell. My sister Daniela von Blum has been a continuous source of emotional support and stability. I thank Guru Sosale and Heide Humburg for providing a home in Mannheim and for unwavering friendship when I was out of aces. Finally, my biggest thanks go to my son Kjell for love, happiness, and purpose.

This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences.

Contents

List of Figures

List of Tables

Preface

This thesis asks the question: How can we design public policies or individual programs that alleviate some of the many plights associated with poverty in developing countries? Upon first sight, some solutions seem straightforward. But in practice, limited absorptive capacity, caused among other factors by missing human or social capital, inefficient institutions, or adverse policies, often imply that postulated impact chains do not fully unfold.

In this thesis, I evaluate three different interventions which all sought to better the lives of poor people, be it through health insurance or entrepreneurship training. Whereas the programs achieved some positive impact, these are more limited than expected during program design. In my dissertation, I therefore aim to not only quantitatively measure the impact of the programs, but to also understand the factors which challenge or contribute to the programs' success in a particular setting.

The first two chapters, of which one co-authored with Andreas Landmann, analyze the effects of two very similar health insurance schemes. Both schemes are supported by the German Development Bank KfW and both are implemented in Pakistan, albeit in different regions, namely in Khyber Pakhtunkhwa and Gilgit Baltistan. Yet, while very similar in design, the two schemes showed very different effects on the beneficiary population. In Khyber Pakhtunkhwa, the program managed to involve the private sector hospitals in the insurance scheme, leading to a shift in hospital choice from the public to the arguably higher-quality private hospitals. In contrast, private hospitals were de facto not participating in the insurance scheme in Gilgit Baltistan and there is some indication that it was not entirely effective in public hospitals either, rendering the program ineffectual. Moreover, tensions between different Islamic sects seem to have inhibited the enrollment of the population into the insurance scheme to begin with. The comparison of these two programs illustrate the importance of an intervention's contextual setting: Whether a well-designed program takes effect as intended also depends on absorptive capacity, such as public-private relations, capacity in public infrastructure, and cultural factors.

In the third chapter, co-authored with my supervisors Markus Frölich and Alexandra Avdeenko, we evaluate a training program for owners of micro and small enterprises in Indonesia. As training is costly, the program in Indonesia partnered with financial service providers and applied a train-the-trainer approach, where the loan officers then provided the actual client training. The program failed to achieve any sizable impact, except for one financial service providers, whose clients answered more knowledge questions correctly and applied better business practices after the training. This suggests that the choice of implementing partners is a crucial one in program design, which as yet has received little attention in program evaluations.

Although these three chapters are independent studies of different programs, they share important common ground in both, contribution and methodology. First, as laid out above, all three treatments achieved outcomes different from what was hypothesized in the programs' designs. This will be of little surprise to any practitioner given that the challenges a program faces during implementation are manifold and can change the program's direction. An economic impact evaluation should hence not purely focus on the econometric estimation of treatment effects, but also point out and discuss program inhibitors. As such, my thesis shows the potential of the private sector to overcome capacity constraints in the public sector (Chapter 1), the importance of carefully considering institutional and cultural factors such as weak governance and religious segregation in the target population (Chapter 2), and the differential effects across types of partner institutions used as program mitigators (Chapter 3).

Second, the empirical analysis in all chapters is based on large-scale household surveys, each consisting of two waves. Whereas the analyses in the first two chapters uses quasi-experimental methods, the third chapter is based on a randomized control trial. The latter has the higher internal validity and is often considered the gold standard of impact evaluation. For this study, we also registered a pre-analysis plan, contributing to the robustness of our empirical design, but also limiting the flexibility of our analysis. In contrast, the two studies using quasi-experimental methods have the advantage of using a random sample of the population, thus potentially achieving higher external validity. Further, the combination of propensity score matching and regression discontinuity design, together with an analysis of administrative data add to the credibility of the quasi-experimental methods and allowed a more flexible analysis of program constraints.

In the following, I provide a short summary of the three chapters of this thesis.

Health insurance in Khyber Pakhtunkhwa, Pakistan

Together with my co-author Andreas Landmann I empirically analyze the effects of free hospitalization insurance for the poorest quintile of the population on the quantity and pattern of inpatient care consumed. A novel feature of the paper is studying the differential choice of households between public and private health care providers in a setting where both market sectors participate in the insurance scheme. To this end, we apply two identification strategies which meaningfully complement each other: First, we exploit imperfect roll-out and compare insured and uninsured households using propensity score matching. Second, we exploit that eligibility is based on an exogenous poverty score threshold and apply a regression discontinuity design.

Our paper reveals that insured households do not visit hospitals more often, but have a larger propensity to choose private hospitals. We consider this an important outcome since the program at least in the short term seems to facilitate more equitable access to private hospitals with perceived higher quality - something that might also have longer term ramification for the market structure of health service providers. Our finding is also consistent with a more efficient program implementation in private hospitals, and shows the latter's ability to increase absorptive capacity in large insurance schemes.

This paper is currently under review for publication in the Geneva Risk and Insurance Review.

Health insurance in Gilgit Baltistan, Pakistan

In Gilgit Baltistan, a similar insurance scheme was implemented. Here, too, the Government covered the premiums of a hospitalization insurance for the poorest quintile of the population. However, with Gilgit Baltistan being an autonomous territory within the larger Kashmir region, the program here operates in a unique setting in a part of the world we know little about. I again exploit the exogenous poverty cut-off and apply a fuzzy regression discontinuity design to estimate local average treatment effects on the treated. Although this design has a high internal validity even when sample sizes are small, I find no effect of the program on financial outcomes, such as subjective financial burden or actual costs of health care, or on the usage of health care.

Two factors might have inhibited the program developing its full potential. First, I find some indication that the program was not fully operational at hospital level, suggesting a lack of institutional capacity and insufficient governance structures. Whereas private hospitals were reluctant to implement the scheme, partly because of mistrust in the reimbursement process, I find evidence in administrative data of either misreporting or malpractice in public hospitals. Second, my paper makes a novel contribution by analyzing heterogeneous effects along sectarian affiliation proxied by household location, if only descriptively. Especially the region's capital city Gilgit Town is geographically segregated along sectarian fault lines between conflicting Shia and Sunni Muslims, which effectively led to a sectarian health care infrastructure. A third sect, the Ismailis, is largely perceived as neutrally positioned between the conflicting sects, and an Ismaili-affiliated NGO is in charge of the program's implementation. I find descriptive evidence that the program achieved particularly high enrollment rates in predominantly Ismaili parts of town, and that these areas, too, are the only ones that experienced any increase in hospital care consumption. This finding underlines the importance of considering anthropological studies and cultural factors, such as religion, in econometric analyses.

Entrepreneurship training in East and West Java, Indonesia

In this study, co-authored with Markus Frölich and Alexandra Avdeenko, we measure the impact of a management training program for micro and small enterprises in Indonesia, designed and implemented by the International Labour Organization (ILO). There are a great number of training programs world-wide, ranging from high intensity individual consulting which aim at maximizing impact, to standardized training programs which aim at maximizing outreach. As varied as they are in design, as varied is the impact they achieve.

A novel feature of our study is that the program worked with twelve individual financial service providers. The ILO trained their loan officers to provide training and/or counseling to their clients running micro or small enterprises. The train-the-trainer approach allows a large outreach while operating cost-efficiently, but has the disadvantage of longer impact chains with limited direct control. Indeed, using a randomized control trial and panel data from 3,975 clients we find no evidence of changes in key outcomes, such as profits or household spendings. However, effects vary across partner institutions with rural banks achieving consistent, albeit small, improvements in knowledge and practices. In the absence of heterogeneous effects along observable dimensions, the differences could be explained by client selection along unobservable dimensions or higher quality training delivery.

However, since final outcomes such as profits and spendings still remain unaffected even when knowledge and practices improved, the program might have failed to address the most relevant constraints. Nevertheless, our evaluation highlights the role of partner selection in similar multi-stakeholder interventions.

Chapter 1

Does free hospitalization insurance change health care consumption of the poor? Short-term evidence from Pakistan.

With Andreas Landmann

1 Introduction

In lower- and middle-income countries, economic inequity is linked to inequity in health. One of the chains by which these are bound together is through high out-of-pocket (OOP) expenditures for health. These affect poor households in two ways: First, they create financial distress, in particular in the case of expensive events, such as hospitalizations. Second, they create barriers to health care, contributing to a low health status and therefore potentially also lower ability to generate income. A straightforward approach to breaking this vicious cycle is to provide health insurance to the poor. Many recent reforms in lower-and middle-income countries around the world are thus establishing inclusive health insurance schemes, with the aim of not only reducing financial distress, but also to change health seeking behavior by reducing financial barriers.

In this paper, we explore whether fully subsidized insurance for hospitalization changes health service utilization of low-income households in Pakistan. In particular, we evaluate the Social Health Protection Initiative (SHPI) in the province of Khyber Pakhtunkhwa (KP), which grants fully subsidized health insurance to the poorest quintile of the population. By studying the patterns of inpatient care consumption, we not only investigate changes in the quantity of care consumed, but also study whether the composition of care changes. An especially relevant dimension here is the probability to seek care from private providers, which patients associate with higher quality in our study. To evaluate the effect of insurance coverage, we use two features of the program. First, we exploit incomplete rollout and match insured to comparable, eligible but uninsured households using propensity score matching. Second, we implement a regression discontinuity design, using the fact that eligibility for the program is based on an exogenous poverty score. These approaches allow us to calculate two separate effects: the average treatment effect on the treated for eligible households and the intention-to-treat effect for households close to the cut-off.

The results of both econometric approaches suggest that the SHPI did not have significant effects on the quantity of health care consumption, despite high levels of neglected health care. We find no increase in the propensity of using inpatient health care services, no increase in the share of individuals who visited a hospital more than once in the past year, and no decrease of neglected health care. However, we find evidence suggesting a change of provider choice from public to private facilities. This is consistent with a larger reduction of relative costs of private care versus public care in our data as well as with a small number of claims from public hospitals in administrative data, suggesting that public hospitals implemented the program less efficiently. Given the better resources and higher client satisfaction associated with private hospitals, we nevertheless interpret this as an important positive impact of the program. Should the demand shift from public to private providers not be in the interest of the government, however, additional programs to strengthen the capacity of public hospitals might be necessary.

Several studies have analyzed the effect of protecting low-income households through health insurance. Randomized control trials (RCT) on micro-health insurance programs have shown some promising impacts in terms of financial protection (see Habib, Perveen, and Khuwaja 2016 for a recent review), access to medical services (e.g. Levine, Polimeni, and Ramage 2016; Thornton et al. 2010), and social outcomes (e.g. Landmann and Frölich 2015; Frölich and Landmann 2018). In line with this, there is a move towards universal health coverage via a rapid expansion of state-funded health insurance arrangements across lower- and middle-income countries (Lagomarsino, Garabrant, Adyas, Muga, and Otoo 2012; Reich et al. 2016). However, results from RCTs do not necessarily carry over to larger programs where limited absorptive capacity might hamper the effect (Mangham and Hanson 2010), and not every program design allows for plausible identification strategies. Whereas some quasi-experimental studies exist on health insurance reforms in countries such as India, China, and Indonesia (Wagstaff, Lindelow, Jun, Ling, and Juncheng 2009; Prinja, Chauhan, Karan, Kaur, and Kumar 2017; Vidyattama, Miranti, and Resosudarmo 2014), evidence on the Pakistani program is scarce, even though it is a very relevant case for several reasons.

Pakistan is a lower-middle income country with the sixth largest population in the world, where poverty and the risk of falling into poverty are still widespread. Government spending on health has been well below one percent of GDP until recently, and around 90 percent of private expenditure had to be paid out-of-pocket in 2016 (World Bank Indicators 2016).[1] This situation increases the need for inclusive insurance solutions. While the fragmented nature of the health system with provincial responsibility for the health policies renders reforms more difficult, these might have particularly high effects. In addition, through the fully subsidized scheme with household enrollment based on a pre-existing poverty census, the program achieved remarkably high enrollment rates and mitigated the problem of adverse selection which challenges similar interventions in other countries (Banerjee et al. 2019; Asuming 2013).

At the same time, Pakistan features a dual health sector with both private and public providers operating in the same market. A similar situation exists in India, which has undergone large-scale reforms with far-reaching transformations in the health care market a few years earlier. In this context, large health financing reforms might shape the long-term character of the market and it is therefore worth studying how demand in each sector is affected by insurance. For Cambodia, Levine et al. (2016) find that insured households shift towards public hospitals, but in their case private hospitals were not empaneled by the program, which means that patients simply shift towards participating hospitals. This is also what Thornton et al. (2010) find in Nicaragua, where insured households were more likely to visit health care providers covered under the insurance. We contribute to this literature by studying a setting under which insurance coverage was in principle available at both public and private providers. Note that effective coverage might still have differed between the two sectors, as these face different incentives and dispose of different resources and governance structures for implementation. In fact, our finding of an increased usage of private care is consistent with a more efficient program implementation in private hospitals, suggesting the importance of including the private sector to increase absorptive capacity. Despite the relevance of the research question for public policy, there is virtually no evidence on the impact of large state-funded insurance schemes on public versus private health systems if insurance coverage is offered at both. By looking at demand-side effects, we thus contribute to closing this evidence gap in the context of a nascent health insurance system in Pakistan.

The rest of this paper is organized as follows. In Section 2 we provide the country context and program details. In Section 3 we present details of our dataset and summarize descriptive statistics. In Section 4 we explain our two main identification strategies and assess the plausibility of the underlying assumptions. Section 5 contains our main results on the usage of inpatient care and a brief analysis of heterogeneous effects. In Section 6 we discuss effect channels and challenges in implementation. The last section concludes.

[1] World Bank Indicators until 2016 are available at http://data.worldbank.org/country/pakistan.

2 Rationale of the intervention

2-1 Challenges in health care in Pakistan

Poor health is wide-spread in Pakistan. In its report from 2017, the World Health Organization (WHO) attests Pakistan to have the fifth highest burden of tuberculosis world-wide and the highest rate of malaria in the region, while being one of only three countries in the world where residual poliomyelitis (infantile paralysis) has not been eradicated. Hepatitis B and C, dengue and chikungunya show high prevalence, and leprosy and trachoma are still reported. Regarding non-communicable diseases, cancer, diabetes, respiratory and cardiovascular diseases are among the main causes of death. Maternal and child mortality are among the highest globally (WHO 2017).

With the abolition of the Federal Ministry of Health in 2011, health care management and regulation became the responsibility of the Provincial Governments. These maintain networks of multi-tiered health care providers, yet overall public spending on health care is very low. In consequence, the quality, in particular of the primary health care infrastructure, is limited, suffering from political interference and corruption, shortage of trained personnel, staff absenteeism, non-functioning facilities, and lack of medicines (ADB 2019; WHO 2013). Notably, the non-existence of public family physicians means that hospitals are often the first point of contact with the formal health care infrastructure. But even major district hospitals often lack specialized staff such as gynecologists, anesthetists or pediatricians (TRC 2012). Therefore, households often use private service providers (Government of Pakistan 2016), implying that most of the health expenditures must be borne by the patient (Nishtar et al. 2013; WHO 2017). Also in public hospitals, expenditures, such as for medications, are usually paid out-of-pocket.

Social security systems are not broadly spread and leave the large majority of the population uncovered (Nishtar et al. 2013).[2] Private health insurers, though existing, lack the depth of penetration, in particular into rural and poorer population groups, covering less than 3% of the population (Nishtar et al. 2013). While there are a number of micro health insurance schemes run by non-governmental organizations (NGOs), they have not achieved broad outreach. With one third of the population living of less than 1.5 USD per day and in the absence of affordable insurance, it is reasonable to assume that financial constraints lead to less than optimal health care among the poor population of Pakistan.

2-2 The *Social Health Protection Initiative* (SHPI)

Against this background, the Government of the Province of KP launched a large-scale program to improve access to health care, called the SHPI. With financial and technical assistance of the

[2] According to Nishtar et al. (2013), there are three vertical systems servicing 14.12% of the population: By the Armed Forces, by the Fauji Foundation for retired military servicemen, and by the Employees Social Security Institution for public servants. These are vertical, i.e., they have mutually exclusive service delivery infrastructures.

German KfW Development Bank, the program intends to reduce financial barriers to health care through the introduction of a subsidized health insurance. The program uses a pre-existing national poverty score, which had been assigned to all households in Pakistan based on a proxy means test (PMT) in 2010.[3] All households below a pre-defined cut-off poverty score were selected to receive the insurance card at fully subsidized rates. The first phase of the program was officially launched in December 2015 in the four pilot districts Chitral, Kohat, Malakand, and Mardan. It covered households with poverty scores below 16.17, corresponding to the poorest 21% of households in this area (approx. 0.7 million people targeted). The program delivered the cards to beneficiaries via selected regional NGOs, who were in charge of forward campaigning (including but not limited to banners and call centers providing general information, radio announcements and posters to inform about dates of enrollment at village level) as well as the physical distribution of insurance cards at special card distribution centers (including permanent offices at district level and temporary offices at village level). Following the official enrollment dates, unenrolled eligible households should be contacted directly by the insurer via phone or in person (Oxford Policy Management 2016). In addition, the consulting company advising the program on behalf of the KfW Development Bank verified the distribution of cards via a limited number of spot checks. Six months after the official launch, the insurer reported an enrollment rate of 87.3% among the target population in the two pilot districts considered in our study (Oxford Policy Management 2017).[4]

During our study period, one insurance policy covered a household of seven members (assumed typical case: household head, spouse, four children and one elderly dependent). The benefit package addressed maternity-related care as well as non-maternity hospitalization, up to an annual limit of PKR 25,000 (238.25 USD)[5] per person.[6] This covered treatment for normal delivery and C-sections, as well as a pre-defined list of 497 medical procedures requiring hospitalization. Notably, the program did not cover outpatient care.[7]

The insured households could obtain these services at one of the empanelled hospitals, which include public and private health care providers.[8] Prior to the distribution of insurance cards, the program identified and contacted potential hospitals for empanelment in the program, but was met with skepticism. Private providers were hesitant to join the network due to concerns regarding the reimbursement of costs, religious beliefs, or fear of stricter tax controls (Oxford

[3]The Government revised the poverty score again only after our endline survey, so that during our study period and the five years prior to that the score remained unchanged for all households.

[4]Whereas the program also foresaw voluntary, non-subsidized health insurance to the non-eligible population, no such product was on offer at the time of our study.

[5]Exchange rate on December 31, 2015.

[6]We have administrative cost data only for a short period of time overlapping our study. Between January and July 2017, the median cost of treatment was 15,000 PKR in the two pilot districts considered here.

[7]A second phase of the program, starting in January 2017, saw the gradual roll-out to the remaining districts and raised the poverty cut-off score to 26.75, thus covering approximately 51% of households in the district (approx. 14.4 million people targeted). The program also altered the benefits slightly, covering eight household members, raising the annual coverage limit and including tertiary care providers, but notably still restricting coverage to cases of inpatient care. Table A.1 in Appendix 1-1 provides an overview of the program features in both phases. Following the completion of our study, the Government initiated Phase 3, which extended the program to cover up to 69% of the population in the entire province of KP. Further extensions are planned with the aim of achieving universal health coverage.

[8]Despite there being a number of NGOs active in the health sector in Pakistan, such as the Aga Khan Foundation, there are no NGO-run hospitals in our survey region.

Policy Management 2016; Oxford Policy Management 2017). Public hospitals also showed little interest in the program until Government influence was used to encourage joining the program. Nevertheless, the program was able to empanel around one third of the candidate private hospitals, as well as the two main public hospitals in each district. During our survey period, however, some hospitals were de-paneled due to the use of unnecessary procedures or, in one case, a conflict of interest. Overall, during our study period, there were at least four public and seven private hospitals available for service provision at all times.[9] Before the official launch, the program trained hospital staff and established service desks in each empaneled hospital for identification of beneficiaries, verification of eligible treatment and available balance, and claim management for cashless service provision. For further gatekeeping, a District Medical Officer employed by the insurer visited clients within 24 hours after admission.

Fully subsidized premiums naturally lead to an adverse incentive structure for the insurance company: The Government transfers the insurance premiums for each enrolled household, hence creating a steady flow of income from the Government to the insurer. At the same time, the cost structure of the insurance company, which was also responsible for the distribution of insurance cards, is determined by actual usage. The insurance company would hence benefit from not informing insured individuals of the full benefit package. Therefore, a mandated awareness campaign accompanied each phase of card distribution, carried out by the implementing insurance company as well as the NGOs. A further challenge was the identification of beneficiary households, which were selected based on the poverty census from 2010. This implies not only that the program does not necessarily target the currently poor, but also challenged the localization of households for enrollment given that addresses were partially outdated.[10]

The Government of the Province of KP is spearheading the program, supported by the KfW Development Bank with financial and technical cooperation. Considering the difficult political landscape of Pakistan, the Provincial Government had its own vested interest in the program which likely went beyond the distributional goals: At the time of our study, the Province of KP was governed by a different party than held power of the Federal Government. The Federal Government of Pakistan planned and slowly started rolling out a similar national social health insurance. While the Federal Government had not implemented the national scheme in the Province of KP at the time of our study and hence did not create competition in economic terms, it most certainly imposed political competition. The Provincial Government was hence politically motivated to make the SHPI widely-known and clearly associated with their party. Nevertheless, limited awareness remained a concern, which we further address in Section 6.

[9] Specifically, in Malakand the program started with three public hospitals, and three out of nine identified private hospitals. Later, one public and one private hospital were de-paneled, while one new private hospital joined. In Mardan, the program started with three public and five out of 14 identified private hospitals. Later, one public and two private hospitals were de-paneled, while one new private hospital joined (Oxford Policy Management 2016, Oxford Policy Management 2017).

[10] After our study, the Government initiated a new survey to update the poverty score, which might improve targeting in later phases.

2-3 Intended effects

The rationale behind the SHPI is that the insurance would lower the cost of hospitalization and that this would affect households along two dimensions. On the one hand, lower OOP expenditures should encourage an increased usage of health services and hence the quantity of health care consumed. Thus, the program would contribute to health improvement. On the other hand, lower OOP expenditures directly decrease the households' financial burden and reliance on more stressful coping strategies. Thus, the program would contribute to financial protection against health risks.[11] Whereas we acknowledge the importance of financial protection for the poor in its own right, we concentrate on the first aim in this study, i.e., improving health by increasing health care consumption.

3 Data and descriptive statistics

3-1 Survey data

We make use of household survey data collected specifically for the program evaluation. Four months prior to the start of the first program phase, we collected baseline data (autumn 2015). We carried out the endline survey twelve to fifteen months after the first program rollout (spring 2017). Prior to the design of this evaluation, the Provincial Government had selected four pilot districts for the first phase of the program, where the insurance was to be offered exclusively. We therefore collected data in these four districts as well as in four additional districts, initially intended as control districts. Political dynamics, however, led to an early extension of the program into control districts as well as differences in rollout across the four pilot districts. The data we use in this study therefore is from only two of the four pilot districts, where the initial rollout plan was largely followed and where our identification strategies are still valid (Malakand and Mardan).[12] We also use data of the control districts for some robustness checks.[13] Appendix 1-2 summarizes the timeline of the SHPI roll-out and our surveys in the relevant districts.

Our sampling strategy is a multi-staged clustered approach. We randomly selected 24 union councils as survey clusters in the two pilot districts considered here. The poverty census of

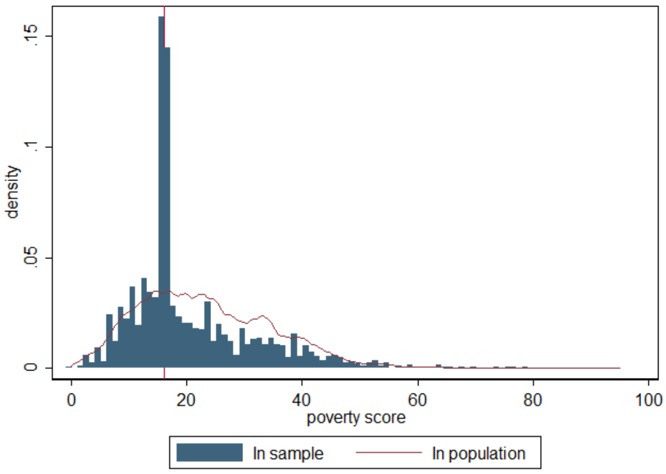
[11] Additionally, the Government aimed at increasing quality and accountability of public hospitals by ensuring a client-based flow of funds through the program. We do not consider supply-side effects in this study.

[12] We also collected data in the two other pilot districts, namely Chitral and Kohat. In Kohat, however, our monitoring during the endline survey revealed several problems. Specifically, we find particularly high differences between official and self-reported enrollment in the urban areas. We also faced the highest attrition rate (7%) in this area. In addition, there were problems in the project implementation in this district with one hospital being suspected of fraud. We thus exclude the data from the whole district out of prudence. The district of Chitral, on the other hand, was hit by a severe flood just prior to the baseline. This negatively affected our data collection in terms of access to some areas. Also the empanellment of hospitals was much delayed, and the program became fully operational only after our endline survey, which led us to exclude this district as well. In Mardan, the second program phase started three months prior to the endline, which might create some first additional effects, but does not invalidate our empirical approach. We discuss implications for the regression discontinuity design in Section 4.

[13] We selected the four control districts using an algorithm matching on publicly available socio-demographic indicators and health infrastructure. The second phase rolled out prematurely in two of these, but using a different cut-off score.

2010 served as a sampling frame for the third and fourth stage: Stratified random sampling of 70 villages and then 1,200 households in the two pilot districts. To increase power for our identification strategies, we additionally sampled 240 households below and 480 closely around the cut-off poverty score in the pilot districts (i.e., an additional 20% and 40% respectively in each survey cluster). Therefore, our baseline sample in the pilot districts consists of 1,920 households of which 828 were eligible for the insurance. Figure 1.1 depicts the distribution of the poverty score in the population and in our sample respectively, in the selected union councils of the two pilot districts, illustrating the degree of oversampling below and around the cut-off score of 16.17.

Figure 1.1: Distribution of poverty score in sample and population in selected UCs of Malakand and Mardan



► *Note*: This figure shows the distribution of the poverty score in our sample (blue bars) and in the sampling frame (red line) in sampled union councils (UCs) of Malakand and Mardan.
► Sample: Household-level sample (panel, N=1,842) and BISP sampling frame (N=71,591).
► Source: Baseline survey (2015) and BISP survey (2010).
► The poverty cut-off score (assigned in 2010) determining eligibility for the first phase of the SHPI is 16.17. The figure illustrates the degree of oversampling below and around this cut-off.

Interviewing the same households in the baseline and endline study, we constructed a household panel dataset. We used computer-assisted personal interviews in both survey waves, allowing the collection of GPS coordinates, an efficient survey administration and, thus, a minimal level of attrition of under 2.5%. An additional 1.2% of the sample were dropped in the data cleaning process, leading to a panel dataset of 1,842 households in the two pilot districts, of which 795 eligible households. We collected information on economic conditions, subjective well-being, the use of health care during childbirth, outpatient care, and neglected health care on household level. In light of the focus of the program on inpatient treatment, we recorded the history of inpatient care, including associated costs, and the subjective health status of each household member individually. This leads to a final panel sample size of 12,862 individuals, thereof 6,007 eligible for insurance, when considering inpatient care. In the endline survey, we administered the same questionnaire, but added questions on the enrollment status and familiarity with the program.

3-2 Data quality and processing

Our local research partner pre-tested, translated, and implemented the questionnaires on tablet computers. To a large extent, items are based on a questionnaire which had been tested repeatedly and demonstrated high validity in previous projects. At the end of each survey day, supervisors uploaded the data from the tablets onto a server and we downloaded data in Germany for monitoring of interviewer performance and data quality. Daily quality control included automated consistency checks, spot checks, and follow-up phone calls. Comparing GPS coordinates of a household at baseline and endline guaranteed that indeed the same household was interviewed.

We winsorized quantitative variables which showed a large variation. The level of winsorizing depends on the initial variation of the specific variable and ranges from the 90th to the 99th percentile. We performed a principal component analysis of asset ownership to derive a variable for socio-economic standing (in the following denoted wealth index) and a principal component analysis of access to amenities such as toilets and drinking water to derive a variable for hygienic condition (in the following denoted hygiene index). For per capita household income, we account for economies of scale within the household and use the square root equivalent scale, i.e., we divide household income by the square root of household size. (An implication is that, e.g., a four-person-household has twice the monetary needs of a single person.)

We note that our survey might suffer from coverage error. This stems from the fact that the best available sampling frame, the poverty census, was collected in 2010 and is hence partly outdated. Moreover, in the absence of official addresses of most households, the identification of sampled households was a challenge and might have led to population subgroups being missing not-at-random. However, one should note that the SHPI used the same frame to determine program eligibility. While our results might not be fully representative, e.g., for young and newly-formed or migrated households, they are internally consistent under the plausible assumption that all groups used for comparison in our identification strategies are likely to be similarly affected.

3-3 Baseline characteristics

Table 1.1 contains selected baseline characteristics of households and individuals in our panel samples, i.e., sampled households and their members in the two districts Malakand and Mardan with baseline as well as endline information. We separately present statistics on the full sample as well as on households eligible for insurance coverage, i.e., with a poverty score below 16.17. Note that the goal here is not to give a representative picture of the population but to describe the samples we are using for our analysis. These samples include oversampling below and around the cutoff, and therefore do not reflect average differences between eligible and non-eligible households in the population. We present statistics for the subsample of randomly selected households in Table A.2 in the Appendix (differences to table below are marginal).

Table 1.1: Baseline characteristics of full sample and subsample of eligible population (selected variables)

	Full sample				Eligible sample			
	Mean (1)	Std. Dev. (2)	Min (3)	Max (4)	Mean (5)	Std. Dev. (6)	Min (7)	Max (8)
Panel A: Household-level variables								
Insurance status at endline	0.45	0.50	0.00	1.00	0.65	0.48	0.00	1.00
Poverty score	20.61	11.15	0.00	79.00	12.53	3.60	0.00	16.17
P.c. monthly income (sqr. root equiv.)	538.65	829.68	0.00	8,888.89	388.80	611.17	0.00	8888.89
Wealth index	-0.25	1.92	-3.40	12.23	-0.56	1.56	-3.21	6.58
HH size	7.43	2.78	1.00	23.00	8.09	2.54	3.00	21.00
Electricity in HH	0.96	0.20	0.00	1.00	0.95	0.21	0.00	1.00
Tab water supply in residence	0.12	0.32	0.00	1.00	0.12	0.32	0.00	1.00
Private flush toilet	0.36	0.48	0.00	1.00	0.29	0.45	0.00	1.00
Reported dist. to next hosp. (minutes, win99)	43.62	26.35	0.00	150.00	44.41	26.90	0.00	150.00
Use of prof. assist. during childbirth	0.89	0.31	0.00	1.00	0.86	0.35	0.00	1.00
Case of neglected health care	0.14	0.35	0.00	1.00	0.16	0.36	0.00	1.00
Case of outpatient care	0.78	0.42	0.00	1.00	0.80	0.40	0.00	1.00
Citing health shock as a risk	0.94	0.23	0.00	1.00	0.95	0.22	0.00	1.00
Dif'ty finding money for health care >= 8 (scale 1/10)	0.48	0.50	0.00	1.00	0.49	0.50	0.00	1.00
Having heard of insurance	0.02	0.14	0.00	1.00	0.01	0.12	0.00	1.00
Observations	1,842				795			
Panel B: Member-level variables								
Age (win99)	23.17	18.17	1.00	90.00	22.25	17.45	1.00	90.00
School-aged (6 to 16)	0.33	0.47	0.00	1.00	0.38	0.49	0.00	1.00
Female	0.48	0.50	0.00	1.00	0.48	0.50	0.00	1.00
Prim. school not comp'd.	0.59	0.49	0.00	1.00	0.63	0.48	0.00	1.00
Comp'd sec. educ. or higher	0.12	0.32	0.00	1.00	0.08	0.28	0.00	1.00
Worked for salary in previous month	0.19	0.39	0.00	1.00	0.18	0.39	0.00	1.00
Usage of inpatient care	0.05	0.22	0.00	1.00	0.04	0.21	0.00	1.00
Cost of last treatment (PKR, win99)	25,086	44,655	0	300,000	25,072	44,769	500	300,000
More than one admittance to hospital	0.26	0.44	0.00	1.00	0.26	0.44	0.00	1.00
Use of private hospital	0.31	0.46	0.00	1.00	0.30	0.46	0.00	1.00
Observations	12,862				6,007			

- *Note*: This table shows the baseline characteristics of the full sample and the eligible subsample (households (members) with a poverty score below 16.17). Selected variables on the left, statistics on top.
- Sample: Households and their members in full or eligible sample (panel, varying N).
- Source: Baseline survey (2015), insurance status from endline (2017).
- Column (1) displays the mean for continuous/shares for binary variables in the full sample, Column (2) the standard deviation, Columns (3) the minimal and (4) the maximal value in the full sample. Columns (5) to (8) display the same statistics for the subsample of eligible households and their members.
- The suffix *win99* indicates that we winsorized the variable at the 99th percentile level. Monetary variables in PKR (100 PKR = 0.953 USD on December 31, 2015).
- Note that the insurance status in the non-eligible sample is non-zero due to the roll-out of the second phase of the program in the district of Mardan shortly before our endline survey.
- Table A.2 in the Appendix contains the statistics for the subsample of randomly selected households (excluding sampling). Differences are small and statistically insignificant.

The average household in our full sample consists of 7.43 members and of 8.09 members in the subsample of eligible households. The members of eligible households are slightly younger (22 versus 23 years), more likely to be of school-aged (38% versus 33%), and a larger share has not completed primary school (63% versus 59%). Conversely, a smaller share of members has completed secondary school or higher (8% versus 12%). Consistently, the per capita household income among eligible households is around two thirds that of the full sample. There is a high gender disparity in education and work (not shown in table): Among male adults, 47.0% in our full sample have no formal education, and this percentage rises to 82.1% among female adults. Similarly, 67.2% of male adults have worked for pay in the year prior to the baseline survey, compared to only 3.3% of female adults. Overall, hygienic conditions are sub-optimal: Whereas

96% of households have electricity in their home, only 36% have a private flush toilet and only 12% have tap water supply in their residence. Travel time to the next hospital averages 44 minutes.[14] Notably, awareness about insurance is virtually non-existing at baseline.

Regarding the use of health care services, 5% of individuals in the full sample reported an overnight stay in a hospital within the twelve months prior to baseline. To understand the socioeconomic drivers of using inpatient services, we run three logit regressions including individual and household covariates with different proxies for poverty, namely the poverty score, the per capita household income, and the wealth index (results shown in Table 1.2). Older and female individuals are consistently more likely to consume inpatient care, where the gender effect is driven by childbirth related admissions (effect disappears when childbirth is excluded, see Table A.3 in the Appendix). The results also suggest that poorer households consume significantly more inpatient care when using the wealth index as proxy for poverty. This is consistent with the fact that both wealth as well as health represent outcomes of long-term processes.

Table 1.2: Logit regression of hospitalization on individual and household characteristics, baseline

Admission to inpatient care	Logit 1 Coef. (1)	Std. Error (2)	P-value (3)	Logit 2 Coef. (4)	Std. Error (5)	P-value (6)	Logit 3 Coef. (7)	Std. Error (8)	P-value (9)
Poverty score	0.004	0.005	0.497						
Per capita monthly HH income				-0.009	0.006	0.134			
Wealth index							-0.095	0.031	0.002
Female	0.229	0.085	0.007	0.228	0.086	0.008	0.222	0.084	0.008
Age	0.028	0.001	0.000	0.028	0.002	0.000	0.028	0.002	0.000
Household size	-0.025	0.022	0.248	-0.027	0.021	0.209	-0.005	0.021	0.815
Hygiene index	0.031	0.040	0.449	0.009	0.040	0.822	-0.028	0.044	0.532
Dist. to next hospital (min.)	-0.002	0.002	0.441	-0.002	0.002	0.466	-0.001	0.002	0.523
Const.	-3.643	0.294	0.000	-3.499	0.217	0.000	-3.794	0.249	0.000

- *Note*: This table shows the coefficients of logit regressions of a dummy indicating admission to hospital on individual and household covariates. Covariates on the left, statistics on top.
- Sample: Member-level sample (panel, N = 12,852).
- Source: Baseline survey (2015).
- Columns (1), (4), (7) display the coefficient estimates from the logit regressions, Columns (2), (5), (8) the standard errors and Columns (3), (6), (9) the p-value of the two-sided tests that the coefficient is equal to zero, with one of three different proxies for poverty respectively. Standard errors are adjusted for 24 clusters in union councils.
- Table A.3 in the Appendix contains the results excluding childbirth related hospitalization. The general direction and significance of coefficients remains unchanged, but females are no longer significantly more likely to be admitted to hospital.

The conclusion that in our sample, the less wealthy are more likely to consume inpatient health care, does not necessarily imply that poor households are not restricted in their access to health care. Instead, the finding could be driven by higher health needs, as health and poverty are related by causality running in both directions (Wagstaff 2002). We therefore also check the relation of the wealth index with other important outcomes of interest, namely, a measure of subjective health status, using a private facility (conditional on being admitted), and neglected health care in Table 1.3. To do so, we repeat the regressions, controlling only for the evidently important covariates age and gender, but also including squared terms for a more flexible form. The wealth index and its square are correlated not only with admission to inpatient care

[14] We did not ask to specify the medium of transport, so this likely differs across households.

(Column (1)), but also with the subjective health status, which improves for individuals in wealthier households (Column 4). Also, wealthier households are more likely to visit a private hospital (Column 7), where care is frequently perceived to be of higher quality, and less likely to report an incident of neglected health care (Column 10).[15] Our data therefore supports the hypothesis that poor households are indeed restricted in their access to health care, both in quantity and perceived quality.

Table 1.3: Usage patterns and health care needs

	Admission to inpatient care			Health status			Use of private versus public hospitals			Neglected health care		
	Coef.	Std. Error	P-value	Coef.	Std. Error	P-value	Coef.	Std. Error	P-value	Coef.	Std. Error	P-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Wealth index	-0.094	0.032	0.003	0.004	0.017	0.813	0.296	0.079	0.000	-0.291	0.062	0.000
Wealth index sq.	0.005	0.006	0.479	0.003	0.002	0.082	-0.031	0.019	0.105	0.033	0.010	0.001
Female	0.223	0.084	0.008	-0.126	0.017	0.000	-0.447	0.167	0.007			
Age	0.022	0.006	0.001	0.005	0.002	0.008	-0.008	0.015	0.591			
Age sq.	0.000	0.000	0.294	-0.000	0.000	0.000	0.000	0.000	0.893			
Const.	-3.834	0.193	0.000	4.551	0.055	0.000	-0.225	0.304	0.458	-2.078	0.187	0.000
N	12,852			12,852			589			1,849		

- *Note*: This table shows the coefficients of regressions on various outcomes (logit for binary outcomes, linear regression for health status variable). Covariates on the left, outcomes and statistics on top.
- Samples: Member-level and household-level samples (panel, varying N). Columns (7) to (9) conditional on reporting a case of inpatient care.
- Source: Baseline survey (2015).
- Columns (1), (4), (7), (10) display the coefficient estimates from the regressions, Columns (2), (5), (8), (11) the standard errors and Columns (3), (6), (9), (12) the p-values of testing that the coefficient is equal to zero, on different outcomes respectively. Standard errors are adjusted for 24 clusters in union councils.
- We measured neglected health care on household level, hence we regress only on household level covariates. Results are robust towards including the share of female, old, or young household members as covariates.

4 Econometric approach

We use two identification strategies, which estimate different effects. First, we match insured and non-insured individuals and households on the propensity to receive insurance estimated from baseline values. This provides an estimate of an Average Treatment Effect on the Treated (ATT). Second, we apply a sharp Regression Discontinuity Design (RDD) using the poverty score as running variable. This provides an estimate of an intention-to-treat (ITT) for observations around the cut-off. Table 1.4 illustrates the different samples considered for the two estimators.

[15]The result on the usage of private hospitals shown in the table is obtained by restricting the sample to individuals with a case of inpatient care. It also sustains, albeit less pronounced, when running the regression on the full sample, unconditional of a case of inpatient care.

Table 1.4: Treatment and control groups in two estimators

	Treatment group	Control group
Propensity score matching (PSM)	• insured HH/members from two pilot districts • poverty scores $\in [0, 16.17]$	• uninsured HH/members from two pilot districts • poverty scores $\in [0, 16.17]$
Regression discontinuity design (RDD)	• insured and uninsured HH/members from two pilot districts • poverty score $\in [16.17-B, 16.17]$	• insured and uninsured HH/members from two pilot districts • poverty score $\in [16.17, 16.17+B]$

► *Note*: This table illustrates the different samples considered for the propensity score matching and regression discontinuity design estimators respectively.

4-1 Propensity score matching (PSM)

Among eligible households in our PSM sample, the program achieved self-reported enrollment rates of 65.2% of households. This is remarkably high,[16] yet a sizable number of households targeted by the program did not report themselves insured in our survey, likely due to imperfections in program roll-out. It is important to note that we make use of the self-reported insurance status, instead of the official status as per administrative data. We believe that households which are officially insured but not aware of this are more likely to behave as if they were uninsured and should hence be part of the control group.[17] For ease of notation, we will use the term *(un-)insured* to refer to the self-reported insurance status from now on.

We exploit the fact that a third of the target population remains uninsured, and estimate the ATT using the following propensity score matching estimator:

$$\beta_{PSM} = \frac{1}{|I_1|}\sum_{i\in I_1}\left\{Y_{1i} - \frac{\sum_{j\in I_0} Y_{0j} G(\frac{P_j - P_i}{B_n})}{\sum_{k\in I_0} G(\frac{P_k - P_i}{B_n})}\right\}, \tag{1.1}$$

where β_{PSM} is the statistic of interest, the average treatment effect on the treated. I_1 is the set of insured households within the region of common support, I_0 is the set of uninsured households, Y_{1i} is the outcome for an insured household, Y_{0j} for an uninsured household, $P_j = Pr_j(insured|Z)$ is the propensity score, i.e., the probability of being insured conditional on a set of covariates Z, $G()$ is the epanechnikov kernel, B_n the bandwidth.

Two assumptions are key to this approach (Todd 2010): Conditional mean independence and common support. Whether the conditional mean independence assumption is fulfilled is not directly testable, but hinges on the considered set Z for calculating propensity scores (Smith

[16] As comparison, Banerjee et al. (2019) report enrollment rates of 8% in the Indonesian national health insurance program of their study, which they managed to increase to 30% under a treatment arm with full premium subsidization and assistance in the enrolment process.

[17] Only 2.5% of ineligible households who report themselves insured in our survey are not insured according to administrative data. In contrast, 74.4% of eligible households who report themselves uninsured are registered as insured in administrative data. The numbers are in line with programs in other countries such as Philippines and Rwanda, where Lagomarsino et al. (2012) note that Government enrolment figures do not always match household survey data. In our case, three factors likely contribute to the deviance: (i) The household never received the card and the administrative data is fraudulent. (ii) The household was enrolled after our endline survey. (iii) The household was enrolled, but the interviewed household member was not aware of it.

and Todd 2005). The lowest bias arises when Z includes all variables that simultaneously affect insurance status and considered outcomes.[18] Many of these variables, such as education, prior insurance knowledge or household size, are observable to us from the baseline survey and we systematically include them in our matching procedure. Tables A.4 and A.5 in the Appendix show means of all collected baseline variables for insured and uninsured households. The two groups differ significantly only in their willingness to take financial risks, with the uninsured households being more willing to bear risks. This is in line with the theory that risk-averse individuals have a higher incentive to seek insurance coverage. In addition, there are a number of unobservable variables, such as geographic accessibility, quality of accessible health care, intensity of the awareness campaign, quality of education, or interviewer effects. Many of these are however likely to be geographically clustered. Indeed, enrollment rates in our sample of eligible households range from 40% to 80% in the 24 union councils of the two districts, as depicted in Figure 1.2.[19] Correspondingly, we tested whether the set of union council dummies contributes to explaining enrollment and find this to be the case (p-value of an f-test testing joint significance=0.001). We therefore estimate propensity scores using a probit model with a vector of union-council dummies and selected baseline variables in linear and quadratic terms.[20] Tables A.4 and A.5 in the Appendix demonstrate the achieved balancing on union councils and baseline variables. Figures A.2 and A.3 in the Appendix show the distribution of the poverty score among insured and matched uninsured samples, underlining the credibility of the conditional mean independence assumption.

To ensure common support, we restrict our treatment sample to individuals or households with a propensity score above the 99th percentile score among the control group, as suggested in Caliendo and Kopeinig (2008). This eliminates 10.23% of households and 5.53% of household members in our respective treatment groups, for whom we have no suitable control observations. Furthermore, there are some gaps in the density of propensity scores in the control group. This is no concern though, as we have sufficient density to the left and right of these gaps for kernel matching. Nevertheless, we follow Smith and Todd (2005) and additionally drop 1% of our treatment observations at which the propensity score density of the control group is at its lowest. Figures A.4 and A.5 in the Appendix demonstrate that common support is thus sufficiently ensured.[21]

For the calculation of standard errors we account for the fact that propensity scores are estimated and that variables are clustered on the union council level by providing clustered

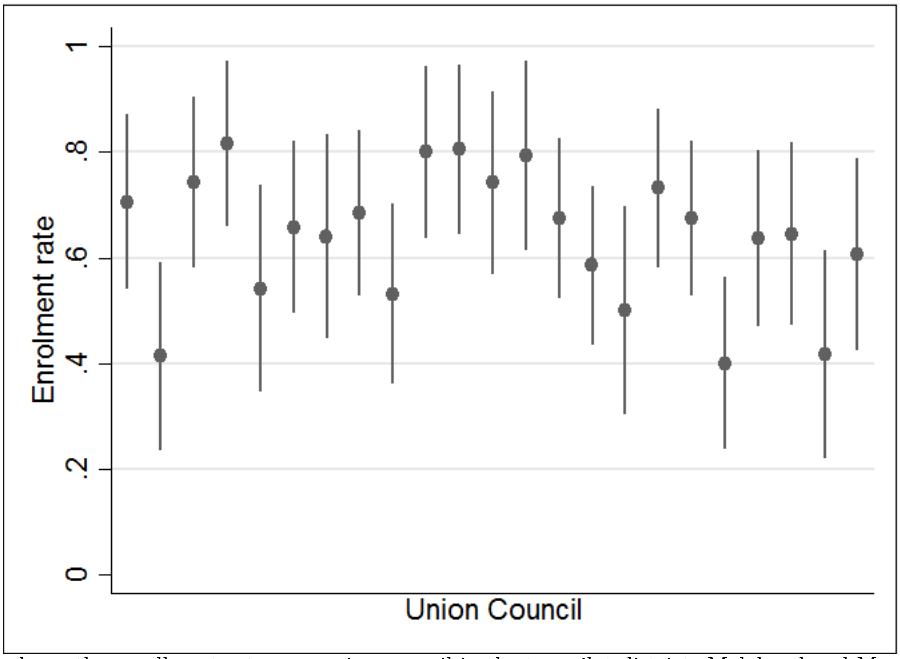
[18] We see three factors that are important here: Non-random targeting (e.g. due to infrastructure or social status), non-random acceptance of the card (e.g. due to lack of education or trust in the government), and non-random awareness of having received a card (e.g. due to low valuation or knowledge about insurance).

[19] Union councils are administrative units between the district and village level, which also served as survey clusters (second stage sampling unit).

[20] The selection of baseline covariates to be included in the propensity score estimation is an important step in our econometric approach. As Caliendo and Kopeinig (2008) note, omitting important variables can increase the bias in the estimates, which suggests including as many covariates as possible. However, over-parameterized models suffer from a lack of common support, which potentially increases the variance of the propensity score estimate. To balance the risk of bias and variance, we follow the procedure described in Imbens and Rubin (2015) (chapter 13) as summarized in Appendix 2-1.

[21] We concentrate on estimation of average treatment effects on the treated, hence we need not drop control observations for whom there is no match in the treatment group.

Figure 1.2: Enrolment rates in main sample of eligible households across 24 union councils



► *Note*: This figure shows the enrollment rates per union council in the two pilot districts Malakand and Mardan.
► Sample: Household-level PSM sample (panel, N=795).
► Source: Endline survey (2017).
► Union councils are geo-administrative units two levels below the district which served as sampling clusters (second stage sampling unit).

bootstrapped standard errors (9,999 repetitions).[22] Note that we bootstrap the whole process of estimating propensity scores, imposing common support, matching observations, and estimating effects. To check robustness, we estimate various specification tests (clustering standard errors on household-level, excluding oversampling around the cut-off and using a smaller, pre-deterimend set of covariates for propensity scores). Our results are robust regarding these different specifications, as shown in Table A.6 in the Appendix.

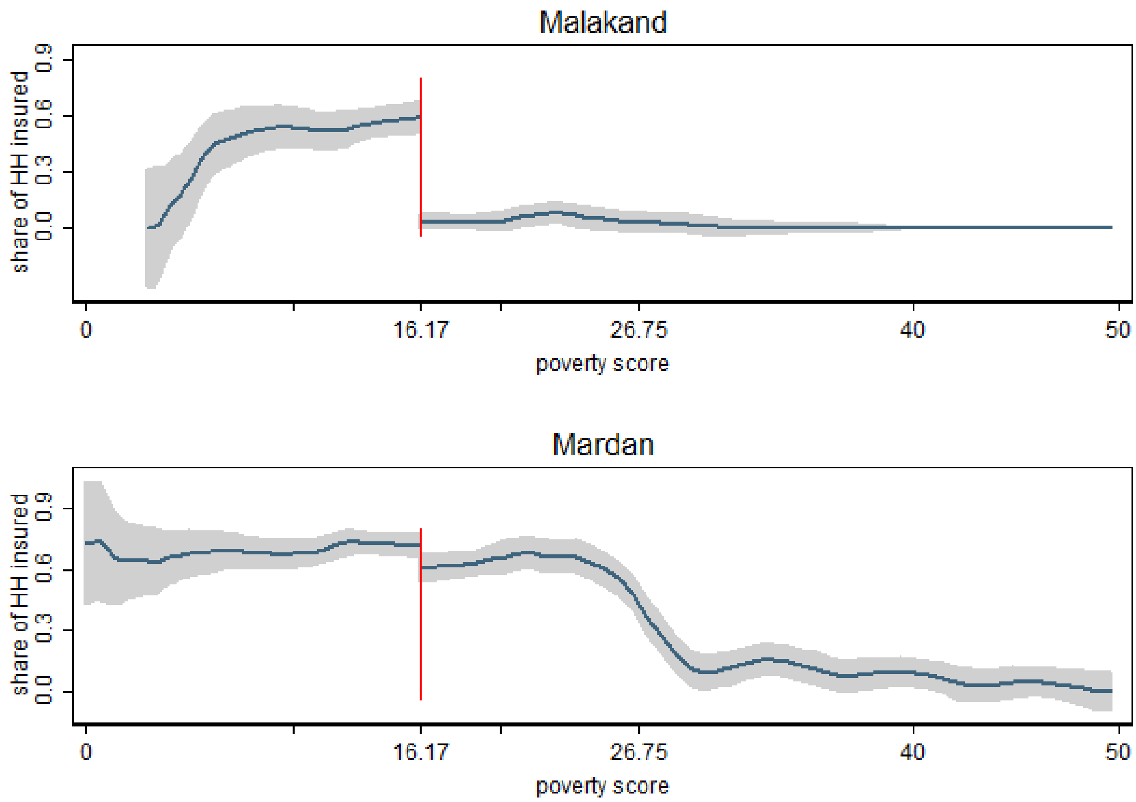
4-2 Regression discontinuity design (RDD)

We exploit the fact that there exists a pre-defined poverty cut-off score which exogenously determines program eligibility, creating an ideal set-up for an RDD approach. Figure 1.3 depicts the self-reported insurance status by poverty score using local polynomial smoothing in both considered districts. In Malakand, there is a large and significant drop in insurance enrollment at the cut-off. This drop is smaller in Mardan due to a pre-mature roll-out of the second phase, which led to enrollment of households with poverty scores between 16.17 and 26.75 in this district three months prior to our endline survey. The figure displays the self-reported insurance status, hence also including enrollment under the second phase. Since our main outcome of interest, the usage of inpatient care, relates to a period of twelve months, it is more appropriate to consider households covered under the second phase as (largely) uninsured.

[22] Whereas Abadie and Imbens (2008) show that bootstrapping is invalid for nearest-neighbor matching, they anticipate that the bootstrap is valid for kernel-based matching (which we use) due to its asymptotic linearity.

If anything, this should lead to a slight downward attenuation of the estimated affect.[23] In our main model specification we hence calculate intention-to-treat effects using a sharp RDD design with treatment determined by the poverty score only.

Figure 1.3: SHARE OF HOUSEHOLDS INSURED (SELF-REPORTED) BY POVERTY SCORE



▶ *Note:* This figure shows the average insurance rate, conditional on the poverty score. The solid blue lines and shaded grey areas are the predicted values and associated 95%-confidence intervals, respectively, based on local mean smoothing. The red vertical line indicates the cut-off score of 16.17.
▶ Sample: Household-level RDD sample (panel, $N_{Malakand} = 617$, $N_{Mardan} = 1,232$).
▶ Source: Endline survey (2017).

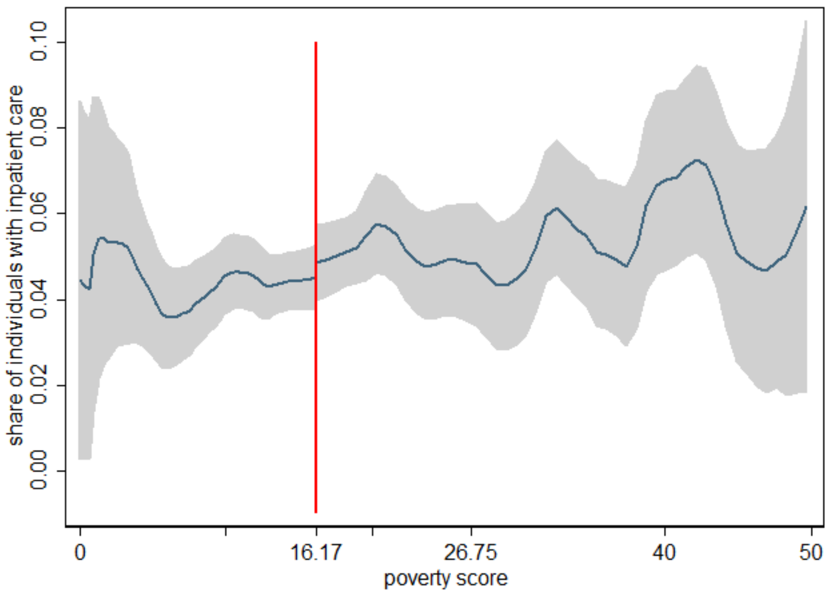
We calculate local linear regression models to the left and right of the cut-off score using a triangular kernel, where the bandwidth is estimated to minimize the mean squared error as suggested in Calonico, Cattaneo, Farrell, and Titiunik (2017). We provide standard errors using heteroskedasticity-robust nearest neighbor variance estimators as provided by the Stata command rdrobust by the same authors. To check robustness, we estimate various specification tests (clustering standard errors, including covariates, fuzzy design, local constant regression models, bandwidth estimators suggested by Imbens and Kalyanaraman (2012)). Our results are robust regarding these different specifications, as shown in Table A.7 in the Appendix.

The key assumption for the internal validity of the RDD approach is that the distributions of potential outcomes are smooth around the cut-off score. As this assumption is not directly testable, it is often instead tested whether treatment assignment is as good as random in

[23] For a quick back-of-the-envelop calculation, note that the effect in our model is reduced approximately by the average time the control group was covered (estimated as 2/12 months) times the share of recently insured individuals above the cut-off (0.4 across both districts) over the share of insured below the cut-off (0.67), so by around 10%.

some window around the cut-off, e.g. by applying the McCrary test or showing continuity of covariates. As-good-as-random-assignment is a sufficient, but not necessary condition for the continuity of potential outcomes. Our running variable, the poverty score, however is not randomly assigned, but instead constitutes a function of sociodemographic and economic covariates, albeit the precise functional form is unknown to us. As this functional form could well be non-smooth, we believe less in balancing checks around the cutoff, specifically not if they involve covariates used to construct the poverty score. Instead, we test the (in our view more appropriate) assumption of smooth potential outcomes by testing for discontinuities in outcomes at baseline. In line with the RDD assumption, we find no significant effects in any of the outcomes of interest, as illustrated in Table 1.5 and Figure 1.4. That is, whereas the unknown functional form might cause discontinuities in individual covariates used to construct the score, this does not translate into discontinuities in outcomes at baseline. We believe this continuity of baseline outcomes to be sufficient evidence for the continuity of potential outcomes around the cut-off, and hence the validity of the RDD approach. Nevertheless, we include the covariates age, gender, and wealth in our regressions using the method proposed by Frölich and Huber (2019). With this, we find point estimates to be slightly larger in magnitude. Since inference remains however unaffected, we provide the more conservative estimates excluding covariates as main results and estimates controlling for covariates as robustness checks in Table A.6 in the Appendix. Further pseudo-effect calculations on those covariates not used for calculating the poverty score give insignificant results, see Table A.9 in the Appendix.

Figure 1.4: Inpatient care by poverty score at baseline



► *Note:* This figure shows the share of individuals with a case of inpatient care by poverty score, using local polynomial smoothing. Red vertical line indicates the cut-off score of 16.17.
► Sample: Member-level random sample (panel, N=12,863).
► Source: Baseline survey (2015).

Additionally, note that the poverty score was initially derived to determine eligibility to a nation-wide social program, which includes among other benefits an unconditional cash transfer. In fact, 84% of eligible households in our panel claim to have received transfers from the BISP

Table 1.5: Pseudo-effects on inpatient care consumption at baseline

Outcome	RDD β_{ITT} (1)	S.E. (2)
Individual outcomes		
Usage of inpatient care	**0.001**	0.008
N (left/right of cut-off)	*3,217/2,557*	
conditional on usage of inpatient care		
More than one admittance	**0.006**	0.080
N (left/right of cut-off)	*134/125*	
Usage of private versus public hospitals	**0.021**	0.079
N (left/right of cut-off)	*161/132*	
Household outcomes		
Neglected health care	**-0.011**	0.030
N (left/right of cut-off)	*418/361*	

- *Note*: This table shows results from RDD estimates on baseline variables, main outcome variables on the left, the different statistics on top.
- Sample: RDD sample (panel, varying N).
- Source: Baseline survey (2015).
- Columns (1) and (2) show the coefficient and standard errors for the pseudo-intention-to-treat effect for households just below the cut-off score, estimated using a sharp regression discontinuity design. Estimated using local linear regression models with a triangular kernel and bandwidth estimated to minimize the mean squared error; S.E. as proposed in Calonico, Cattaneo, and Titiunik (2014) (corresponding to our main model specification). Reported sample size refers to observations within selected bandwidth.
- The statistical significance is given as follows: * indicates p < 0.1, ** p < 0.05, and *** p < 0.01, with the null hypothesis being a zero effect size.

program. However, the transfer was small (10% reporting 1,000 PK and another 85% reporting 1,500 PKR) and, most importantly, 80% of households claim to have received these transfers already at baseline. The continuity of baseline outcomes however suggests that this national social program does not confound our analysis.[24]

5 Results

5-1 Main results

Our main outcome variables concern the use of inpatient care. In our endline survey, we asked for each household member separately whether that member experienced a case of inpatient

[24] Additionally, the national social program might have created incentives for self-sorting into treatment when the poverty scores were assigned in 2010. However, the density test proposed by McCrary (2008) fails to detect a significant discontinuity, as illustrated in Figure A.6 and Figure A.7 in the Appendix, which suggests that there was no systematic manipulation of the assignment variable.

care in the past twelve months (admittance to hospital). If answering affirmatively, we also asked how often the individual was admitted to hospital within that timeframe, and what type of hospital she visited (private versus public). Furthermore, on household level we asked whether any household member faced an accident or illness where inpatient care was considered but not sought within the past twelve months (neglected health care). For all these four key outcomes, we estimate the effects of providing insurance coverage using both, Propensity Score Matching (PSM) and a Regression Discontinuity Design (RDD).

Table 1.6 contains the results from the PSM and the RDD estimations. For example, in the first line the outcome considered is whether an individual has sought inpatient care in the past twelve months. The mean among the matched control group in the PSM sample is 0.059 and we estimate a negative and insignificant coefficient of -0.002 in the PSM estimation, with a standard error of 0.011. The sample consists of 2,526 uninsured and 3,638 insured household members. Our RDD estimation also yields a coefficient of -0.002 with a standard error of 0.007, where we rely on 3,118 observations below and 2,526 observations above the cut-off. Note that the reported sample size refers to the area of common support (PSM sample) and the observations within the selected bandwidth around the cut-off score (RDD sample) respectively, implying that these numbers change across regression specifications.

Table 1.6: Effects on inpatient care consumption

Outcome	Control Mean (1)	PSM β_{ATT} (2)	PSM S.E. (3)	RDD β_{ITT} (4)	RDD S.E. (5)
Individual outcomes					
Usage of inpatient care	0.059	**-0.002**	0.011	**-0.002**	0.007
N (uninsured/insured, left/right of cut-off)		*2,526/3,638*		*4,336/3,227*	
conditional on usage of inpatient care					
More than one admittance	0.195	**0.022**	0.077	**0.062**	0.054
N (uninsured/insured, left/right of cut-off)		*107/212*		*255/174*	
Usage of private versus public hospitals	0.383	**0.068**	0.075	**0.237*** **	0.073
N (uninsured/insured, left/right of cut-off)		*101/202*		*186/131*	
Household outcomes					
Neglected health care	0.062	**0.007**	0.026	**0.001**	0.022
N (uninsured/insured, left/right of cut-off)		*277/461*		*581/466*	

- *Note*: This table shows our main results, the effect of free hospitalization insurance on inpatient care consumption. Outcome variables on the left, different econometric models and statistics on top.
- Samples: Member-level and household-level PSM and RDD samples (panel, varying N).
- Source: Endline survey (2017).
- Column (1) displays the mean for the matched controls of uninsured, but eligible households (poverty score below 16.17). Columns (2) and (3) show the coefficient and standard error for the average treatment effect on the treated, estimated using propensity score kernel matching. Columns (4) and (5) show the coefficient and standard errors for the intention-to-treat effect for households just below the cut-off score of 16.17, estimated using a sharp regression discontinuity design.
- *Note on PSM:* S.E. are derived by bootstrapping the whole process of estimation of propensity scores, restricting the sample to common support, matching, and ATT estimation. Unit of clustering is the union council. Number of bootstraps: 9,999. Reported sample size refers to area of common support (overall sample size: 795 households with 6,007 members). Table A.6 in the Appendix contains results of various robustness checks. Inference remains unchanged across various alternative specifications.
- *Note on RDD:* Estimated using local linear regression models with a triangular kernel and bandwidth estimated to minimize the mean squared error; S.E. as proposed in Calonico et al. (2014). Reported sample size refers to observations within selected bandwidth (overall sample size: 1,842 households with 12,862 members). Table A.7 in the Appendix contains results of various robustness checks. Inference remains unchanged across various alternative specifications.
- The statistical significance is given as follows: * indicates p < 0.1, ** p < 0.05, and *** p < 0.01, with the null hypothesis of the two-sided test being a zero effect size.

Despite the large number of observations at our disposition, we find no significant effects of the program on the usage of inpatient care, neither averaged across all treated (PSM) nor locally around the cut-off (RDD). Even when accounting for clustering effects, standard errors are limited to one percentage point, such that we would have detected effect sizes of less than two percentage points as significant (one third of the control mean). In other words, we can exclude short-term transformative changes in seeking hospitalization in our sample.

Among individuals who reported a case of inpatient care, we also look at the share of individuals with more than one stay at a hospital and also find no effect here. As we did not find an effect of the program on the probability of using any inpatient care before, we believe in the validity of this result, even though the sample restriction to those with inpatient care might in principle be endogenous. Furthermore, note that our sample size is much smaller here.[25]

We also do not observe a decrease in the share of households with neglected health care. Again, precision of the coefficients is limited, but both the PSM as well as the RDD point estimates are very close to zero. Note that 90 percent of households reporting a case of neglected health care stated that this was because they could not afford the cost of treatment in a hospital, suggesting that a functioning insurance scheme could have had an impact on this variable.

Suggestive evidence in line with these null effects also comes from households with childbirth. Given that the insurance explicitly covers maternity care, we would expect a particularly strong increase in the usage of professional assistance during childbirth in these households. Unfortunately, there are too few childbirths in our sample to run a proper matching procedure, but a simple comparison of beneficiary groups in the PSM and RDD samples does not reveal any significant differences.[26]

While the quantity of inpatient care consumed seems to remain largely unchanged, usage patterns may nevertheless have changed. Specifically, we find a significant increase in the usage of private versus public hospitals in the RDD estimation. This result is robust against different bandwidth specifications, as illustrated in Figure 1.5.

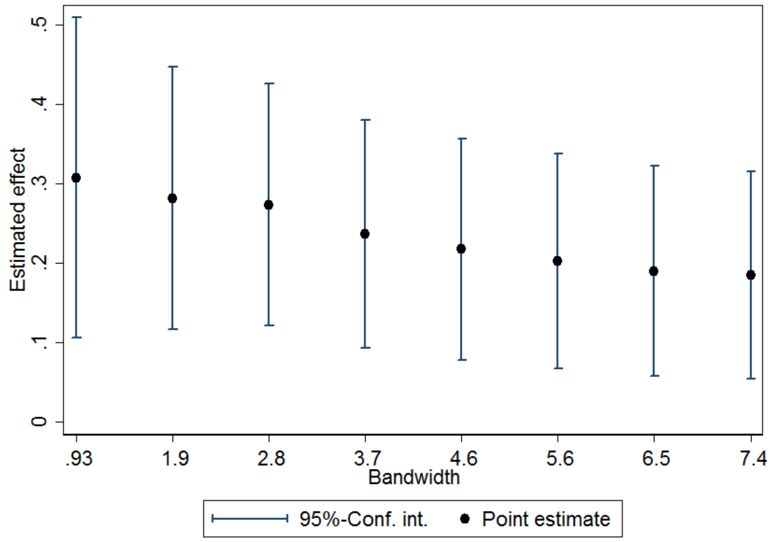
In addition, the effect of 6.8 percentage points calculated in the PSM estimation is, albeit insignificant, sizeable and in the expected direction, increasing the share of individuals visiting a private instead of a public hospital by 18.06%. This result is in line with administrative data: In their progress report for January to June 2017, the consultancy supporting the program on behalf of the KfW Development Bank notes that 95.59% of admissions in the two districts were registered in private hospitals (Oxford Policy Management 2017).[27] We draw further descriptive evidence from a separate section of the questionnaire, where we asked households whether they have used the card, at what type of hospital, and whether this was the first time they visited that facility. Among respondents who used their card at a private facility, 79.45% visited this

[25] To avoid overfitting, we therefore repeat the calculation of propensity scores for this subsample and include only linear terms and no interaction terms in the estimation model.

[26] In our PSM sample, we observe only 35 uninsured and 80 insured households with childbirth, and this sample size is not sufficient to ensure common support for PSM estimation. Regressing the use of professional assistance during childbirth on the insurance status among the PSM sample yields no significant result. We have 113 cases of childbirth within a 2-points interval around the cut-off score. Yet, the RDD estimation also does not find a significant effect on the usage or professional assistance at childbirth.

[27] At the time of our study, the program had empanelled seven private and four public hospitals in the two districts, and for each private facility there is one public facility in immediate proximity (Oxford Policy Management 2017).

Figure 1.5: RDD estimates for use of private versus public hospitals, different bandwidths



▶ *Note*: This figure shows the point estimates and confidence intervals for the impact of the insurance on the use of private versus public hospitals for different bandwidths and accounting for clustering on household level. Main specification uses bandwidth of 3.71.
▶ Sample: Member-level RDD sample with case of inpatient care (panel, varying sample size depending on bandwidth).
▶ Source: Endline survey (2017).

facility for the first time, compared to only 36.67% among public-facility card users.

The shift from public to private hospitals constitutes an improvement of health for the beneficiaries, if private hospitals provide better quality of care. However, whereas public hospitals are hardly monitored, private hospitals do not even register, rendering it notoriously difficult to measure quality of care.[28] Though private hospitals seem to perform better regarding governance and resources, it is unclear whether this transforms into better health outcomes, as private hospitals may overtreat common diseases while referring difficult cases to public tertiary hospitals. Therefore, we restrict ourselves to subjectively perceived quality of care. Evidence that patients associate higher quality of care with private hospitals comes from our endline survey, where 65.82% of respondents in our PSM sample rather agreed than disagreed with the statement that private facilities provide better quality of service than public facilities.[29] Most importantly however, as we have laid out in Section 3, wealthier clients are significantly more likely to visit private hospitals. Specifically, 17.93% of individuals with a case of inpatient care in the lowest wealth quintile visited a private hospital at baseline, whereas that share increases to 44.87% in the highest wealth quintile. We reasonably assume that individuals would not be willing to pay higher prices in private hospitals if these were not perceived to provide better care. Therefore, we associate the observed behavior change in provider choice caused by the

[28] A recent assessment of hospitals in the province of KP led by the Asian Development Bank paints a rather daunting picture of health care quality, listing among other challenges political interference and corrupt practices, serious lack of space, workforce, and drug supplies, as well as issues related to infection control (ADB 2019). The review comprised 37 hospitals, including two private ones, and while this is hardly a representative review of the private sector, the described governance challenges related to nepotism and corruption are likely to be dominant in the public sector.

[29] Shabbir and Malik (2016) provide further circumstantial evidence by finding patients of private hospitals in Islamabad to be more satisfied than patients of public hospitals.

insurance with an increase in subjective quality of care.

5-2 Heterogeneous effects

Average treatment effects might mask heterogeneity regarding demographic or socio-economic characteristics. We therefore repeat the estimation of treatment effects on our main outcome variable, the propensity to use any inpatient care, for selected subsamples with particularly high health financing needs.[30] We look at female household members, at adults above the age of 16, at members with self-rated health status below median at baseline (i.e., below perfect health), and at households with below median wealth.[31] Table 1.7 contains the results of the subsample analysis. We find no significant effects for any of the four subgroups. Note that the control mean in the overall PSM sample was 0.059, underlining that the subgroups considered here are the high-risk groups.

Table 1.7: Hetereogeneous effects on usage of inpatient care

Subgroup	Control Mean (1)	PSM β_{ATT} (2)	PSM S.E. (3)	RDD β_{ITT} (4)	RDD S.E. (5)
Female household members	0.064	**0.005**	0.013	**0.004**	0.012
N (uninsured/insured, left/right of cut-off)		*1,017/1,740*		*1,633/1,253*	
Adults above 16 years	0.097	**-0.020**	0.022	**-0.021**	0.014
N (uninsured/insured, left/right of cut-off)		*1,051/1,823*		*1,584/1,266*	
Baseline health status below median (< 5)	0.093	**-0.002**	0.017	**-0.000**	0.015
N (uninsured/insured, left/right of cut-off)		*825/1,281*		*1,419/1,108*	
Wealth index below median ($< -.60$)	0.079	**-0.015**	0.023	**-0.002**	0.114
N (uninsured/insured, left/right of cut-off)		*1,065/1,772*		*1,574/ 1,362*	

- *Note*: This table shows heterogeneous effects of hospitalization insurance on the usage of inpatient care. The different subgroups considered are indicated on the left, econometric models and statistics on top.
- Samples: Subgroups of the PSM and RDD samples (panel, varying N).
- Source: Endline survey (2017).
- Column (1) displays the mean for the matched controls of uninsured, but eligible households (poverty score below 16.17). Columns (2) and (3) show the coefficient and standard error for the average treatment effect on the treated, estimated using propensity score kernel matching. Columns (4) and (5) show the coefficient and standard errors for the intention-to-treat effect for households just below the cut-off poverty score of 16.17 estimated using a sharp regression discontinuity design.
- *Note on PSM:* S.E. are derived by bootstrapping the whole process of estimation of propensity scores, restricting the sample to common support, matching and ATT estimation. Unit of clustering is the union council. Number of bootstraps: 9,999. Reported sample size refers to area of common support.
- *Note on RDD:* Estimated using local linear regression models with a triangular kernel and bandwidth estimated to minimize the mean squared error; S.E. as proposed in Calonico et al. (2014). Reported sample size refers to observations within selected bandwidth.
- The statistical significance is given as follows: * indicates p < 0.1, ** p < 0.05, and *** p < 0.01, with the null hypothesis being a zero effect size.

6 Discussion

In this section, we discuss our finding of shifts towards private care without an increase in overall hospitalization. Let us first emphasize that the PSM and RDD approaches meaningfully

[30] Note that for the other outcomes analyzed before, subsample analyses suffer from the limited number of observations.

[31] For these subsamples, we calculate propensity scores based on covariates Z_j, which are selected anew for each subsample from the set of all baseline covariates following the same procedure as for the complete PSM sample, described in Appendix 2-1.

complement each other, because they allow us to look at effects on two different populations (average treatment effect on the treated versus intention-to-treat effect at the cutoff), and thereby provide a more complete picture.[32] Also, they complement each other in overcoming relative weaknesses. For example, PSM relies on comparing households based on self-reported enrollment. We argued that self-reported might be more relevant than official coverage in a context of imperfect rollout and awareness. Nevertheless, both self-reported as well as official coverage might be noisy measures of 'effective' coverage; and it is therefore valuable to have the RDD estimate, which is based on an exogenous eligibility cutoff, to confirm results.

One important aspect in the interpretation of the findings is that the endline survey took place twelve to fifteen months after the distribution of insurance cards. This might be too short for results to materialize, for example, because households might need longer to change behavior, or because hospitals might need longer to set up the required procedures. Whereas we agree that the program likely needed more time to reach its full potential, we do not believe that inertness to change behavior is the main reason for this in this setting.[33] In line with this, absolute claim numbers in our two study districts reach relatively stable levels within the first two to three months of insurance introduction and only increase after the second phase of the program is introduced (around the timing of the endline survey). We illustrate this fact in Figure 1.6, where we plot the number of claims in the two districts respectively as per administrative data of the program. Note that Phase 2 started at different points of time in the two districts and saw a notable increase of enrollment from 21% of the population to 51%.

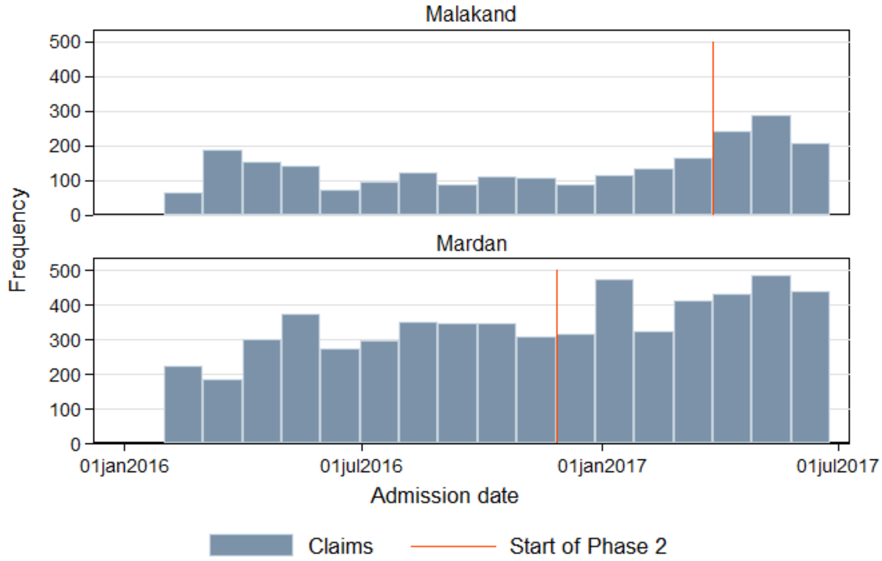
The overwhelming majority of these claims come from private providers, in line with our finding of an increased propensity to visit private rather than public hospitals. A plausible explanation might be that public hospitals were slow to implement required procedures and hence effective coverage was only provided in private hospitals. In fact, private hospitals might have faced greater incentives to participate in the program as the insurance allowed them to attract clients who could previously not afford their services, while at the same time possessing a more flexible governance structure. We therefore cannot reject the possibility of supply side constraints in the public sector driving the observed change in provider choice. This is a relevant consideration as it illustrates the importance of empaneling private health care providers in large-scale programs with potential public capacity limitations. We have to keep in mind, however, that despite the apparent capacity to absorb patients in private facilities, we do not observe an increase in overall inpatient care consumption. This prompts the question whether there are other bottlenecks in the new health insurance scheme.

One possibility is that a lack of information restricts beneficiaries from effectively using the insurance. Given that the government pays full premiums to the insurer based on insurance

[32]Note that we do not focus on the validity of the assumptions underlying our empirical approach here. Those are discussed in Section 4, where we present supporting evidence as far as possible. Specifically, we test the assumptions wherever our data allows, conduct a range of plausibility and sensitivity tests (see Table A.6 and Table A.7 in the Appendix), and run placebo analyses using baseline variables, alternative cut-offs, and control districts (see Table A.9 in the Appendix).

[33]At baseline, 72% of households with a case of outpatient care four weeks prior to our survey received this care at a (public or private) hospitals. Households are hence familiar with hospital visits and the required behavioral change is smaller than it might be in settings with a functioning primary health care system.

Figure 1.6: Absolute number of claims over time (admin data)



► *Note*: This figure shows the absolute number of claims per district over the first one and a half years of the program.
► Source: Administrative data of the SHPI program.
► Red lines indicate the start of Phase 2 in the respective districts (December 2016 in Mardan, April 2017 in Malakand).

cards distributed, and that the costs faced by the insurer are driven by actual usage, the insurer has little incentive to provide comprehensive information to facilitate utilization. Information provided by the regional NGOs might also be incomplete in this principal-agent setting. Our endline survey contains knowledge questions about the insurance program, in particular which treatments are covered (inpatient and/or outpatient) and which hospitals would accept the card (public and/or private). To test whether information is indeed an important factor, we restrict our sample to those households who answered both these questions correctly and repeat the estimation. We display results in Table 1.8 (first line of Panel A). The PSM estimate is negative and insignificant, while the RDD estimate is insignificant as well, and very close to zero. These results do not suggest that the program led to higher utilization among those with better knowledge of insurance details.

Another barrier might be that the program restricts the choice of care providers to specific, empaneled hospitals. At the time of our endline survey, these included two public and three private hospitals in the district of Malakand, and two public and four private hospitals in the district of Mardan. All hospitals are in city centers, hence accessibility remains an issue in rural areas. We measure the distance to these hospitals using GPS data and find a median of 9.6 km. Note that this is the geographic distance calculated from GPS coordinates and likely only proxies accessibility. We also asked respondents about their travel time to the next hospital, including non-empaneled ones, and report a median of 40 minutes. To analyze to which extent distance to hospitals restricted the program's impact, we repeat the estimation of effects for households which live within a below-median distance to a hospital, i.e., within 10 km to an empanelled hospital or within 40 minutes away from any hospital. We display results in Table

1.8 (second and third line of Panel A). Again, we find no evidence of a program impact on overall inpatient service utilization in these subsamples.

An increase in the consumption of inpatient care, however, is only plausible if two conditions are fulfilled. First, the insurance should achieve financial protection, i.e., it should decrease costs of seeking inpatient care. The second condition is that costs of treatment should actually influence inpatient service utilization. In the endline survey, we asked respondents about the cost of treatment born out-of-pocket, which we use to assess the first condition. We estimate the effect of the program on total expenditures and on the individual cost positions for diagnosis and treatment, and medication.[34] We present results in Table 1.8, Panel B.[35] Note that we have only a very limited sample size, as we only consider individuals who reported a case of inpatient care, while at the same time here considering a variable with high variation. Therefore, coefficients are not significant, even though we estimate negative and sizable effects (suggesting a cost decrease of around 30 percent). Additionally, we asked whether those with a hospitalization case experienced sleepless nights due to the related costs. In this case, the coefficient is positive, though insignificant. An explanation for the counterintuitive result on this subjective measure might be an attention bias, given that we previously had asked only insured households about their insurance status and understanding of insurance principles. In summary, our data is inconclusive when it comes to financial protection achieved by the program. It is consistent with a possible decrease in hospitalization costs, though.

Even if the program was successful in decreasing the financial burden of inpatient care, it does not necessarily lead to more utilization. In Section 3, we showed that using inpatient care in general does not increase with financial wealth, suggesting a low risk of moral hazard for an insurer. With higher wealth, however, we observe an increase in private care, which patients often associated with higher quality. In other words, individuals with urgent health problems might visit a hospital irrespective of their wealth. This is consistent with evidence that hospitalization (in contrast to outpatient care visits) is not very sensitive to cost sharing by an insurer (e.g. Finkelstein 2007). Poorer individuals, however, seem to seek care at cheaper public facilities. We compare average costs of public and private hospitalization in our data. Indeed, we see that private care is much more expensive than public care at baseline in our sample of interest (39,000 vs. 23,000 PKR).[36] Interestingly, it descriptively seems like this difference shrinks after the program rollout only for the insured, driven by a strong decrease in private care costs.[37] Given these observations, it may not be too surprising that instead of

[34] We asked separate questions for total costs and individual cost positions to test for possible side payments demanded by the hospital staff. We find no evidence for this. Note that we did not measure opportunity costs such as forgone wages, except for transportation, meals, and accommodation for accompanying family members, which we find to be a negligible portion of total expenditures.

[35] Due to the highly screwed distribution of the quantitative cost variables, we use log values as outcomes.

[36] The diseases treated in public and private facilities at baseline are largely the same in our data, but sample size per disease is too low to comment on statistical significance. A notable exception is that whereas only 5.22% of patients were treated for appendicitis in public facilities, the rate is 19.02% among patients in private facilities.

[37] After the insurance rollout average private care costs are 20,000 PKR and public care costs 13,000 PKR for the insured, while the difference is much larger for the non-insured (30,000 vs. 13,000 PKR). So in particular the relative decrease in costs for private care seems to be larger for the insured at endline. The difference in cost for private care between insured and non-insured was in fact the reverse at baseline (41,000 PKR for the insured vs 35,000 PKR for the noninsured).

an overall increase in hospitalization, we measure a shift towards private facilities as a result of the program. For a dual health system with public and private providers operating in the same market, this is a highly relevant result, as the insurance program might also shape the composition of the market in the long term.

Table 1.8: Evidence on program limitations

Outcome	Control Mean (1)	PSM β_{ATT} (2)	PSM S.E. (3)	RDD β_{ITT} (4)	RDD S.E. (5)
Panel A: Effects on usage of inpatient care in different subsamples (member-level)					
HH with good knowledge on program details	0.069	**-0.018**	0.015	**-0.001**	0.008
N (uninsured/insured, left/right of cut-off)		*2,007/2,313*		*2,863/2,755*	
HH living within 10 km distance to next empaneled hos.	0.066	**-0.014**	0.022	**-0.020**	0.011
N (uninsured/insured, left/right of cut-off)		*1,061/1,881*		*1,861/1,431*	
HH living within 40 min travel distance to next hospital	0.049	**-0.005**	0.012	**0.007**	0.010
N (uninsured/insured, left/right of cut-off)		*1,061/1,419*		*1,713/1,326*	
Panel B: Effects on financial outcomes conditional on inpatient usage (member-level)					
Log out-of-pocket expenditures (PKR, win99)	8.947	**-0.256**	0.310	**-0.266**	0.256
N (uninsured/insured, left/right of cut-off)		*107/212*		*178/136*	
Log of cost for diagnosis and treatment (PKR, win99)	4.833	**-0.261**	0.562	**-0.601**	0.567
N (uninsured/insured, left/right of cut-off)		*107/212*		*207/151*	
Log of cost for medicines (PKR, win99)	8.238	**-0.316**	0.398	**-0.030**	0.344
N (uninsured/insured, left/right of cut-off)		*107/212*		*204/147*	
Sleepless night due to hospital costs (PKR, win99)	0.486	**0.083**	0.095	**0.028**	0.066
N (uninsured/insured, left/right of cut-off)		*107/212*		*235/170*	

- *Note*: This table shows evidence for the discussion on program limitations. Different subgroups (Panel A) or outcomes variables (Panel B) on the left, statistics on top.
- Samples: (Subsamples of) PSM sample and RDD sample (panel, varying N).
- Source: Endline survey (2017).
- Column (1) displays the mean for the matched controls of uninsured, but eligible households (poverty score below 16.17). Columns (2) and (3) show the coefficient and standard error for the average treatment effect on the treated, estimated using propensity score kernel matching. Columns (4) and (5) show the coefficient and standard errors for the intention-to-treat effect for households just below the cut-off score of 16.17, estimated using a sharp regression discontinuity design.
- *Note on PSM:* S.E. are derived by bootstrapping the whole process of estimation of propensity scores, restricting the sample to common support, matching and ATT estimation. Unit of clustering is the union council. Number of bootstraps: 9,999. Reported sample size refers to area of common support.
- *Note on RDD:*Estimated using local linear regression models with a triangular kernel and bandwidth estimated to minimize the mean squared error; S.E. as proposed in Calonico et al. (2014). Reported sample size refers to observations within selected bandwidth.
- The statistical significance is given as follows: * indicates p < 0.1, ** p < 0.05, and *** p < 0.01, with the null hypothesis of the two-sided test being a zero effect size.
- The suffix *win99* indicates that we winsorized the variable at the 99th percentile level.
- Table A.8 shows results for non-logarithmic financial outcomes.

7 Conclusion

Providing free health insurance to a large number of poor households is an intuitive approach to increase health care consumption. The rationale is that high OOP expenditures not only pose a financial risk, but also restrict poor households' access to inpatient care. We analyze the effect of the SHPI, which provided free hospitalization insurance to the poorest 21% of the population in the Pakistani province of KP, on health care seeking behavior. To this end, we apply a propensity score matching approach, comparing insured and uninsured but eligible households, and a regression discontinuity approach, comparing households just above and just below the exogenous poverty cut-off score. While the latter has a high internal validity, it provides estimates of intention-to-treat effects for households around the cut-off poverty score only. In contrast, propensity score matching relies on more restricting assumptions, but provides average treatment effects on the treated. In this sense the two identification strategies complement each other.

In our study, we find that insured households do not increase the quantity of inpatient care consumed and have the same propensity to neglect their health care as uninsured households. Large-scale multi-stakeholder programs like the SHPI naturally face many challenges in implementation, including limited awareness and insufficient empanelment of hospitals. Yet, we find no support in our data that these factors seriously impaired the program's impact. Also, we measure impact only one year after program introduction. Whereas we concur that the program might develop larger impact over a longer period of time, administrative data confirms that the program was largely operational within the considered time period, in particular in private hospitals. To check whether heterogeneity masks effects for some subgroups, we repeat estimations separately for several high-risk groups, but also fail to detect significant increases in inpatient care.

Importantly, we do however observe a sizable increase in the propensity of visiting a private instead of a public hospital. This result is in line not only with administrative data, but also with a larger decrease of reported care costs for insured individuals in private compared to public hospitals. Since patients in Pakistan often consider private hospitals to provide higher quality of care, this is an important and policy-relevant effect of the program, which might thus contribute to a more equitable access to high-quality care. Whether strengthening the private sector to overcome possible supply side constraints in the public sector leads to desirable outcomes in the long run is an open question, though. Given that there are a number of countries with mixed health systems moving towards universal insurance coverage, including India and Indonesia, further research on the long-term effects on public versus private market sectors seems promising.

Chapter 2

Subsidized health insurance in a multi-conflict setting: An impact evaluation in Northern Pakistan

1 Introduction

An increasing number of developing countries implement social health insurance schemes to help poor people overcome financial barriers to health care. Whereas these have shown the potential to meaningfully increase access to health care, people in conflict-affected areas face additional barriers to health care. These might persist even if financial constraints are relaxed and hence reduce the effectiveness of insurance schemes. In addition, the implementation of large-scale health insurance is a challenging task in any country, with potential bottlenecks existing on regulation, demand and supply side levels. These challenges can be difficult to overcome in conflict-affected areas, where trust in government might be weak and threats to everyday peace persist.

In this study, I evaluate a health insurance scheme for the poorest households in Gilgit Baltistan, a geopolitically and culturally complex territory with wide-spread poverty and poor health. Part of the Kashmir region, households here live in harsh high-altitude conditions with limited infrastructure, worsened by the late intensification of a sectarian conflict between Shia and Sunni Muslims, in particular in the capital Gilgit Town. This setting implies two particularly relevant dimensions I consider in my study: the implications of additional physical barriers to health care faced especially by rural households, and, uniquely, the implications of the sectarian setting for program implementation especially in Gilgit Town.

To analyze whether the program reduced financial constraints and increased the use of hospital care among beneficiary households I use two waves of household survey data. Given that program eligibility is determined through an exogeneous poverty cut-off score, a regression discontinuity design is the obvious choice. Despite the high internal validity of the approach, I fail to detect any evidence for the intended effects on key outcome variables, including cost of hospitalization, subjective burden of health cost, use of hospitals, and incidences of neglected health care.

A possible explanation might be that physical constraints to health care might be binding. Since average effects might mask heterogeneity in this respect, I repeat my estimation for population subgroups disaggregated by geographic indicators. I find effects do not differ between rural and urban populations, and even households living in proximity to the best available hospital did not report changes in costs or health care consumption. Whereas the regression discontinuity approach has a high internal validity, its external validity is limited as only households close to the cut-off score are considered. It is therefore possible that positive effects of the program remain undetected if these only materialized for poorer households. Importantly, however, I find various pieces of descriptive evidence for deficiencies in program implementation.

First, on hospital level, administrative data and program reports suggest that private hospitals did not fully participate in the program and that public hospitals engaged in fraudulent behavior. Specifically, on the one hand, private hospitals were hesitant to join the program as they did not trust the government to reimburse their claims. On the other hand, almost half of the official insurance claims from public hospitals filed acute appendicitis as reason for treatment. This is a significant deviance from both, the survey data and claims filed in private hospitals, and indicates fraudulent behavior by the hospitals, which could be either misreporting or malpractice. Overall, the inefficient program implementation on hospital level is in line with households reporting in my survey that their insurance card was not accepted.

Second, my survey shows that the program achieved highest enrollment rates in areas of Gilgit Town which are known to house predominantly Ismaili households. It is also only in these areas that the share of individuals reporting a case of inpatient care increased significantly between base- and endline.[1] This is an important finding given that the consortium of development NGOs, the Aga Khan Development Network (AKDN), in charge of overall program implementation is closely associated with the Ismaili faith. The Ismailis are third largest sect in Gilgit Baltistan, who consider themselves neutral during sectarian tensions between Shia and Sunni Muslims and hence seem well suited to act as conciliatory force (Hunzai 2013). Assuming that decision-making power in the enrollment process largely rested with the Ismaili implementing consortium, the biased distribution of the publicly funded insurance could be considered a form of elite capture.

With these two findings, I contribute to a growing body of literature on state-funded health insurance in developing countries. Whereas studies from China and India (Wagstaff et al. 2009; Prinja et al. 2017) show that large-scale subsidized health insurance can meaningfully increase access to health care and decrease the financial burden for the poor, it is beyond doubt that

[1]Due to the small sample size especially around the cut-off, I unfortunately cannot estimate effects for Ismaili households only using the regression discontinuity appraoch.

implementing such schemes is a difficult task.[2] Therefore, identifying factors that can affect the success of large-scale health insurance programs is crucial. In the case of Gilgit Baltistan, the program seems to have been unable to overcome the challenges posed by sectarian dynamics combined with a weak health care infrastructure, as demonstrated in sectarian preferences in the enrollment process, limited participation of private hospitals, and partially fraudulent behavior at public hospitals.

This study is novel in that I explicitly take sectarian conflict into consideration and show that insurance enrollment might be biased towards people of the same religious affiliation as the implementing NGO, who are generally considered neutral. However, this bias might not translate into sizable impact in health and financial outcomes as the program was largely not operational on the hospital level. In this regard the program is in stark contrast to the very similar program operating in the Pakistani Province of Khyber Pakhtunkhwa (KP), which was also designed and supported by the German KfW Development Bank. In KP however, the program managed to engage a number of private hospitals, who offered services to the insured even when public hospitals were still hesitant. This led to a significant shift of provider choice among insured individuals from public to private hospitals, which households perceive to be of higher quality (Helmsmueller and Landmann (2020)). In contrast, the program in GB, though facing similar absorptive limitations in the public sector, was not able to increase absorptive capacity by including the private sector.

Finally, from a methodological point of view, my study illustrates two additional points: First, to not only measure impact but also identify the bottlenecks in implementation, it is important to supplement rigorous econometric methods with an analysis of administrative data. As such, my study may contribute to the debate on which data should be collected to meaningfully inform policy, as discussed in Gugerty and Karlan (2018). Second, especially in areas where social tensions affect everyday peace, my study illustrates the importance of providing context-specific analyses.[3] To do so, anthropological studies and sociological theory can give valuable insights complementing and inspiring the econometric analysis.

The rest of this paper is organized as follows. In Section 2, I describe the health sector, the program, and the survey data. In Section 3, I provide details on program implementation and related descriptive evidence from administrative data and the survey. In Section 4, I lay out my identification strategy and discuss its underlying assumptions. In Section 5, I provide the estimated effects on health care financing and usage. Section 6 contains the results of the heterogeneity analysis along geographic dimensions. The last section concludes.

[2] Also see Habib et al. 2016 for a recent review.

[3] See Mac Ginty (2014) for a discussion of the concept of everyday peace.

2 Setting, intervention, and dataset

2-1 Culture and health care system

As part of the larger Kashmir region, which remains subject of conflict between Pakistan, India and China, Gilgit Baltistan (GB) is an autonomous territory administered by Pakistan. Furthermore bordering Afghanistan in the North and China to the East, it is a strategic point of interest within a troubled part of the world for various global players (Holden 2019), rendering its governance an inherently difficult task. Figure 2.1 depicts its location in the region. Its estimated population of 1.9 million, who at the time of this study are not allowed to participate in national elections, belong to various linguistic, ethnic, and religious sects (UNPO 2017). More than 80% of these live in rural areas, which constitute of valleys in the Karakoram, Himalaya, Pamir and Hindu Kush ranges and are hence separated by some of the largest mountains of the world. Accordingly, life is characterized by harsh winters and major obstacles to transportation. Given the difficult living conditions, it is little surprising that the health status of the people in Gilgit Baltistan remains low, with an under-5-mortality rate of 91.8 per 1,000 live births, which is even higher than in Sub-Sahara Africa (MICS 2017).

There are only two larger cities, Gilgit Town and Skardu. The former is the region's capital, counting an estimated 260,000 inhabitants (Varley 2015). Shia Muslims form the majority group in Gilgit Town and the area is the only Shia-majority region in Sunni-majority Pakistan (Sökefeld 2015). Although Sunni Muslims constitute the majority on a national level, they are only the second largest population group in GB.

There are repeated outbursts of sectarian violence between the two religious groups, referred to as tensions, which have resulted in a spatial division of the town along sectarian fault lines into Shia and Sunni settlements (mohallas).[4] The third largest religious group are the Ismailis, who perceive themselves and are generally perceived by others to maintain a neutral position during tensions (Hunzai 2013).

Like the rest of Pakistan, GB has a mixed public-private health care system, with the public sector being the main provider of preventive care throughout the region and the major provider of curative services in the rural areas. Public primary health care units are meant to provide eight hours of outpatient services at six days a week for preventive and some curative services (mainly maternal and antenatal care).[5] The five no- or low-cost public District Headquarter Hospitals (DHQs) and 27 Civil Hospitals in the district are in charge of secondary health care (Asif 2017).

Being an autonomous territory, the government of GB assumes responsibility for health planning, main financing, implementation, management, supervision, regulation, as well as medical

[4] Ali (2010) describes how Shia-Sunni sectarianism manifests itself beyond violent outbreaks in everyday forms, including in secular education, and points out the subtle ways in which sectarian anxieties are experienced and reproduced.

[5] The district's public primary health care system consists of 2 rural health centers, 15 basic health units, 93 mother and child health centers, 154 sub health centers, and 190 rural dispensaries for the total area of GB.

Figure 2.1: Location of Gilgit Baltistan within Pakistan

► *Note:* The figure shows the location of the territory of GB within Pakistan, the district of Gilgit, and Gilgit Town.
► Source: Benz (2016), changed highlights, available via license: relative Commons Attribution 4.0 International

education. While the role of the federal government is limited also in other regions of Pakistan,[6] the conflict-ridden territory faces additional challenges due to sectarian hostilities extending to hospitals. As Varley (2010) describes, in 2005, the Shia religious leader was assassinated by Sunni militants. Before succumbing to his injuries, he was treated in the DHQ in Gilgit Town, which is situated in a Shia neighborhood. In the city, members of the Shia sect responded by targeting Sunnis, including patients and physicians at the DHQ. Given the resulting challenges of the Sunni community to safely access medical facilities, two Sunni hospitals were subsequently constructed, marking the beginning of what effectively resulted in a sectarian medical infrastructure. Varley (2015) reports that Shia religio-political flags and graffitis were still found around the DHQ in 2013. Inequity in access hence is a key concern in the town. Figure B.1 in the appendix illustrates the sectarian separation of the Town.

In addition, according to a health facility assessment of 2012, lack of qualified and specialized staff is a major challenge faced by the hospitals, with few people willing to serve in the far-flung areas where difficult living conditions persist. The quality of services is further threatened by the lack of equipment, drugs and supplies, insufficient waste management and infection control, and suboptimal general management practices. While private and NGO-run hospitals exist, these are unregulated and expensive, and their quality unknown (Contech International Health Consultants 2012, Asif 2017).

[6]In all provinces, health care fell into the responsibility of the provincial health departments after the Eighth Amendment to the Constitution of Pakistan in 2011. Averaged over all provinces, the federal government assumed only around one third of health expenditures, mainly for key national programs such as vaccination campaigns (Jalal and Haq 2014).

2-2 The *Sehat Hifazat Social Health Program* in Gilgit Baltistan

To increase access to medical health care and reduce the associated financial burden, the Government of GB started implementing the Sehat Hifazat Social Health Program in the pilot district Gilgit in 2016.[7] Via this program, the Government provides free health insurance to the poorest 21% of the population, corresponding to 5,350 households or 35,671 individuals. Of the approximately one million Euro in costs, 75% are provided by the German Government through the KfW Development Bank, whereas the Government of GB contributes 25% of funds.[8]

Following a competitive tendering process, the Government charged the Aga Khan Development Network (AKDN) with the implementation of the program, a consortium comprised of Aga Khan Foundation Pakistan, Jubilee Life Insurance, Aga Khan Rural Support Programme, and Aga Khan Health Service. The consortium is in charge of awareness raising, beneficiary enrollment, gatekeeping, and claim processing. In exchange and after verifying enrollment in spot-checks, the Government paid the full amount of annual premiums to the insurer.

The insurance is household-based, extending coverage to seven household members. To determine eligibility of households, the Government uses a pre-existing poverty score with the eligible poorest 21% of the district population having poverty scores below 16.19. AKDN acquired the poverty data for these households in February 2016 and by December 2016 had distributed Sehat Hifazat cards to 90% of the eligible households. Six different local support and civil society organizations furthermore supported the roll-out and provided information on the program to the eligible households, using wall chalking, advertisement on cable network and local newspapers, and ground level promotional activities.

With the exception of maternity care, for which also outpatient treatment is included, the insurance covers only hospitalization services and day care surgeries upon referral, up to an annual limit of 25,000 PKR per person or 175,000 PKR per household. The scheme furthermore covers transportation charges up to 1,000 PKR and costs for medicines for five days after discharge as prescribed by a doctor.

AKDN also selected, negotiated with, and finally empaneled private, NGO, and public sector hospitals, where cashless service delivery begun in early 2017. Whereas initially 21 hospitals and primary health care facilities were contacted, many were reluctant to take part in the program due to limited resources and mistrust towards the cashless payment mechanism. Note that this experience is akin to the project beginning in the neighboring province Khyber Pakhtunkhwa, where a similar insurance scheme was implemented. However, here, private hospitals became eager to join the scheme after the reimbursement scheme proved successful in the first few

[7] At the time of my study, the autonomous territory of GB was administratively divided into ten districts (see Figure 2.1). The district of Gilgit has the smallest area but is the most populous district. Subsequent to this study, the program was extended to other areas and then integrated with a national insurance scheme, called the Prime Ministers Scheme.

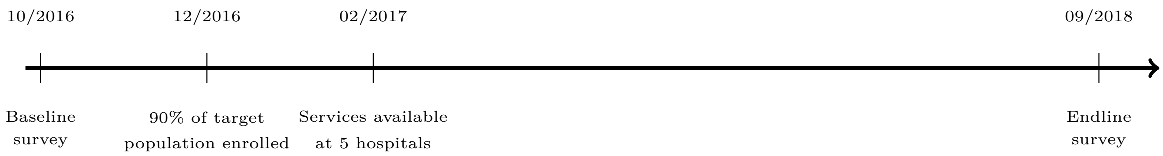
[8] Further program information is available on the program's website: http://sehathifazat.gog.pk/index.php/aboutprograme/ accessed on 04/20/2020.

months. Public hospitals on the other hand remained reluctant to implement the scheme, but were later forced to participate using Government influence (Helmsmueller and Landmann 2020). These mechanisms did not take effect in Gilgit Baltistan, suggesting limited program ownership or weak governance. In the end, AKDN signed agreements with five hospitals, of which two public, two private and one NGO-run.[9] These hospitals established counters for card verification and trained staff on the program details.

2-3 Survey data

The existence of an external poverty cut-off score determining eligibility to the program makes this an ideal set-up for a RDD design. For this, I rely on two waves of household survey data from the treatment district Gilgit and a control district Ghizer, collected in October 2016 (baseline) and September 2018 (endline). Figure 2.2 illustrates the timeline of the two survey waves relative to some program milestones.

Figure 2.2: Timeline



The district of Gilgit further subdivides into eleven union councils, including Gilgit Municipal Area (Gilgit Town). I applied a multi-stage clustered stratified sampling, where I first selected Gilgit Town and sampled villages from two randomly selected rural union councils in Gilgit district. Given that the original evaluation design also foresaw a difference-in-difference design, I also sampled villages from two rural union councils in Ghizer district, where the program was not implemented at the time of this study and which I use to check model specification. Subsequently, I selected households stratified according to poverty score from the original poverty census of 2010, with oversampling around the cut-off (total of 320 households in Gilgit Town, 160 in rural areas of Gilgit district and of Ghizer district respectively). This implies that the sample is not representative for Gilgit.[10]

[9]Specifically, the participating public hospitals are DHQ Hospital Gilgit and City Hospital Gilgit; the NGO-hospital is Aga Khan Medical Center; the private hospitals are Family Health Hospital and Sehat Foundation Hospital.

[10]Initially, the evaluation foresaw an RDD in Gilgit Town, and an additional difference-in-difference strategy in rural areas. Accordingly, in rural Gilgit and Ghizer, I randomly selected 10 households per village, then additionally two households below the cut-off at random, and two households each below and above the cut-off, but closest to the cut-off (non-random). However, the RDD in Gilgit Town would not require random sampling. At the same time, it was not clear whether the cut-off score would be raised above 16.19 during the study period. Therefore, I randomly selected households below the cut-off score, and then randomly selected households in poverty score strata of width 1, between 16.19 and 22.19. For logistic reasons, I then grouped households according to settlement and in some of these, additional households with higher poverty scores were sampled to ensure a sufficient number of observations per settlement. With the BISP scores having been assigned to households in 2010, there is significant undercoverage of the city in my sampling frame. Additionally taking into account the difficulties in identifying sampled households and the relatively high attrition in Gilgit Town, I cannot claim to have a representative sample of the city.

I use endline data collected two years after baseline and 20 months after program start, where the same households were interviewed. The overall attrition rate lies at 7.93% and hence within acceptable range. The main reason for attrition is noncontact as interviewers were unable to locate households despite GPS data from baseline. At 9.81%, attrition rate is highest in Gilgit Town, where household mobility is likely greater than in rural areas. I restrict the sample to households and its members with poverty scores below 32.38 so as to consider the same interval of poverty score left and right of the cut-off. I furthermore drop less than twenty households due to quality concerns. My final panel dataset hence comprises 531 households (401 in treatment group) and, for questions asked regarding each household member separately, 4,195 individuals (3,230 in treatment group). Table 2.1 summarizes the net sample sizes attained in the different areas.

Table 2.1: SAMPLING AREAS WITH SAMPLE SIZES

Area	N (households)	N (individuals)	Program implemented	Location
Gilgit Town	278	2,222	yes	urban
Rural Gilgit	123	1,008	yes	rural
District Ghizer	130	965	no	rural
Total	**531**	**4,195**		

▸ *Note*: This table shows the different sampling areas with respective sample size. Overall sample size in treatment areas is hence 401 households and 3,230 individuals (treatment sample).

The survey covered questions on the socio-demographic characteristics of the household and its members, in particular questions on asset ownership, which I combined into an indicator (hereafter called wealth index) using a principal component analysis. The survey focused on the usage of health care services, including use of outpatient and inpatient care, where the latter was measured for each household member individually. The questionnaire also contained a number of questions on health care financing, including quantitative measurements as well as subjective attitudes.

2-4 Baseline characteristics

Table 2.2 (Panel A and B) presents an overview of the variables considered with respective baseline statistics for my sample, separately for eligible and ineligible households in the treatment district. For example, the first line says that the average poverty score among eligible households is 14.12 with a standard deviation of 2.78 (Columns (1)-(2)), whereas it is 19.83 with a standard deviation of 3.09 among ineligible households (Columns (3)-(4)). This difference is significantly different from zero with a p-value of 0.00 (Column (5)). Overall, the variable ranges from 0 to 32.17 (Column (6)-(7)).

Panel (A) illustrates that households in my sample are indeed poor, with an average household income among eligible households of 165 USD and a maximum of under 800 USD.[11] Among eligible households, almost a third contains only members without primary school education. Note that the average household size is around 9, whereas the insurance covers a maximum of eight members. Although around two third of my sample live in the urban areas of Gilgit Town, the reported average distance to the next hospitals is over four hours among eligible households. This is however driven by households in the rural union council of Charkarkote. If excluding these households, the average reported distance to hospitals is 60 minutes.

Panel B illustrates that health care financing is indeed a serious concern for many households. Around three quarters of households state health shocks as the main financial risk, followed by loss of job or wages, which around 20% perceive as the main risk. This is in line with reported costs of the last inpatient treatment averaging the mean household income of two months. Nevertheless, health care usage rates are high: 13% of individuals report a case of inpatient care within the last twelve months, corresponding to almost 75% of households having at least one member admitted to a hospital in that time frame. More than a third of all individuals have visited a hospital more than once within twelve months. In addition, 74% of eligible households report a case of outpatient care within the previous four weeks, and 35% claim that they did not visit a hospital despite being recommended to do so. Almost all of these households report financial constraints as reason for neglecting health care.

Comparing eligible and ineligible households, on average, eligible households in my sample are less likely to live in urban areas, contain more family members, are more likely to lack a member who completed primary education, have a smaller monthly income and lower wealth, and live further from hospitals. Eligible households in my sample are also more likely to report difficulties in finding money for health care and worries about the cost when a household member was hospitalized. Despite using more outpatient care and similar levels of inpatient care, eligible households are also more likely to report a case of neglected health care in the past six months. While eligible households report higher costs for inpatient care, this difference is insignificant.

[11] Exchange rate from June 30,2016: 1 USD = 104.63 PKR.

Table 2.2: Baseline characteristics in treatment district

Variable	Below cut-off		Above cut-off		p-value	Overall	
	μ_b	Std.Dev	μ_a	Std. Dev.	H_0: $\mu_b = \mu_a$	Min.	Max.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Covariates							
Poverty score	14.12	2.78	19.83	3.09	0.00	0	32.17
Urban	0.63	0.48	0.74	0.44	0.02	0	1
Household size	9.11	2.73	8.60	3.09	0.09	1	23
Avg. Monthly HH income (PKR, win99)	17,246	11,719	19,934	14,042	0.04	0	76,000
Wealth index	-0.36	2.47	0.66	2.26	0.00	-5.83	7.65
Reported distance to next hosp. (min., win99)	249.64	591.48	115.31	376.94	0.01	1	2,200
Having heard of insurance	0.14	0.34	0.13	0.33	0.77	0	1
No HH member completed primary school	0.28	0.49	0.14	0.35	0.00	0	1
Some HH member with higher educ.	0.34	0.47	0.42	0.49	0.10	0	1
N	*170*		*231*				
Age	21.11	17.46	22.25	17.33	0.07	1	100
Female	0.47	0.50	0.48	0.50	0.38	0	1
N	*1,408*		*1,822*				
Panel B: Outcomes							
I. Health care financing							
Fin. Satisfaction ≤ 5 (scale 1 to 10)	0.09	0.29	0.27	0.45	0.00	0	1
Citing health shock as main risk	0.77	0.42	0.71	0.46	0.15	0	1
Dif.ty finding money for health care ≤ 8 (scale 1 to 10)	0.58	0.49	0.41	0.49	0.00	0	1
N	*170*		*231*				
Total cost of last treatment (PKR, win99)	34,487	48,368	40,327	56,582	0.26	0	300,000
Very worried re. costs of inp. care	0.60	0.49	0.38	0.48	0.00	0	1
N	*181*		*256*				
II. Health care usage							
Case of inpatient care	0.13	0.33	0.14	0.35	0.32	0	1
N	*1,408*		*1,822*				
Neglected health care	0.35	0.48	0.26	0.44	0.06	0	1
Case of outpatient care	0.74	0.44	0.65	0.48	0.06	0	1
N	*170*		*231*				
Use of private hospital	0.15	0.36	0.31	0.46	0.00	0	1
More than one admittance	0.36	0.48	0.38	0.49	0.74	0	1
N	*181*		*256*				
Panel C: Program implementation							
Having heard of the program	0.61	0.49	0.23	0.42	0.00	0	1
Insured by program	0.49	0.50	0.08	0.27	0.00	0	1
N	*170*		*231*				
Knew that ins. accepted only at empaneled hos.	0.35	0.48	0.50	0.51	0.22	0	1
Knew that only inpatient services covered	0.61	0.49	0.56	0.51	0.69	0	1
Used card at hospital	0.48	0.50	0.50	0.51	0.86	0	1
N	*84*		*16*				

- *Note*: This table contains baseline characteristics of the treatment sample. Variables on the left, statistics on top.
- Sample: Treatment sample (panel, varying N)
- Source: Baseline survey (2016) and endline survey (2018, Panel C only)
- Columns (1) and (2) display means for continuous/shares for binary variables, and standard deviation among eligible households (i.e., with poverty scores below 16.19). Column (3) and (4) display means for continuous/shares for binary variables, and standard deviation among ineligible households (i.e., with poverty scores above 16.19). Column (5) contains the p-value of testing equality of means above and below the cut-off. Columns (6) and (7) display the minimum and maximal value over the whole treatment sample.
- The suffix *win99* indicates winsorizing at the 99th percentile level.

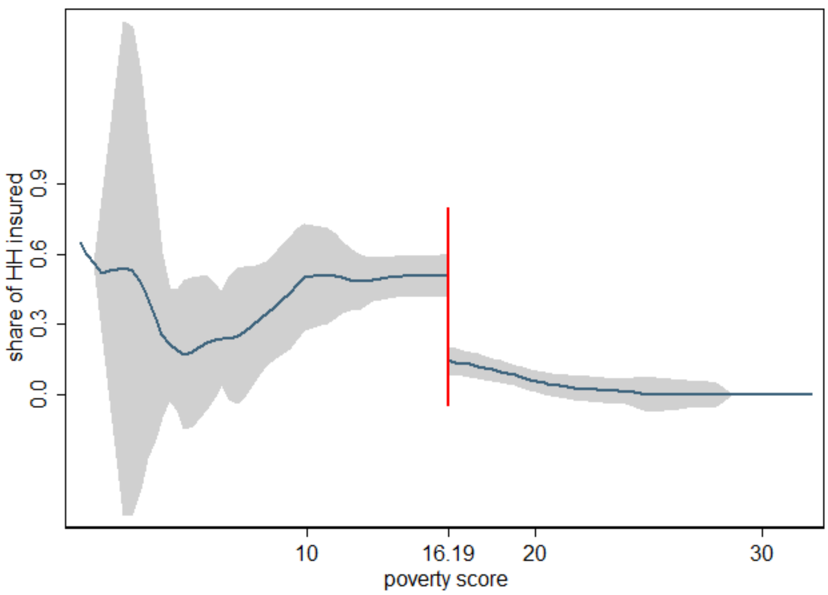
3 Program implementation

3-1 Overall enrollment, awareness, and card usage

In the endline survey, I gathered information on enrollment in and knowledge of the program. Table 2.2, Panel C, contains summary statistics of these, which I describe in the following. From now on, the terms *enrollment* and *insured* describe self-reported measures. Administrative enrollment data is available but could not be matched to survey data.

The average enrollment rate among eligible households in my sample is 49%.[12] Some households with poverty scores above the cut-off of 16.19 also report themselves insured, mostly in Gilgit Town. This is in line with administrative data: Reportedly, the identification of eligible households was a challenge during program roll-out particularly in urban areas.[13] In consequence, the program selected households with poverty scores up to 19 as replacements. Figure 2.3 illustrates the enrollment rate by poverty score in rural and urban treatment areas jointly. In the control district Ghizer, only two households (1.54%) also reported themselves insured, even though the program had not been rolled out there. This speaks in favor of a small magnitude of measurement error regarding insurance status. For the RDD approach, I therefore use a fuzzy design in my main model specification.

Figure 2.3: Share of households insured (self-reported) by poverty score



► *Note:* The figure shows the average insurance rate, conditional on the poverty score, in treatment areas. The solid lines and shaded areas are the predicted values and associated 95%-confidence intervals, respectively, based on local mean smoothing. Red vertical line indicates the cut-off score of 16.19.
► Sample: Household-level panel in treatment areas ($N = 401$)
► Source: Endline survey (2018)

Despite the awareness campaign foreseen in the program design, only 61% of eligible households in my sample report having heard of the program (23% among non-eligible households). Note

[12]This rate does not necessarily reflect the program's overall outreach due to non-random sampling.
[13]I also find the highest attrition rate in Gilgit Town, also suggesting high mobility of households in urban areas.

that this is only 12 percentage points above the insurance rate - given that the program had been running for twenty months at the time of the endline survey, word-of-mouth spread of information seems to be rather limited. Conditional on being both, aware of and eligible for the program, 78% received an insurance card. That the program did not achieve full enrollment among targeted households might hence be due to limited information rather than refused uptake of the product, at least in my sample. Accordingly, the question arises whether enrolled households were informed about the full set of benefits. Indeed, only 35% of insured eligible households knew that only empaneled hospitals offer services, whereas 39% were not aware that some private hospitals can also be visited. At the same time, 61% of insured eligible households knew that only inpatient treatment was covered by the insurance, while 22% could not answer the question at all.

The consultancy OPM, which advised the consortium on behalf of KfW Development Bank, reported that by July 2017, the two private hospitals together had served only four insured persons, and the NGO-run hospital only 29. In line with these low numbers in private hospitals, the annual report 2017 published on the program's website[14] states that 82% of claims were made by public hospitals.[15] However, the claim data also reveals that 46% of claims in public hospitals were filed for appendicitis, whereas only 7% of claims in private hospitals stated that cause of admission, see Figure B.2 in the appendix. For comparison, in my survey data, 9% of hospitalization cases stated abdominal pain as reason, and 2% appendicitis. Two explanations are possible, namely actual malpractice, in that doctors conducted medically unnecessary appendectomies, and fraudulent claim filing, in that hospitals filed claims for non-existing treatments. In both cases, there seems to be a lack of control mechanisms within hospitals and among program stakeholders, suggesting weak governance in government hospitals and a lack of commitment to the program.

Consequent to an only marginal implementation in private and fraudulent implementation in public hospitals, I conclude that the insurance program did not work as intended. This is echoed in my survey, where almost half of the insured households report having used the card at a hospital at least once, and 21% even more than once. However, half of these respondents also reported problems in using their card, most often their card was not accepted by the hospital.

3-2 Differential program implementation by location

Of the five hospitals empaneled by the program, three are in Gilgit Town, one (private) just East of the city, and one (NGO) just North. All five hospitals are however within reach of each other in the same valley, suggesting possible inequity in access between urban and rural households, see Figure B.3 in the appendix.

In addition, my data contains evidence for geographic differences in enrollment. This does not

[14] http://sehathifazat.gog.pk/index.php/downloads/, retrieved on May 21, 2020.
[15] Unfortunately, claim data spanning the complete time period covered in this study is not available to me.

only refer to the urban/rural divide, but also extends to geographic differences within Gilgit Town. To show this, I code the settlements in Gilgit Town according to predominant sect affiliation.[16] My data thus contains households from three predominantly Ismaili settlements, four predominantly Shia settlements, and seven predominantly Sunni settlements within Gilgit Town.

Table 2.3 contains statistics on program implementation along geographic dimensions. To allow better comparison despite non-random sampling, I restrict the sample to eligible households. The first three columns show statistics for the urban and rural subsample separately, and the p-value of testing the null-hypothesis of equality of means. The last three columns show statistics for Ismaili and non-Ismaili (i.e., Shia and Sunni) areas separately, and the p-value of testing equality of means. Although the main divide in the city runs between Shia and Sunni sects, I pool households from predominantly Shia and predominantly Sunni areas because the consortium AKDN, which implements the program, is an Ismaili NGO. Whereas AKDN subscribes to a secular culture, it has historically been active predominantly in Ismaili regions (Varley 2015, Miller 2015).[17] As Hunzai (2013) reports, the contribution of the generally pro-development Ismaili community is recognized and appreciated by most people. Yet, the vertical organization of their community in general and the AKDN in particular has contributed to them being accused of promotion of own interests and opportunism.

Table 2.3: HETEROGENEOUS ENROLLMENT

Variable	Urban		Rural		p-value	Ismaili		Non-Ismaili		p-value
	μ_u	Std.Dev	μ_r	Std. Dev.	H_0: $\mu_u = \mu_r$	μ_I	Std.Dev	μ_{NI}	Std. Dev.	H_0: $\mu_I = \mu_{NI}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Having heard of the program	0.66	0.48	0.52	0.50	0.07	0.83	0.38	0.60	0.49	0.03
Insured by program	0.51	0.50	0.46	0.50	0.50	0.66	0.48	0.46	0.50	0.08
N	*107*		*63*			*29*		*78*		
Knew that insurance accepted only at empaneled hos.	0.29	0.46	0.45	0.51	0.15	0.21	0.42	0.33	0.48	0.35
Knew that only inpatient services covered	0.58	0.50	0.66	0.48	0.52	0.58	0.51	0.58	0.50	0.98
Used card at hospital	0.51	0.50	0.41	0.50	0.41	0.53	0.51	0.50	0.51	0.86
N	*55*		*29*			*19*		*36*		

- *Note*: This table contains indicators of program implementation for different subsamples. Variables on the left, statistics on top.
- Sample: Different subsets of eligible households (panel, varying N)
- Source: Endline survey (2018)
- Columns (1), (3), (6) and (8) display the shares and Columns (2), (4), (7) and (9) the standard deviation among eligible households (i.e., with poverty scores below 16.19) for the subsamples of urban, rural, Ismaili and non-Ismaili areas. Column (5) and (10) contain the p-value of testing equality of means in rural and urban samples, respectively in Ismaili and non-Ismaili areas.

The first two lines of Table 2.3 illustrate that program enrollment indeed varies between different locations. The program achieved larger outreach in urban than in rural areas. While the difference in general program awareness is weakly significant, the difference in enrollment rates is not. This latter difference is contrary to program reports which claim that identification of urban households for enrollment was more difficult than in rural areas.

Even larger and significant differences exist in program awareness and enrollment rates across the different areas of Gilgit Town. Households living in predominantly Ismaili settlements are

[16] For this I rely on information provided by Emma Varley, Associate Professor at the Department of Anthropology, Brandon University, who conducted five years of ethnographic fieldwork in Gilgit-Baltistan.

[17] The founder, his Highness the Aga Khan, is the 49th hereditary Imam of the Ismaili Muslims and their spiritual leader. He is based in France

much more likely to have heard of and to be enrolled in the program than households living in the predominantly Shia and Sunni parts of town. This finding is in line with AKDN's affiliation with the Ismaili sect, and suggests a sectarian bias in program implementation.[18]

Regarding knowledge of insurance coverage and card usage among insured households, no clear picture emerges. Rural households as well as households in non-Ismaili settlements seem to have a higher knowledge, but use the card less often. None of these differences are significant though and they might simply be due to chance.

4 Regression discontinuity design (RDD)

Given that eligibility for the program is defined through an exogenous and pre-existing national poverty score Z_i, I use a regression discontinuity design for estimating the effect of the program on a number of outcomes Y_i. For this, I calculate local linear regression models to the left and right of the cut-off score with an epanechnikov kernel, using the bias-corrected bandwidth estimator suggested in Calonico et al. (2017) and implemented in the Stata command rdrobust. As before mentioned, incomplete enrollment below and additional enrollment above the cut-off renders the fuzzy design the appropriate choice. Therefore, I use the poverty score Z_i as instrument for the actual insurance status D_i. The standard errors in the sharp design are naturally smaller than in the fuzzy design as the fuzzy design introduces additional uncertainty in the first stage estimation. Since my result is that the program had no impact, one can argue that the smaller standard errors are the more conservative choice here. Accordingly, I also show results for the sharp design.

Depending on the outcome Y_i considered, I use three different samples: The household sample I use for estimating effects on outpatient care, neglected health care, and subjective variables, such as financial statisfaction and citing health as main financial risk. Since questions pertaining to inpatient care were asked relating to each household member separately, I use the member-level sample to measure an effect on the extensive margin of inpatient care. The fact that one respondent answered questions for each household member might lead to reporting bias, for which I account by clustering standard errors on the household level. Finally, I use what I call the conditional sample, consisting of those household members who experienced a case of inpatient care, to analyze effects on care consumption patterns conditional on being admitted to hospital, specifically, hospitalization costs, associated worries, use of private hospitals, and the intensive margin of inpatient care. Here, too, I cluster standard errors on household level. In Table B.5 in the appendix, I show results without clustering as a robustness check.

The poverty score Z_i was assigned to each household in all of Pakistan in 2010, initially to determine eligibility of households for a national social welfare program, called the Benazir

[18] Note that the same NGO already offered a health insurance product prior to my study. Nevertheless, at baseline only 1.9% of households in Gilgit Town reported being insured. Here, too, the share is highest within Ismaili-dominant settlements (4.4%).

Income Support Program (BISP). Accordingly, the poverty score is referred to as BISP score. The BISP score Z_i is based on a proxy means test using the following variables X_{BISP}: Number of household members, number of household members under age of 18 or above age of 65, highest education level of household head, number of children attending school, number of rooms in house, type of toilet, as well as ownership of ten specific household assets, vehicles, agricultural land or livestock. The final poverty score is a function f of these variables, $Z_i = f(X_{BISP})$, where the functional form of f is unknown to me. Households with a poverty score below 16.17 are eligible for benefits under the BISP, the largest single social program in Pakistan. This program includes various features, the most prominent being an unconditional cash transfer of 1,000 PKR per month, which aimed at cushioning the adverse effects of the 2008 financial crisis.[19]

The design of the BISP score and the fact that eligibility for BISP is based on a cut-off score very close to the one that determines eligibility for the insurance (16.17 for BISP, 16.19 for Sehat Hifazat) has three consequences for this study, which I discuss in the following subsections.

4-1 Self-selection and manipulation of BISP score

In theory, it is possible that households manipulated the poverty score in an attempt to self-select into the social program. Therefore, I run the McCrary test on the original sampling frame of BISP data, which gives an insignificant discontinuity estimate. Note that I use the complete BISP dataset for the union councils considered in this study instead of my sample due to non-random and non-symmetric sampling around the cut-off.[20] Figure 2.4 visually confirms that there is no discontinuity in the BISP score density around the cut-off. In practice, this result seems reasonable as the exact function f which assigns PMT answers to poverty scores is unknown and it is hence unlikely that households could temper with its calculation with any degree of precision.

4-2 Continuity of covariates

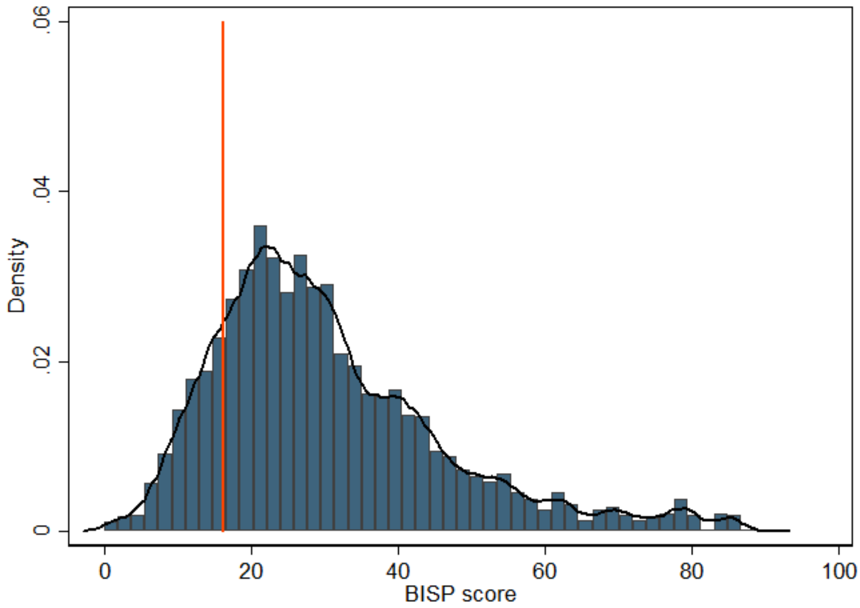
The score is not random and the set of score-determining covariates X_{BISP} might well also influence the outcomes considered in this study. If the design of the poverty score led to discontinuities in these covariates at the cut-off, the RDD estimates would be biased if covariates are not properly accounted for (Frölich and Huber 2019).

To test for discontinuities in covariates, I run RDD estimations with the covariates as outcome variables on the three samples considered here. I use the same specification as in my main

[19]See https://bisp.gov.pk/ Officially, the program also includes a small loan program for women, a vocational training program, and a health insurance component. The breadth and depth of outreach of these additional features is however smaller than the cash transfer component. In particular, no household in my sample reported to be insured under BISP. The monthly cash transfer was raised to 1,500 PKR per month after the survey period.

[20]I sampled more households to the right of the cut-off than to the left, because it was ex-ante not clear which precise cut-off the Government would chose.

Figure 2.4: Density of BISP score in survey areas



► *Note:* The figure shows the distribution of poverty scores in the survey areas. The black lines illustrates a kernel density estimates, the red line indicates the Sehat Hifazat cut-of score of 16.19.
► Sample: Households in survey areas within treatment district (frame population, $N = 91,825$ in three union councils in district Gilgit)
► Source: BISP data (2010)

estimation, including a fuzzy design and clustering of standard errors on household level in the member and conditional sample. Table 2.4 contains the results. These are robust towards specifying a triangular kernel and clustering standard errors for urban and rural areas respectively. Table B.1 in the appendix contains the results using the sharp design.

Table 2.4: Continuity of covariates

Variable	Household sample			Member sample			Conditional sample		
	β^H_{RDD} (1)	S.E. (2)	p-val. (3)	β^M_{RDD} (4)	S.E. (5)	p-val. (6)	β^C_{RDD} (7)	S.E. (8)	p-val. (9)
Urban	-0.037	0.288	0.894	-0.123	0.339	0.716	-0.773	1.623	0.635
Household size	-3.663	2.216	0.128	-6.415	3.621	0.101	-22.283	34.317	0.516
Avg. Monthly HH income (PKR, win99)	-13,116	9,332	0.160	-16,843	12,327	0.172	-72,587	100,000	0.487
Wealth index	-2.689	2.042	0.172	-3.101	2.140	0.147	-14,075	18,533	0.448
Reported distance to next hosp. (minutes, win99)	238.10	374.63	0.333	196.31	352.68	0.578	947.17	2,260.50	0.675
No HH member completed primary school	0.398	0.244	0.104	0.313	0.255	0.219	1.726	2.843	0.544
Some HH member with higher educ.	-0.426	0.339	0.208	-0.383	0.373	0.305	-2.820	4.069	0.488
Age				-0.896	3.406	0.793	-45.845	72.051	0.525
Female				-0.014	0.102	0.893	-0.904	1.545	0.559
N (left/right of cut-off	*170/231*			*1,408/1,822*			*182/326*		

► *Note*: This table contains rdd estimation results using covariates as outcomes. Variables on the left, statistics on top.
► Sample: Columns (1) to (3): Household sample (panel, N=401); Columns (4) to (6): Member sample (panel, N=3,230); Columns (7) to (9): Conditional sample (panel, N=508)
► Source: Baseline survey (2016)
► Columns (1), (4), and (7) display the coefficient estimates, Columns (2), (5), and (8) the associated standard errors, and Columns (3), (6), and (9) the p-value of the null hypothesis of zero effect size, for the three samples respectively. All estimates are obtained by running local linear regression models to the left and right of the cut-off, using the bias-corrected bandwidth that minimizes the mean squared error, as implemented in the Stata command rdrobust by Calonico et al. (2017), and specifying a fuzzy design, where the poverty score serves as instrument for the self-reported insurance status. For the member and conditional sample, standard errors are clustered on household level.
► Table B.1 in the appendix contains corresponding results using the sharp design.

There are no significant jumps in the conditional expectation of covariates in the fuzzy design, hence I do not account for covariates in my main model specification. As a robustness check, I include the covariates wealth index, household size, age and gender, since this might improve precision. However, when using the sharp design there are discontinuities in household size in all subsamples, see Table B.1 in the appendix. In the conditional sample, there are significant discontinuities in four additional variables when using the sharp design. Whereas Frölich and Huber (2019) provide an estimator that eliminates the potential bias caused by these discontinuities, the bandwidth used in their approach is too large for my setup, especially in the conditional sample, potentially creating additional bias. Therefore, I do not apply their estimator, but note that the estimates obtained in the sharp design for the conditional sample might be biased and should be interpreted with care.

4-3 Confounding program

The poverty score was assigned in 2010, shortly after which identified beneficiaries received the benefits of the BISP program. By the time of my endline survey, households with scores below 16.17 had hence received benefits of the social program for up to eight years. If the BISP had a significant impact on health care usage and financing, the RDD approach in this study would pick up the net effect of the BISP and the Sehat Hifazat treatment. To test for a potential BISP impact, I therefore run the estimation on baseline variables and show results in Table 2.5.

There is a significant discontinuity in reporting neglected health care, with households just below the cut-off being more likely to report that they chose not to visit the hospital although it was recommended. There is no apparent reason why the existing BISP program should have such a positive effect. If anything, the effect should have been negative (i.e. reducing neglected health care), since the reason for neglected health care almost always was lack of financing and the BISP eased financial constraints below the cut-off. An alternative explanation is an anticipation effect, in that insured households neglected their health care in the months preceding the baseline in anticipation of free health care after the program's start.[21]

4-4 Specification checks

I check the plausibility of my model by calculating three types of pseudo results: First, I estimate effects on the control sample in the district Ghizer, where the program was not implemented. Here, I do not consider variables conditional on a case of inpatient care due to a small sample size around the cut-off.[22] Second, I estimate effects on plausibly invariant endline variables, namely age and whether a child was born in the past year. Third, I estimate effects at other poverty scores, namely at 17.19 and 19.19 (sharp design). Tables B.2, B.3, and B.4 in the appendix contain the respective results. Of the thirty estimates I thus calculate, only two are

[21] An anticipation effect would also imply a reduced likelihood of inpatient care, which I do not observe in my data. However, the survey asked about inpatient care in the past twelve months, whereas neglected health care referred to the past six months. As such, the two measures might not be directly comparable.

[22] The effective sample size here would be 14 observations below and 12 observations above the cut-off.

Table 2.5: Pseudo effects on baseline outcomes

Variable	Fuzzy design β_{LATE} (1)	S.E. (2)	p-val. (3)	Sharp design β^M_{ITT} (4)	S.E. (5)	p-val. (6)
I. Health care financing						
Fin. Satisfaction $\leq$ 5(scale 1 to 10)	-0.066	0.249	0.789	-0.025	0.093	0.787
Citing health shock as main risk	0.080	0.288	0.781	0.037	0.114	0.743
Dif.ty finding money for health care $\leq$ 8 (scale 1 to 10)	0.177	0.323	0.583	0.067	0.122	0.582
N (left/right of cut-off)			*170/231*			
Total cost of last treatment (PKR, win99)	-25,434	47,946	0.596	-6,376	12,050	0.597
Very worried re. cost of inp. Care	0.222	0.587	0.705	0.055	0.162	0.732
N (left/right of cut-off)			*181/256*			
II. Health care usage						
Case of inpatient care	0.107	0.087	0.222	0.043	0.028	0.122
N (left/right of cut-off)			*1,408/1,822*			
Neglected health care	0.834	0.365	0.022	0.303	0.097	0.002
Case of outpatient care	0.026	0.300	0.929	0.016	0.129	0.901
N (left/right of cut-off)			*170/231*			
Use of private hospital	-0.246	0.385	0.523	-0.046	0.103	0.653
More than one admittance	0.492	0.472	0.297	0.124	0.105	0.236
N (left/right of cut-off)			*181/256*			

- *Note*: This table contains rdd estimation results on baseline values of outcome variables. Variables on the left, statistics on top.
- Sample: Household and member sample (panel, varying N)
- Source: Baseline survey (2016)
- Columns (1) and (4) display the coefficient estimates, Columns (2) and (5) the associated standard errors, and Columns (3) and (6) the p-value of the null hypothesis of zero effect size, for the fuzzy and sharp design respectively. All estimates are obtained by running local linear regression models to the left and right of the cut-off, using a bandwidth that minimizes the mean squared error, as implemented in the Stata command rdrobust by Calonico et al. (2017). In the fuzzy design specification, the poverty score serves as instrument for the self-reported insurance status. For variables measured in the member and conditional sample, standard errors are clustered on household level.

weakly significant . Given the large number of hypotheses tested and the size of the p-values (0.060 and 0.065 respectively), the effects might well be due to multiple hypotheses testing. Hence, I conclude that there is no indication of model misspecification.

5 Local average treatment effects

Table 2.6 contains the results from my estimations for all outcome variables. As before, I distinguish between effects regarding health care financing (Panel I) and health care usage (Part II). I show effect estimates, standard errors, and p-values testing the hypothesis of no effect for the fuzzy design (Columns (1) to (3)) as well as for the sharp design (Columns (4) to (6)). The results suggest that the program did not provide the intended financial protection and did not lead to an increased quantity of health care consumed. These results are robust towards different model specifications, as demonstrated in Table B.5 in the appendix.

Table 2.6: ESTIMATION RESULTS

Variable	Fuzzy design			Sharp design		
	β_{LATE} (1)	S.E. (2)	p-val. (3)	β^{M}_{ITT} (4)	S.E. (5)	p-val. (6)
I. Health care financing						
Fin. Satisfaction $\leq$ 5(scale 1 to 10)	0.320	0.344	0.353	0.133	0.124	0.284
Citing health shock as main risk	0.074	0.298	0.802	-0.005	0.115	0.961
Dif.ty finding money for health care $\leq$ 8 (scale 1 to 10)	-0.237	0.282	0.402	-0.094	0.112	0.400
N (left/right of cut-off)			*170/231*			
Total cost of last treatment (PKR, win99)	32,775	110,000	0.767	-289	12,407	0.981
Very worried re. cost of inp. Care	-0.150	1.287	0.907	-0.033	0.152	0.828
N (left/right of cut-off)			*182/326*			
II. Health care usage						
Case of inpatient care	-0.110	0.106	0.300	-0.052	0.038	0.165
N (left/right of cut-off)			*1,408/1,822*			
Neglected health care	-0.092	0.208	0.655	-0.044	0.079	0.576
Case of outpatient care	0.249	0.288	0.388	0.069	0.123	0.573
N (left/right of cut-off)			*170/231*			
Use of private hospital	0.856	1.453	0.556	0.097	0.096	0.312
More than one admittance	1.123	1.673	0.502	0.186	0.134	0.165
N (left/right of cut-off)			*182/326*			

- *Note*: This table contains RDD estimation results on main outcome variables. Variables on the left, statistics on top.
- Sample: Household and member sample (panel, varying N)
- Source: Endline survey (2018)
- Columns (1) to (3) contain results applying a fuzzy design, where insurance status is instrumented via the poverty score. Columns (4) to (6) contain results applying a sharp design, where treatment is strictly determined by poverty score. Columns (1) and (4) display the coefficient estimates, Columns (2) and (5) associated standard errors, and Columns (3) and (6) the p-value of testing the null hypothesis of zero effect size. All estimates are obtained by running local linear regression models to the left and right of the cut-off, using the bias-corrected bandwidth that minimizes the mean squared error, as implemented in the Stata command rdrobust by Calonico et al. (2017). For the member and conditional sample, standard errors are clustered on household level.
- Table B.5 in the appendix contains robustness checks.

5-1 Effect on health care financing

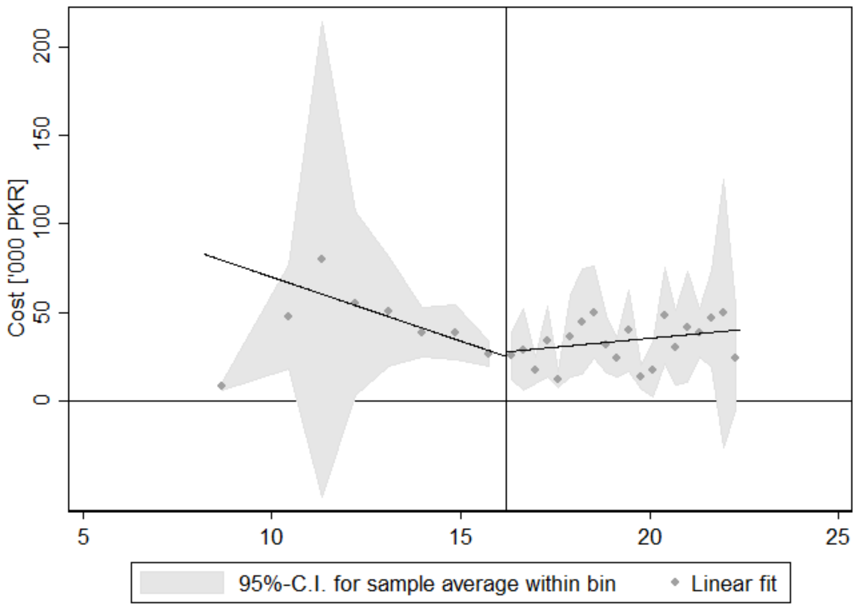
The Sehat Hifazat program is based on the postulated impact chain that the insurance reduces expected out-of-pocket payments for inpatient care, and that the reduced payments in turn lead to an increased usage of hospitalization services. As I look at rather short-term effects, the latter impact would also unfold if the program only reduced the expected but not the actual cost of inpatient care. I therefore aimed at measuring this ex-ante feeling of protection by asking questions on the financial satisfaction (rating scale 1 to 10), on the main financial risks faced, and on their agreement with the statement that finding the money to pay for health care of household members has been difficult during the past twelve months (rating scale 1 to 10). The first three lines of Table 2.6 show that the program did not have a significant impact on these subjective indicators of financial well-being.[23]

In addition, for each household member who was admitted to hospital (conditional sample), I asked about the total costs incurred in the most recent admission. I estimate an effect

[23] Note that I transformed the ordinal measurements on rating scales into dummy variables indicating an answer above or below the respective median values.

of 32,775 in the fuzzy design, fourth line in Table 2.6. The effect is insignificant and in an unanticipated direction. Conversely, the estimate in the sharp design is much smaller but negative. Not unexpectedly, the reported costs vary greatly. This variance together with the relatively smaller sample size in the conditional sample leads to a large standard error, as illustrated also in Figure 2.5 for the sharp design. Therefore, I also asked respondents for their agreement with the statement that he/she experienced sleepless nights when the respective household member was admitted to the hospital out of fear that they could not afford the treatment. I obtain an effect size of -0.150 in the fuzzy design, see fifth line in Table 2.6. Whereas the direction of the estimate points to a decrease of the actual financial burden, the effect is again statistically insignificant.

Figure 2.5: Total cost of inpatient treatment by poverty score



► *Note:* The figure shows the total cost of treatment by poverty score, as 95%-confidence intervals around sample averages in bins (gray areas) and as linear regression function fit (black line).

As with any insurance, moral hazard and/or fraud by hospitals could be an issue: Hospitals might charge higher prices for insured individuals on items not covered by the insurance, or elicit unofficial side payments. In fact, the Annual Report 2017 by AKDN suggests that at one of the empaneled public hospitals, insured households could not obtain medicines but had to visit private dispensaries to redeem their prescriptions, which can be expensive.[24] To obtain a precise picture of the OOP expenditures, I therefore collected data on costs incurred for the following separate items: transportation, admission, diagnosis and treatment, medicines, further prescriptions and documents, or meals and accommodation (incl. for accompanying relatives).[25] Table 2.7 contains the results of the effect estimation on these outcomes. The estimates vary greatly both in magnitude and direction, and none of the effects are significant.

[24] http://sehathifazat.gog.pk/index.php/downloads/, retrieved on May 21, 2020.

[25] These variables were only measured at endline, hence I cannot check for pre-existing discontinuities at baseline.

Table 2.7: Estimation results on individual OOP cost positions

Variable	Fuzzy design			Sharp design		
	β_{LATE} (1)	S.E. (2)	p-val. (3)	β^{M}_{ITT} (4)	S.E. (5)	p-val. (6)
Transportation	908	17,186	0.958	-68	1,734	0.968
Admission (e.g. forms)	251	3,639	0.945	-74	495	0.881
Diagnosis and treatment	12,701	23,000	0.581	1,279	3,975	0.748
Medicines	20,359	60,499	0.736	2,320	5,717	0.685
Prescriptions and documents	-11,827	21,736	0.586	-1,544	1,751	0.378
Meals and accommodation	11,071	18,977	0.560	1,332	1,280	0.298
N (left/right of cut-off)	*182/326*					

- *Note*: This table contains RDD estimation results on the individual OOP cost positions. Variables on the left, statistics on top.
- Sample: Conditional sample (panel, N=408)
- Source: Endline survey (2018)
- Columns (1) to (3) contain results applying a fuzzy design, where insurance status is instrumented via the poverty score. Columns (4) to (6) contain results applying a sharp design, where treatment is strictly determined by poverty score. Columns (1) and (4) display the coefficient estimates, Columns (2) and (5) associated standard errors, and Columns (3) and (6) the p-value of testing the null hypothesis of zero effect size. All estimates are obtained by running local linear regression models to the left and right of the cut-off, using the bias-corrected bandwidth that minimizes the mean squared error, as implemented in the Stata command rdrobust by Calonico et al. (2017). For the member and conditional sample, standard errors are clustered on household level.

I conclude that the program did not significantly contribute towards financial protection against health risks, neither regarding a subjective ex-ante feeling nor on actual costs for hospitalization, be it total costs or individual cost positions.

5-2 Effects on health care usage

Given that the program did not provide real or expected financial protection against health care cost for households in my sample, it is unsurprising that health care consumption did not increase. My main outcome variable is the use of inpatient care, as this is what the insurance covers primarily. Table 2.6, Panel II, shows that I estimate an insignificant effect of -0.110 in the fuzzy design. Note also that at baseline, the pseudo effect was positive (0.107), suggesting a net negative effect of the program. However, the difference between baseline and endline effect is not significant. When including covariates in the sharp design, the coefficient becomes significant at the 10% level but remains negative. I hence conclude that the program did not increase the consumption of hospital care along the extensive margin as intended.

Similarly, the program did not affect the likelihood of reporting a case of neglected health care in the household (insignificant estimate of -0.092 in the fuzzy design). In this variable I had observed a significant effect of 0.834 at baseline (Table 2.5). To obtain the net effect of the Sehat Hifazat program on neglected health care, I could in principle run a difference-in-discontinuity analysis using baseline and endline data. However, as explained above, it is reasonable to assume that the baseline data captures anticipation effects, and hence a difference in discontinuity analysis would overestimate the program's impact. Therefore, the null effect estimated at endline seems the more conservative and credible estimate.

Whereas the program mainly covers inpatient care, an exception is maternity care, where also outpatient services are covered. Therefore, I also estimate whether the program had an impact on the usage of outpatient care and obtain a positive but insignificant estimate of 0.249 in the fuzzy design. Most reported cases of outpatient care are however related to either stomach problems or fever, and only seven households reported the use of antenatal care or childbirth. In fact, I also asked households with childbirth whether they used professional assistance and what type of assistance. In both, baseline and endline survey, over 80% of babies were delivered with the help of professional care, the large majority using inpatient services. Given the high level of usage already at baseline, it is little surprising that I find no effects of the program on these variables using the regression discontinuity specification.[26]

Finally, I analyze the patterns of seeking hospital care in more detail and estimate effects on the usage of private versus public hospitals and whether an individual visited a hospital more than once in the conditional sample. For both variables, I obtain positive and sizeable but insignificant effects. As laid out in Section 2, program reports also suggest that the program did not unfold its full potential in private hospitals, which is in line with the insignificant effect I estimate on hospital choice.

Overall, the results suggest that the proposed impact chain did not unfold as expected in the program within the time period considered for households in my sample.

6 Heterogeneous effects

A key obstacle to health care in GB is accessibility of hospitals. There are two ways in which accessibility is restricted.

First, as depicted in Figure B.3 in the appendix, the empaneled hospitals are all located within or just outside of Gilgit Town. This implies that rural households have long distances to travel to reach a hospital: Whereas households in Gilgit Town report a median distance of 30 minutes travel to the next hospital, the median travel distance increases to 120 minutes for rural households in the district of Gilgit. The program design foresees a co-payment to transportation costs, however, I find no reports on these actually being paid out nor on foreseen payment mechanisms. Even if the payments indeed reached the households, it is unclear to what extend they covered actual transportation costs. It is therefore reasonable to assume that access to hospitals remained particularly challenging for rural households, and hence I look at heterogeneous effects along the rural/urban divide.

Second, even for residents within Gilgit Town, de facto accessibility of hospitals is restricted due to a sectarian divide along Shia and Sunni lines. As mentioned above, sectarian tensions in 2005 centered around the DHQ hospital, which is situated within a majority Shia neighborhood, see Figure B.1 in the appendix. In response, a Sunni hospital was set up in a Sunni neighborhood. Both hospitals are empaneled in the program, but the Sunni City Hospital reportedly offers

[26] This does not rule out that the program increased quantity of care along the intensive margin, i.e., days spent at hospital after delivery, as I do not have data on this variable.

lower quality services than the Shia DHQ (Varley 2010). Given the results of anthropological studies such as Varley (2015), I assume that households living closer to the City Hospital are more likely to visit this hospital, not so much due to geographic proximity but sect affiliation, whereas households living closer to the DHQ Hospital are more likely to visit the higher quality DHQ. Therefore, I segregate households in my urban sample by the geographic linear distance to the (Sunni) City or the (Shia) DHQ hospitals as measured by GPS data, and analyze heterogeneous effects for the two subgroups.

I calculate heterogeneous effects by restricting my sample to the respective subgroup. As the sample size for each subgroup is then rather small, I do not look at outcomes which are only defined for the conditional sample. In addition, I restrict attention to the sharp design to keep standard errors small. I present results for the fuzzy design in Table B.6 in the appendix.

Table 2.8: HETEROGENEOUS EFFECTS - SHARP DESIGN

Variable	Urban		Rural		Near DHQ		Near City Hosp.	
	β_u	S.E.	β_r^M	S.E.	β_{DHQ}	S.E.	β_{CH}	S.E.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
I. Health care financing								
Fin. Satisfaction ≤ 5 (scale 1 to 10)	0.168	0.266	0.057	0.180	0.256	0.538	0.527	0.372
Citing health shock as main risk	-0.141	0.198	0.132	0.184	-0.057	0.301	0.101	0.391
Dif.ty finding money for health care ≤ 8 (scale 1 to 10)	0.193	0.163	-0.231	0.158	0.312	0.328	0.131	0.187
N (left/right of cut-off)	*107/171*		*63/60*		*61/83*		*46/88*	
II. Health care usage								
Case of inpatient care	-0.067	0.062	-0.046	0.050	-0.038	0.081	-0.074	0.095
N (left/right of cut-off)	*875/1,347*		*534/475*		*520/651*		*355/696*	
Neglected health care	0.379	0.230	-0.042	0.108	0.015	0.087	0.362	0.338
Case of outpatient care	0.388*	0.232	-0.218	0.194	0.442	0.473	0.962***	0.337
N (left/right of cut-off)	*107/171*		*63/60*		*61/83*		*46/88*	

- *Note*: This table contains heterogeneous effect estimation on main outcome variables. Variables on the left, statistics on top.
- Sample: Subsamples of household and member sample (panel, varying N)
- Source: Endline survey (2018)
- Columns contain the effect estimate and standard errors for the following subsamples:
  - Column (1)-(2): Households/members living in Gilgit Town
  - Column (3)-(4): Households/members living outside of Gilgit Town
  - Column (5)-(6): Households/members in Gilgit Town living closer to the DHQ than to the City Hospital, implying a higher propensity to belong to the Shia sect
  - Column (7)-(8): Households/members in Gilgit Town living closer to the City Hospital than to the DHQ, implying having a higher propensity to belong to the Sunni sect

  All estimates are obtained by running local linear regression models to the left and right of the cut-off, using the bias-corrected bandwidth that minimizes the mean squared error, as implemented in the Stata command rdrobust by Calonico et al. (2017), in the sharp design. For the variable *Case of inpatient care*, standard errors are clustered on household level.
- Statistical significance is given as follows: *** indicates $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
- Table B.6 in the appendix contains results of the fuzzy design, where insurance status is instrumented by poverty score.

Table 2.8 contains the results. There are no significant effects on any of the subsamples on variables concerning health care financing, inpatient care or neglected health care. There is a significant effect on the usage of outpatient care in the urban sample, which is driven by households living closer to the City than the DHQ Hospital. For this group, the effect is highly significant, positive and sizeable. Note that 93% of my sample reported receiving outpatient care at a hospital (75% at a public hospital, 18% at a private hospital, 7% at a doctor's practice, pharmacy or traditional healer). One explanation for the increase in outpatient care could therefore be an overly strong gatekeeping mechanism, especially at City Hospital: Household might try to use the card at a hospital but receive outpatient instead of inpatient care. The effect is even larger in magnitude when using the fuzzy approach but due to the larger standard errors it is then insignificant, see Table B.6 in the appendix. Further research would be needed to decide whether the effect is due to strong gatekeeping or sampling error.

Overall, the results above suggest that even households with the least restrictions on accessibility, i.e., households in Gilgit Town living near the DHQ Hospital, did not benefit from the program as expected. However, these households live in predominantly Shia settlements, and, as shown in Section 3-2, are hence least likely to have received the insurance card in the first place. It is therefore worthwhile to analyze whether the program had a positive effect on Ismaili households, who are the most likely to have received the card (in line with the implementing consortium being an Ismaili NGO).

As before, I therefore proxy sect affiliation of households in my urban sample by the reported settlements they live in. This leaves a small number of households, particularly around the cut-off, rendering the RDD approach unfeasible.[27] Instead, I look at changes in my main variable between baseline and endline for Ismaili and non-Ismaili subgroups separately. I look at the usage of inpatient care only because this is the only variable with a sufficiently large sample size. As before, I cluster standard errors on household level. Table 2.9 contains the results.

Table 2.9: Use of inpatient care at baseline and endline in non-Ismaili and Ismaili settlements

	Baseline				Endline			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Non-Ismaili		Ismaili		Non-Ismaili		Ismaili	
	μ_{N-I}	S.E.	μ_I	S.E.	μ_{N-I}	S.E.	μ_I	S.E.
I Case of inpatient care	0.141	0.010	0.145	0.016	0.152	0.012	0.223	0.028
N	1,607		615		1,607		615	
	$\mu_I - \mu_{N-I}$		S.E.		$\mu_I - \mu_{N-I}$		S.E.	
II Difference	0.004		0.019		0.071**		0.030	
N	2,222				2,222			

- *Note*: This table contains statistics on the use of inpatient care.
- Sample: Subsamples of household and member sample (panel, varying N)
- Source: Endline survey (2018)
- Columns (1) to (4) contain baseline statistics, Column (5) to (8) endline statistics
  - Line I: Shares of individuals with a case of inpatient care with standard errors, for predominantly non-Ismaili and majority-Ismaili settlements respectively.
  - Line II: Difference in the shares of individuals with a case of inpatient care between predominantly non-Ismaili and majority-Ismaili settlements, with standard errors.

  All standard errors are clustered on household level and homoscedasticity robust.
- In Line II, statistical significance is given as follows: *** indicates $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

At baseline, the share of individuals admitted to hospital within the previous twelve months was very similar for households living in predominantly non-Ismaili and majority-Ismaili settlements (0.141 and 0.145 respectively, Line I), with the difference being a precisely estimated zero (Line II). At endline, the share increased slightly for individuals living in non-Ismaili settlements but much more for individuals living in Ismaili settlements (0.152 and 0.223 respectively). The endline difference between the two groups is significant with a p-value of 0.018. This finding is in principle in line with the Sehat Hifazat program improving access to hospitals particularly for Ismaili households.

Whereas this analysis captures the idea of the difference-in-difference estimation of average

[27] The RDD approach works for Shia and Sunni settlements respectively and estimates null effects for both subsamples. It is unfeasible however for the Ismaili subgroup, which is of greatest interest here, because there are no Ismaili households directly to the right of the cut-off.

treatment effects, it is here meant to provide descriptive evidence only. A causal interpretation would require making the common trend assumption. This might be an overly strong assumption here, given that Ismaili and non-Ismaili groups might have been affected differently either by sectarian conflicts and political decisions, which could have restricted access for non-Ismailis, or by health shocks and additional NGO support which could have increased hospital usage for Ismailis. Therefore, the findings above should be interpreted as suggestive evidence, which should serve to motivate future research rather than provide conclusive answers in themselves.

7 Conclusion

The Sehat Hifazat program in Gilgit Balistan aimed at decreasing the out-of-pocket payments at hospitals and hence increase the consumption of inpatient care. To do so, the program provided fully subsidized health insurance cards to households with a poverty score below a fixed cut-off. This creates the ideal set-up for a treatment effect estimation using a regression discontinuity design. Given that there is incomplete enrollment to the left and contamination to the right of the cut-off, my main specification uses a fuzzy approach, where the self-reported insurance status is instrumented by the poverty score. Using baseline values, data from a control district and a range of additional pseudo tests, I confirm that the key assumptions underlying the RDD approach are fulfilled in this setting.

I find no significant local treatment effect of the program on a variety of outcome variables. Reported costs of inpatient treatment remain unaffected, both in total and regarding individual cost positions. Whereas the variables suffer from a high variance, subjective indicators such as financial satisfaction and worrying about hospitalization costs also show no improvement due to the program. In line with this, the use of inpatient care did not increase and the incidents of reported neglected health care did not decrease. This suggests that the program did not meaningfully change the health care behavior of the target population. It should be noted however that the RDD approach is only locally valid and I hence cannot rule out that larger effects might have materialized for households with low poverty scores.

Key factors which might inhibit the program are the limited accessibility of hospitals for the rural population and the differential quality of the two public hospitals which provide services along sectarian lines. I therefore also look for heterogeneous effects, but find no benefits from the program for the population group with comparably direct access to the best available hospital. For households living close to the City Hospital, I find an increase in outpatient care, which might be due to strong gatekeeping.

However, I find some indication that the program was not implemented as designed. Administrative data reveals that the program was not fully operational at hospital level: Whereas private hospitals submitted only few claims to the insurer, public hospitals made an unusual high number of claims for treatment of appendicites, which suggests fraudulent behavior. In addition, at endline survey almost half of the households who tried using the insurance card complained that the card was not accepted. Furthermore, I find the highest enrollment rates in those settlements of Gilgit Town, where a predominantly Ismaili community lives. Descriptive evidence also suggests that this sub-group saw the largest increase in the usage of hospital services between baseline and endline. While I cannot claim causality, the finding is in line with the fact that the consortium selected to implement the program is an Ismaili NGO.

In conclusion, the program did not impact the health care consumption and financing of beneficiary households close to the cut-off. In my data, I find no indication that limited physical access to hospitals is the main constraint. In contrast, sectarian preference in the enrollment process, limited participation of private hospitals, and partially fraudulent behavior at public hospitals might all have combined to inhibit the program from developing its full potential. This highlights the various difficulties faced by governments and implementing agencies operating in a complex setting such as in Gilgit Baltistan. From a methodological point of view, my study underlines the importance of using administrative or monitoring data to complement survey-based evaluations, and advocates for considering cultural factors such as religious affiliation as potential program constraints.

Chapter 3

Assisting the Growth of Micro and Small Enterprises - Evidence from an RCT with 12 Financial Service Providers

With Alexandra Avdeenko and Markus Frölich

1 Introduction

Accounting for an overwhelming majority of businesses, Micro and Small Enterprises (MSEs) are the backbone of many developing and emerging economies (Ayyagari, Beck, and Demirguc-Kunt 2007; De Mel, McKenzie, and Woodruff 2008). Supporting MSE owners increase their profits is therefore a noble aim of the self-help advocating development cooperation. The struggle against micro-level inhibitors to MSE growth has long focused on financial constraints (e.g., Banerjee, Duflo, Glennerster, and Kinnan 2015), but the focus has recently shifted to addressing the lack of business and managerial skills. To this end, donors and practitioners pioneer a vast range of training courses, be it as stand-alone entrepreneurship program or as addition to financial services. One of the largest training courses world-wide is the Start and Improve Your Business (SIYB) of the International Labour Organization (ILO), having been used by over 65,000 trainers in around 100 countries.

In our study, we evaluate a training and counseling intervention in Indonesia based on the SIYB course, implemented in cooperation with twelve different Financial Service Providers (FSPs), including savings and credit cooperatives, rural banks, and a development bank. Here, the ILO trained more than 150 loan officers of these FSPs to become trainers and/or counselors, with a focus on financial management and marketing. The trained loan officers then provided two-day classroom training sessions and/or five one-on-one counseling sessions to their existing MSE clients.[1] Our research design is a randomized control trial (RCT) with a two-wave panel of 3,975 clients from twelve participating FSPs, who were randomly assigned to the control group or one of three treatment arms: (1) Only classroom training, (2) only individual counseling, or (3) classroom training with subsqeuent individual counseling. With this design we tackle the following research questions: First, calculating average treatment effects, we investigate whether Indonesian MSEs with pre-existing access to finance on average benefited from this intervention. Second, we investigate whether a subgroup benefited more than others, be it because of entrepreneur or firm characteristics, treatment intensity, or serving FSP. For this, we exploit our relatively large sample size, the fact that some FSPs offered three different treatment arms, and - as novel contribution - the fact that the program worked with twelve different FSPs. To comply with best research standards, we uploaded a Pre-Analysis Plan (PAP) on the AEA website (RCT ID: AEARCTR-0003625) before the endline data was collected.[2]

Our results suggest that the program did not change the profits, household spending or loan behavior of the MSEs. In principle, this might be the case because either the program did not focus on relevant skills, or was unsuccessful in teaching the skills it focused on. Our results cannot rule out the latter explanation, since we find no increase in knowledge and no changes in business practices among the treatment group, except for an increase in the share of clients whose business plan includes a cash flow analysis. This holds true on average as well as for subgroups identified by entrepreneur or enterprise characteristics, and does not depend on the treatment form (training and/or counseling).

Importantly, however, our results vary across the different FSPs. To our knowledge, this is the first study working simultaneously with more than one partner institution and the first to directly compare the impact achieved in different types of financial institutions. We find that rural banks achieve consistent, albeit small, improvements in knowledge and practice outcomes, whereas there is no consistent impact in credit cooperatives.[3] In the absence of heterogeneous effects along observable dimensions, this might imply that rural banks are more successful than cooperatives in identifying high-potential entrepreneurs along dimensions unobservable to us, or have more highly qualified loan officers who acted as trainers. Furthermore, one rural bank achieved particularly large impact on intermediate outcomes, and this bank also demonstrated the highest implementation fidelity in our monitoring data. Yet, even for this bank, final outcomes such as business profits or household spending remain unchanged.

With our study we contribute to a growing body of literature on MSE growth in developing countries. Along with access to finance, improved business practices and efficiency can potentially increase profitability and business longevity (Bloom, Eifert, Mahajan, McKenzie, and

[1] This intervention was a pilot within the broader Promoting Micro and Small Enterprises through Improved Entrepreneurs' Access to Financial Services (PROMISE-IMPACT) initiative of the ILO in Indonesia.

[2] https://www.socialscienceregistry.org/trials/3625 *Note:* Our PAP builds on our original intention to analyze data from thirteen participating FSPs. However, one FSP, a development bank, changed its offer of treatment arms and it is unclear what services were offered to whom at what time. Therefore, we disregard the data from this FSP in this paper.

[3] Credit cooperatives also operate predominantly in rural areas. The difference between the institutions is foremost in the ownership structure.

Robert 2013, McKenzie and Woodruff 2015). But causal evidence on how to improve business practices is inconsistent, which is little surprising given that the evaluated interventions differ substantially. Whereas in an early work, Karlan and Valdivia (2011) find no overall impact of training on profits or employment, Drexler, Fischer, and Schoar (2010) find a significant impact in a rule-of-thumb training, and Lafortune, Riutort, and Tessada (2017) find that visits of role models can increase training impact. McKenzie and Woodruff (2013) provide a comprehensive overview of similar evaluations and the general caveats in providing classroom training. In practice, achieving large outreach in a cost-efficient manner is another important dimension. In that respect, the ILO program we evaluate has distinct advantages: training existing clients of FSPs ensures that access to finance is not another binding constraint, using well-tested training material guarantees its high quality, and cooperating with loan officers as trainers/counselors establishes cost-efficiency while maintaining incentive compatibility. On the downside, trainings are short and not provided by experienced trainers, so the average effects were expected to remain small. To account for this, we use a comparably large sample size of 3,975 clients (thereof 2,650 in treatment groups).[4] That we nevertheless find no average effects on business practices, suggests that the training-of-trainers (ToT) approach with loan officers is not a successful training model, and/or the SIYB training content was ill suited for this setting.

Regarding the SIYB, to our knowledge, De Mel, McKenzie, and Woodruff (2014) and Fiala (2013) provide the only two rigorous evaluations of applications of ILO's SIYB, in Sri Lanka and Uganda respectively. De Mel et al. (2014) find changes in practices for existing businesses, but no impact on financial outcomes over a period of two years, except for some positive effects on the subgroup of start-up entrepreneurs. Fiala (2013) combines the training treatment with either a loan or a grant, hence tackling the constraints in finance and managerial capital at the same time. He finds significant increases in profits six months after the intervention ended, but only for male participants and only when combined with loans and not grants.[5] While these two studies show that the intervention can positively impact at least a subset of participants, there are fundamental differences in the studies' set-ups. First, our sample differs from the existing study by focusing on existing enterprises, both female- and male-owned, who are borrowing at commercial rates, which may imply that they are already established businesses. Second, our treatment arms differ from previous studies by including individual counseling sessions, but no additional financing. This is inspired by findings of Valdivia (2015), who finds a stronger impact on monthly sales of combined training with counseling treatment as opposed to only training treatment in Peru.[6] Third, to our knowledge we are the first to cooperate with twelve

[4]As McKenzie and Woodruff (2013) point out in their review study, most existing studies face challenges regarding their power. With a sample size typically ranging between 300 and 500 individuals in a treatment group and these being quite heterogeneous in baseline outcomes, the authors estimate the power to detect a 25% increase in revenues to be as low as 7% and seldom above 25%.

[5]Much of the analysis of heterogeneous effects in the literature has focused on the gender divide. In this regard, Berge, Bjorvatn, and Tungodden (2015) and Gine and Mansuri (2014) both find that male participants are more likely to benefit from training in terms of business outcomes, with neither study finding significant changes in income or assets for women. Focusing exclusively on women, Field, Jayachandran, and Pande (2010) find training increased business income for upper-caste Hindu women in India, but not for lower-castes or Muslim women. In our study we cannot confirm the findings in previous literature where the need for technical assistance was divided along the lines of gender, educational background, age of the business, or size as measured in profit or loan volume.

[6]Also see Bloom et al. 2013; Bruhn, Karlan, and Schoar 2013; Karlan, Knight, and Udry 2015; Lafortune et al. 2017 for the effect of counseling interventions.

different FSPs to provide the training and/or counseling.[7]

The finding that impact differs between types of FSPs and individual FSPs is an important contribution to the literature, where RCTs are likely to work with institutions which differ in motivation and general management from other potential implementing sites.[8] But any training program that aims to achieve large outreach is prone to work with several implementing partners to overcome capacity constraints. As we show in our study, partner choice might well be the bottleneck in upscaling. This might be due to limited absorptive capacity in the institutions (e.g. limited human or managerial capital), or due to the lack of control inherent to longer impact chains. Whereas our study is not suited to identify institutional characteristics important for success, we highlight the importance of carefully discussing partner choice in similar interventions.

We organize the rest of this paper as follows. In Section 2 we describe the intervention. In Section 3 we detail the experimental design, including a description of the sample and implementation fidelity. In Section 4, we present our estimation method. Our results are contained in Section 5 for average treatment effects on intermediate and final outcomes. Section 6 contains results of our heterogeneity analysis by entrepreneur and business characteristics, treatment arms, and across FSPs. The last section concludes.

2 The program and intended effects

2-1 Micro and small enterprises in Indonesia

Considered an emerging middle-income country, Indonesia has managed to cut its poverty rate substantially from 19% in 2000 to below 10% in 2018 (Statistics Indonesia 2018). Yet, of the over 59.27 million firms, 99.89% were MSEs in 2014 (OECD 2018).[9] Their relatively small contribution to the gross domestic product (GDP) (43%) illustrates the MSEs' low productivity (International Labour Organization 2018), which persists despite a healthy macroeconomic environment and light taxation (OECD 2018).

Indonesia's large and innovative MSE finance sector has been recognized as a global leader for decades (OECD 2018, Rosengard and Prasetyantoko 2011).[10] However, the fact that only few

[7] Swain and Varghese (2013) show that training delivery mechanisms affect the impact of the training with training provided by non-governmental organizations achieving greater impact than training provided by government officials in bank linkage groups in India. Their setting is however non-experimental and potentially suffering from self-selection.

[8] For example, Allcott (2015) find microfinance institutions cooperating with the Jameel Poverty Action Lab, Innovations for Poverty Action, and Financial Access Initiative to be older, larger, and more likely for profit. As he then demonstrates in the context of an energy conservation program, site selection can lead to systematically biased estimates.

[9] By law 20/2008, the Ministry of Cooperatives and SMEs characterizes micro (small) enterprises as having net assets below 50 (500) mln IDR or annual revenues below 300 mln (2.5 bln) IDR. The Central Bureau of Statistics (Badan Pusat Statistik, BPS) follows an employment-based definition, with microenterprises employing 1-4 people and small enterprises 5-19 people. The clients in our study fall under MSEs in both definitions.

[10] Most famously, Bank Rakyat Indonesia is one of the largest microfinance service providers in the world with more than 4,400 units serving the rural population.

entrepreneurs transition from micro to small businesses suggests that factors other than financing also inhibit growth.[11] Potentially, some of the microenterprises are *necessity entrepreneurs* who start their own business because they lack formal employment opportunities. These individuals are less likely to identify promising investment opportunities or to manage business growth effectively (Atmadja, Su, and Sharma 2016; Aldianto, Rudito, Mirzanti, Situmorang, and Larso 2010). The productivity of microentrepreneurs in Indonesia might hence also be inhibited by a lack of managerial capital, i.e., the owner's ability to use the firm's capital and labor resources most efficiently (Bruhn, Karlan, and Schoar 2010). Another way to assist growth of MSEs in Indonesia is hence to improve their productivity by combining access to financial services with business advice or training.

2-2 The program

With the aim to increase entrepreneurial skills of MSE owners, the ILO partnered with FSPs in East Java and West Java. Financed by the Swiss development agency State Secretariat for Economic Affairs (SECO) with 3.1 million USD, the program geographically focused on West and East Java, an area with a high number of manufacturing MSEs, predominantly in the textile as well as food and beverage industries. More than 150 loan officers of 12 participating FSPs were formally trained to become trainers and/or counselors. These loan officers then gave classroom training and/or individual business counseling to clients selected by our randomization procedure (see below). The intended increase in profitability of the clients' enterprises is expected to benefit (1) the clients directly through greater profits and increased household expenditures, (2) the FSPs through improved loan repayment and larger loan sizes as well as (3) the broader economy through job creation.

In the ToT approach, the ILO relied on its existing methodology, the SIYB. Now one of the largest business training programs worldwide, it has its roots in the 1970s and has experienced constant revision and refinements for its implementation around the world (International Labour Organization 2013).Its ToT was an intensive ten-days course which focuses on concepts such as generating a business idea, determination of costs and prices, bookkeeping, financial planning, stock control, and purchasing/buying. The training also taught adult learning methodologies, including the delivery of practice sessions by the participants, various simulation games, and exercises for presentation skills, concluding with a certification test. The ToT schedule followed in the intervention is attached in Appendix 1-1.

The newly trained trainers then conducted classroom training courses for their clients who were randomly selected by us and individually invited by phone or in person by the loan officer. Up to 20 clients participated in a training session held in a location close to the clients' residence. Given the priorities determined by the FSPs in a previous needs assessment, trainers focused

[11] Shedding some more light on this, Cole, Sampson, and Zia (2009) investigate whether a low demand for financial services in Indonesia is rational (prices are higher than productivity) or constrained by information asymmetries (lack of financial literacy). Their results rather favor the former, given that financial literacy training only marginally affected demand.

on financial management and marketing in their courses for clients.[12] The loan officers trained the clients only once in a two-day training session which ends by asking clients to prepare their business plan. In the "training only" treatment arm, no further advise to carry out their plans was provided.

In contrast, the training-of-counselors (ToC) ran for five days as an adaptation of the SIYB specifically undertaken for this intervention. In addition to the technical skills of marketing and financial planning, the participants learned about the fundamentals of adult learning and counseling, but were not trained to become classroom teachers. In Appendix 1-2 we include the schedule of the five-day ToC training. Loan officers undertook individual counseling sessions at the client's premise, mostly combined with routine loan collection visits. Counseling of clients happened semi-structured: While all loan officers started with a review of the MSE's past activities compared to the ones in the initially agreed-upon business plan, subsequent counseling consisted of individual advise and encouragement as needed.

2-3 Expected impact chain

The intervention builds on the following rational: Participation in training and/or counseling increases knowledge in the focus topics marketing and financial planning. The increased knowledge leads to behavioral changes in business practices among participants. We refer to participation, increased knowledge, and changed practices as *intermediate outcomes*. These outcomes are then expected to affect business outcomes, which in turn lead to better lives for the entrepreneur and her household, as well as to improved loan behavior, which also benefits the FSPs. We refer to business and households outcomes as well as loan behavior as *final outcomes*.

Prior to the follow-up survey, we formulated sixteen hypotheses which we summarized in a PAP, registered on the American Economic Association (AEA) website (RCT ID: AEARCTR-0003625), downloadable at www.socialscienceregistry.org/trials/3625.

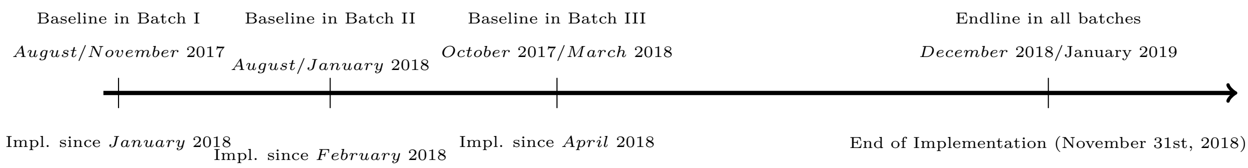
3 Experimental design

We evaluate the impact of the intervention using an RCT implemented by twelve different FSPs. Following a baseline survey, we randomized clients into different treatment or the control group. Subsequently, the intervention started and eight to sixteen months after the baseline, we conducted an endline survey interviewing 3,975 clients.

[12]The ILO supported the FSPs in conducting (not necessarily representative) client surveys in autumn 2016 in order to understand and prioritize the needs of their clients. A total of 2,405 clients were interviewed to this purpose. A staggering 82% believed business support services to be important, yet only a small part had received training and mostly only related to repayment of loans; a topic they themselves deemed of little importance. In contrast, clients prioritized business support services regarding marketing, product quality improvement and financial management. The self-reported willingness to pay for these services is, however, low, with more than half of the respondents stating that they would not pay any fee.

As the ILO formalized the partnerships with the various FSPs at different times, we divided the FSPs into three batches. The six FSPs of Batch I are savings and loan cooperatives, the four FSPs of Batch II and the two FSPs of Batch III are banks. We completed the randomization between January and April 2018, and the implementation of the intervention started shortly after, as illustrated in Figure A.1.

Figure 3.1: Timeline of baselines and randomization



Note: This figure displays the timing of the three waves of randomization, and of the endline survey.

3-1 Baseline survey and sample characteristics

Participating FSPs are six savings and loan cooperatives, five rural banks, and one regional development bank, which all predominantly - but to varying degrees - focus on rural areas. They exhibit similarities in their mission (including commitment to the double bottom line) and scope of operations targeting MSEs in East and West Java. At the same time they are considerably diverse regarding their client outreach (from just under 1,000 to over 433,000 active borrowers, of which 30% to 100% are female),[13] lending methodology (group and individual loans, partly Sharia compliant), and professional experience (founded between ten and over sixty years ago). Table C.1 in Appendix 1 contains some of the key characteristics of the FSPs.

Table 3.1 contains summary statistics from the baseline survey, averaged over all 3,975 clients which we were able to re-interview in the endline survey (hereforth referred to as *estimation sample*). Columns (1) and (2) present the mean and standard deviation of each variable, followed by the minimum (3) and the maximum value (4). Cleaning of baseline data in particular involved the removal of duplicates, observations with missing contact information, missing age or gender, or non-response in more than six items.[14] We accounted for outliers and measurement errors by winsorizing at the 90th, 95th or 99th percentile, depending on the initial variation in the variable.

[13] In our study, the number of participating clients varies between FSPs, from 178 to 582 clients in treatment and control group.

[14] If less than seven items were missing, we imputed the answer using regression imputation based on the set of complete observations.

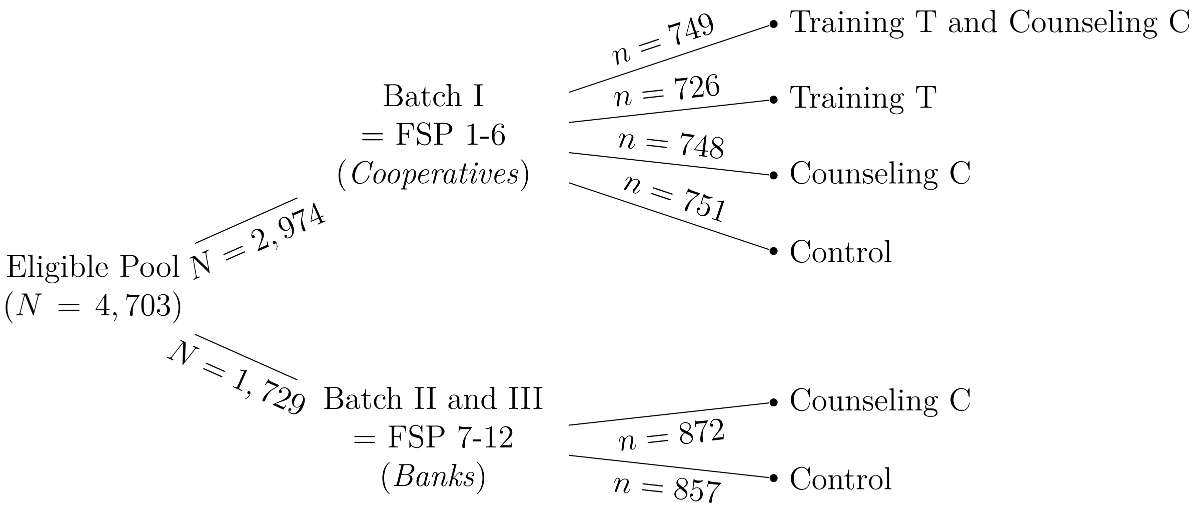
Table 3.1: Baseline characteristics of estimation sample

	Mean	Stand. Dev.	Min. Value	Max. Value
	(1)	(2)	(3)	(4)
Panel A: Client Characteristics				
Share of female clients	0.683	0.466	0	1
Average age of client	44.151	9.223	19	65
Total nr. of people in the household (HH) btw 18-65yrs. (incl. the client)^{99p*}	2.810	1.056	1	6
Share of clients having no education	0.033	0.178	0	1
Share of clients having primary school education	0.300	0.458	0	1
Share of clients having secondary education	0.199	0.399	0	1
Share of clients having vocational education	0.344	0.475	0	1
Share of clients having no add. income	0.570	0.495	0	1
Share of clients having add. income from another business	0.258	0.438	0	1
Panel B: MSE Characteristics				
Nr. of yrs. the business exists^{99p*}	11.646	9.064	0	39
Share of MSEs not registered	0.824	0.381	0	1
Share of MSEs selling directly at the market	0.542	0.498	0	1
Share of MSEs selling through agents	0.164	0.370	0	1
Share of MSEs that are part of the HH	0.696	0.460	0	1
Share of clients preparing a business/financial plan	0.407	0.491	0	1
Share of clients keeping record of all transactions	0.348	0.476	0	1
Share of clients keeping business and HH finances separately	0.517	0.500	0	1
Share of clients investing profit into the business	0.642	0.480	0	1
Total nr. of non-family, permanent workers^{99p*}	1.104	2.554	0	17
Panel E: Loan Behavior				
Average number of loans per year	0.836	0.622	0	10
Share of clients reporting being late with the loan payment at the FSP	0.177	0.381	0	1
Share of clients want to borrow more for business	0.723	0.447	0	1

- *Note*: This table contains selected summary statistics from the baseline survey. The table shows selected baseline variables on the left, the different descriptive statistics on top.
- Sample: Estimation sample, i.e., clients from 12 FSPs whom we were able to re-interview in the endline survey (N = 3,975).
- Source: Baseline survey (2017 - 2018).
- Columns (1) - (2) display the mean and standard deviation. Columns (3) - (4) display the minimum and maximum values.
- The superscript np indicates the winsorizing level. We winsorized variables per batch prior to randomization and following an automated rule to define percentiles. * indicates that we re-defined the winsorizing level manually.
- Table C.2 in Appendix 2 presents selected summary statistics for the three different types of FSPs individually. Table C.3 in Appendix 2 presents summary statistics on all baseline variables, also for the full baseline sample (with attrited households), along with a number of tests for imbalances between treatment and control group.

The average age in our estimation sample is 44 years, about two thirds of our interviewed clients are female and the average household contains around three adults, of which two earn an income. Clients who completed a vocational training form the largest group in our sample with 34.4%. Notably, 57.0% of the interviewed clients earn no additional income themselves, besides their business. These businesses are typically, but not exclusively, informal microenterprises operated from the same location as the household lives in and selling their products mainly directly at the market or as single merchant. Whereas the average business has been operational for 11.7 years, there is a lot of variation in this variable, and our sample covers businesses between 0 and 39 years of age with a median of 9 years. On average, a client employs 1.1 non-family, permanent workers in her business, and costs for raw material outweigh worker's salaries by

Figure 3.2: Random assignment of clients to treatment arms



Note: This figure displays the allocation of clients to treatment arms. Cooperatives (Batch I) offered three treatment arms, whereas the banks (Batch II and III) offered only counseling treatment. Within an FSP, equal number of clients were assigned to the offered treatment or the control group.

the factor five. Consistently, the largest share of clients (33.8%) report expensive raw materials as the main barrier to business.

Table C.2 in Appendix 2 contains the summary statistics by type of FSP separately. These show that cooperatives have the largest share of female clients (26.1% in the development bank, 45.9% in rural banks, and 82.9% in cooperatives), and the smallest share of clients with vocational education (51.1% in the development bank, 42.1% in rural banks, and 29.5% in cooperatives). They are also most likely to serve businesses that are part of the household (43.4% in the development bank, 54.8% in rural banks, and 79.0% in cooperatives) and their clients report lower revenues (45.4 mln IDR in the development bank, 19.6 mln IDR in rural banks, and 12.9 mln IDR in cooperatives). However, the profit to revenues ratio is actually largest in cooperatives (22.9% in the development bank, 29.2% in rural banks, and 31.5% in cooperatives).

3-2 Randomization and baseline balance

We offered the participating FSPs a choice of different treatment arms: (1) training treatment arm (T), (2) consulting treatment arm (C) and (3) training and consulting treatment arm (TC). Whereas the six FSPs of Batch I (cooperatives) chose to randomize all three treatment arms, the six FSPs of Batch II and Batch III (banks) offered only the C treatment arm. We randomly allocated individuals to one of the treatment arms offered by their respective FSPs or to the control group. Figure 3.2 illustrates the assignment of observations to the different treatment arms.

In allocating treatments, we followed a re-randomization approach whereby we retained the first randomization vector that passed a balancing threshold. In Table C.3 in Appendix 2, we show summary statistics testing the balancing of the randomization variables between treatment and control groups for all batches jointly. We show t-test and F-test statistics for the full baseline sample and the estimation sample (i.e., net of attrition), and find no systematic differences in observable baseline characteristics between treatment and control group.

3-3 Endline survey and attrition

We collected endline data up to 1.5 years after the baseline. In the endline, we were able to reinterview 3,975 clients, whereas 728 clients could not be located or refused the interview. Although an attrition rate of 15.5% seems quite high at first glance, it is within the range of similar studies (5.3% in Field et al. (2010), 8% in De Mel et al. (2008), 24% in Karlan and Valdivia (2011), 26% in Calderon, Cunha, and De Giorgi (2013) as illustrated in McKenzie and Woodruff (2013)). Clients from the control group were 3.2 percentage points less likely to participate in the endline survey than clients from the treatment group and this difference is weakly significant. This is in line with the fact that the attrition rate was higher in East Java, where most of the banks are located, while at the same time bank clients make up a higher share in the control than in the treatment group. Figure 3.3 illustrates this point. Much of this difference is hence captured by later controlling for FSP fixed effects.

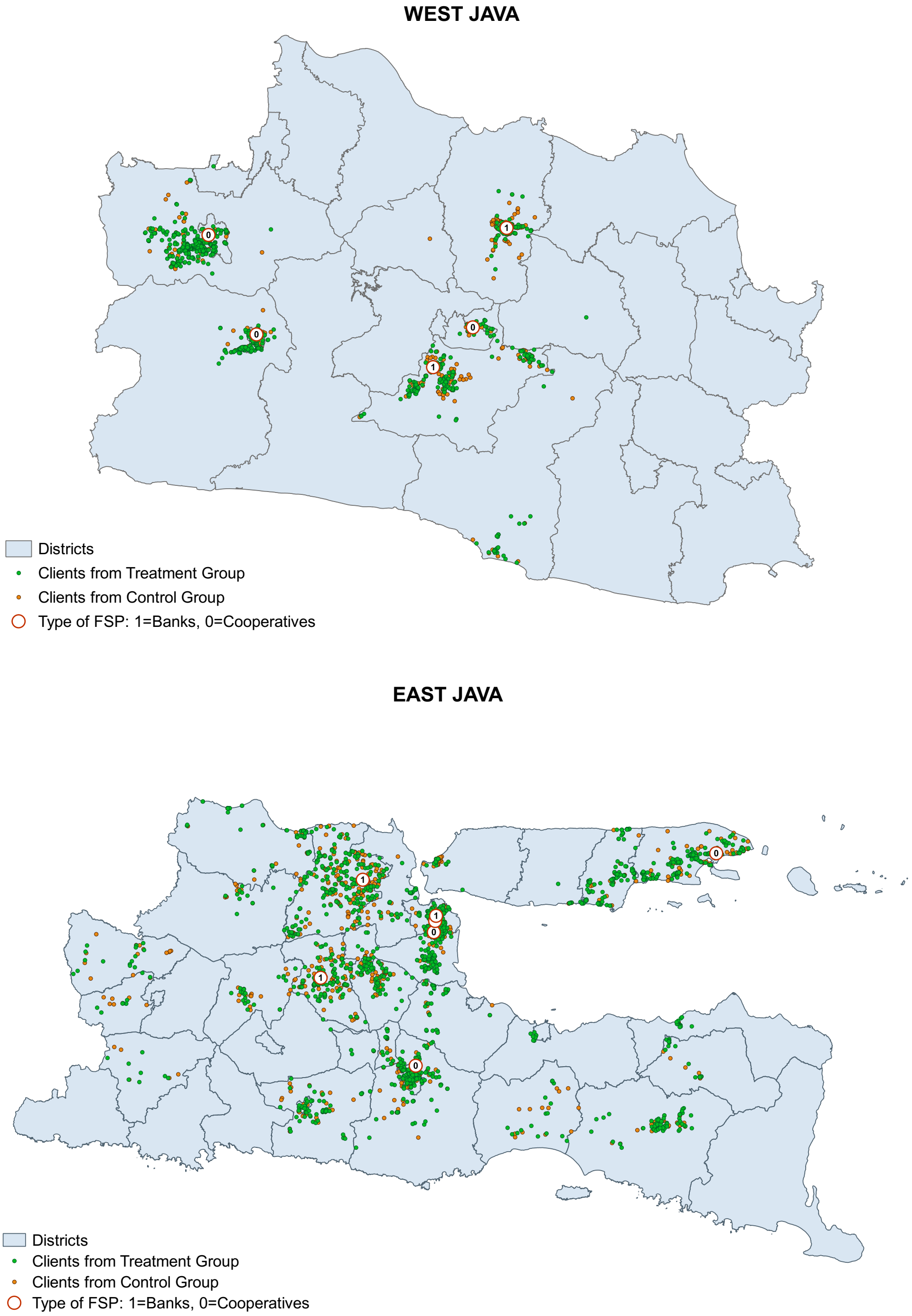
3-4 Monitoring and implementation fidelity

Following randomization, loan officers invited the respective clients to classroom training and/or individual counseling. In the endline survey, we asked whether clients were aware of, had been invited to, and participated in one of the treatments. Since the intervention was advertised differently across FSPs and generally not associated with ILO by clients, we asked generic questions about participation in trainings or counselings.[15] Among the treatment group 57.6% reported being aware of the program. Conditional on being aware, 79.7% were invited to participate. Conditional on being invited, 72.4% of clients participated in the program.[16] As counseling sessions were conducted as part of usual loan collection routines, our survey likely insufficiently captures the implementation of the counseling treatment arm. Indeed, awareness lies at 39.8% in the counseling group, whereas it reaches 74.7% in the training only and 78.9% in the training and counseling group. There is also a rather high contamination among the

[15] We asked all clients: "Are you aware of the existence of services such as classroom training and individual counseling that [*FSP name*] is providing?", where we inserted the FSP name from the baseline survey. Only if the answer was affirmative, we asked the next question: "In the following, we will refer to this programme as *business development program* offered by your financial institution. Were you offered such training or counseling on business development as aforementioned?" Only if the answer was affirmative, we asked: "Did you participate in classroom training or individual counseling sessions or both?" For all three questions, the answer categories were Yes, No, Refused to Answer, Do not know/Not applicable.

[16] The most common reasons for non-participation were that the client could not leave the business unattended (40.5%), or had to attend to household and child care duties (18.8%). Note that female clients were much more likely to participate at treatment when offered: 75% of female clients invited to treatment took up the offer, as opposed to only 52% among male clients. This finding might also be due to the differential reporting as cooperatives offering three treatment arms also have a larger share of female clients. Unaffected by this, we also cannot confirm the finding by Valdivia (2015), where women with young children are less likely to participate than women without young children.

Figure 3.3: Location of clients (12 FSPs)



Note: The maps illustrate the geographic location of the clients as captured in our endline survey within East and West Java. The relative share of control group clients is larger in East Java than in West Java, because rural banks are mostly located here.
▶ Source: Endline survey data (2018 - 2019).

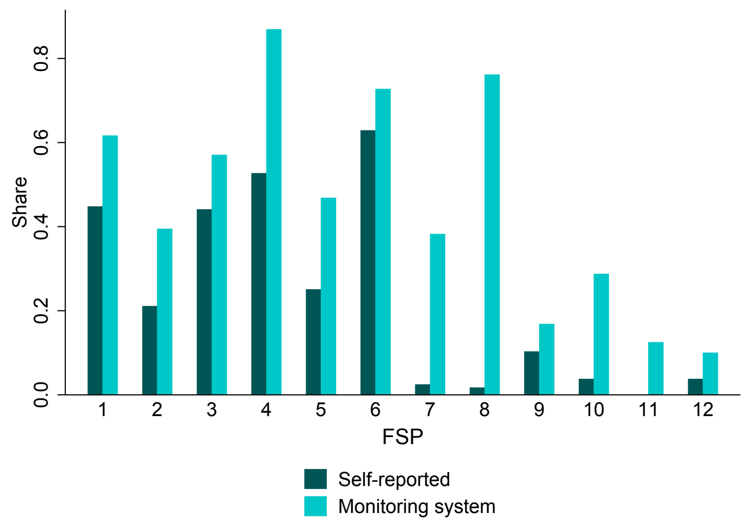
control group, with 13.3% stating that they participated in training and/or counseling.[17] Due to these potential measurement errors, we use initial assignment as main treatment indicator for our analysis and focus solely on ITT effects.

We monitored the implementation process via an online platform starting in April 2018, where

[17]This rate is highest among cooperatives (25.7%) and low among banks (2.15%), and might at least partially be due to the fact that cooperatives also offer other non-financial services.

loan officers filled questionnaires directly after training/counseling sessions for each participating client separately. As loan officers filled questionnaires voluntarily, we have monitoring data only on about half of our treatment group.[18] The degree to which loan officers complied with the monitoring system varies greatly across FSPs. Figure 3.4 illustrates this fact, showing the participation rates among clients assigned to treatment group by FSP. The first bar uses survey data, whereas the second bar uses monitoring data. Note that FSPs 7 to 12 offered only the counseling treatment, where survey data suffers from the above described shortcomings.

Figure 3.4: Participation Rate among Treatment Group by FSP (Survey Data and Monitoring Evidence)



Note: The figure displays the participation rate among the group of clients initially assigned to treatment, disaggregated by the twelve FSPs. The first bar indicates participation as reported by clients in the endline survey. The second bar indicates participation as reported by loan officers in the monitoring system.
► Source: Monitoring data and endline survey (2018 - 2019).
► Note that FSP 1 to 6 offered three treatment arms, whereas FSP 7 to 12 offered counseling only. Self-reported participation in the counseling sessions likely suffers from larger measurement error, likely due to the fact that counseling was done as part of regular loan collection routines.

From Figure 3.4 it is evident that some FSPs were more engaged in the intervention than others, with especially low implementation fidelity in FSPs 9 to 12. This experience is similar to that presented in the work by Karlan and Valdivia (2011): In an entrepreneurship training intervention at different village banks of FINCA Peru only half of the partner banks reached the 17 out of 22 envisioned sessions within two years. The authors also report that these delays are typical for similar interventions and conclude that analysis should focus on intention-to-treat effects to avoid selection bias. While we agree with this conclusion, we analyze in more depth how results differ across institutions in Section 6-3.

Qualitative interviews led by ILO in a roundtable workshop in March 2019 shed light on the reasons for incomplete implementation. Whereas feedback was generally positive, six of the twelve FSPs cited the large geographic spread of the clients as a key challenge, mostly in combination with the limited time available to loan officers. Four FSPs said that some clients

[18] Specifically on 349 clients who received at least one training session, 868 clients who received at least one counseling session and 223 who clients received at least one training and one counseling session.

were little motivated and too passive during counseling, which they expect could be improved by better targeting. Three FSPs reported high drop-out rates of clients who finished their last loan cycle, or loan officers shifting their duty station or resigning after the ToT/ToC. The feedback received does not fully correlate with the different level of compliance illustrated in Figure 3.4. That is, FSP 8 named more challenges than any other FSP, and yet complied with the protocol as per monitoring data, while FSP 10 named the most observed benefits (such as improved reputation and increased loan sizes), but we cannot confirm implementation fidelity with our monitoring data.

Finally, wherever monitoring data is available, we can analyze the nature of the intervention delivered. The intervention design with respect to its content and structure was mostly adhered to, except for the number of counseling sessions which did often not reach the intended five. Two thirds of the participants were actively engaged and questions were answered satisfactorily by the trainer. Correspondingly, the endline survey shows that most of the respondents find the intervention helpful and recommendable. Survey evidence also confirms that the topics covered in classroom training and counseling sessions were largely similar.

4 Estimation method

We focus on calculating ITT effects and regress the relevant outcome variable on a dummy variable indicating treatment assignment, three sets of control variables, and two types of fixed effects, leading to the following main model specification:

$$Y_{i,t=1} = \beta_{ITT} D_i + \bar{X}_i \beta_3 + X_i \beta_2 + Miss X_i \beta_1 + \mu_{FSP_i} + \mu_{Enu_i} + \beta_0 + \epsilon_i. \tag{3.1}$$

Here, $Y_{i,t=1}$ is the outcome variable for entrepreneur i at endline ($t = 1$), D_i is a dummy variable indicating treatment status (i.e., assigned to any of the three treatment groups or control group), and the parameter of interest is the vector β_{ITT}, the intention to treat effect. It gives the average difference in means between the combined treatment group and the control group and is interpreted as the impact of being officially eligible for the training and/or counseling sessions.

In our main specification, we include the following covariates: First, the set $\bar{X}_i$ comprises the covariates age and gender, which we control for in all regressions. Second, we denote by X_i the vector of randomization strata on client level. While in our randomization procedure, we ensured batch-wise balance on a total of 65 baseline variables, we include in the regression those twenty baseline variables, which show the highest imbalance within the respective sample considered. This implies that different covariates X_i are used for different subsamples.[19] Third,

[19]We have used 65 baseline variables for the randomization, i.e., we ensured that treatment and control group are balanced on these variables. However, this was done by batch, so when combining the whole sample together and excluding attrited households, there may be some imbalances on baseline characteristic. We cannot control for all of them due to their large number. Instead, we automatically select those twenty covariates which have the greatest normalized mean difference between treatment and control

there are a few cases of item non-responses in some baseline randomization variables in the raw data. We impute the missing values by estimating a regression model and include the imputed values in the vector X_i. To account for this imputation, we include a set of indicator variables $MissX_i$ which are equal to 1 if a variable in a specific group has been imputed (to account for collinearity). Fourth, we include fixed effects for FSPs and enumerators. The former account for variation within the FSP, which stem from systematic differences in the client base, differential commitment to the intervention, loan officer qualifications, or other unobservable factors. The latter account for reporting differences across enumerators.[20] We present robust standard errors, ϵ_i.

We conduct additional robustness checks, described in Section 3-2 in the appendix. Among other robustness checks, we omit FSP fixed effects and cluster standard errors at FSP level, include batch instead of FSP fixed effects, use ten or thirty most imbalanced control variables, add other control variables, and estimate effects on unwinsorized data. We also repeat the analysis on the subsample of observations with only non-missing values, in which case $MissX_i$ is dropped. Additionally, we calculate a model specifications without enumerator fixed effects and one where we drop data from five enumerator who report at least one outlier value in more than 40% of their respondents. While inference on weakly significant estimates changes with the model specifications, our highly significant results are robust.

In addition to comparing the combined treatment group to the control group, we estimate average treatment effects for each treatment arm separately and test whether there are significant differences between these. We also look at heterogeneous effects across sample sub-groups: Denoting by R_i an interaction covariate, i.e., an indicator of the individual, business or FSP characteristics, we estimate the following regression model:

$$Y_{i,t=1} = \beta_{ITT} D_i + \gamma_1 R_i + \gamma_{IA} D_i \times R_i + \bar{X}_i \beta_3 + X_i \beta_2 + MissX_i \beta_1 + \mu_{FSP_i} + \mu_{Enu_i} + \beta_0 + \epsilon_i. \quad (3.2)$$

In this model, the parameter of interest is γ_{IA}, which provides the differential impact of treatment for different values of the interaction covariate. We consider as variable R_i only credibly exogenous pre-treatment variables.

group. For our full sample which pools all FSPs these variables are: indicator for clients having add. income from full time job, average number of loans per year, indicator for stating tough competition as main barrier, last loan amount in mln Indonesian Rupiah (IDR), indicator for having last loan as business/ individual loan, indicator for stating that business brings high income, indicator for not stating any business barrier, indicator for having written contracts with the workers, indicator for having positive spending on durables, indicator for having no additional income, indicator for wanting to borrow more, indicator for business not being registered, indicator for business being active throughout the year, indicator for business being part of the household, cost per day for workers salaries, indicator for stating that business brings respect, indicator for having a university degree, indicator for keeping records of all transactions.

[20] This model slightly deviates from the one specified in the PAP. Here, we had included batch fixed effects and also intended to use the full set of randomization variables. We calculated these alternative models as robustness check and find inference to remain unchanged. The reason we had to deviate from the PAP model is the stark, not anticipated variation in the implementation of the treatment between the FSPs. By controlling for 12-1 FSP fixed effects instead of only 3-1 batch fixed effects, we believe that we can capture more variation within each FSP, reducing a potential omitted variable bias, and thereby estimating a more restrictive model. The enumerator fixed effect also make a difference in results and should hence be included in all estimations. In the PAP we had proposed this as a robustness check, but decided to show the more conservative estimates as main results.

5 Average treatment effects

5-1 Intermediate outcomes

In line with the expected impact chain laid out above, we first discuss whether the program affected intermediate outcomes on knowledge and business practices. Table 3.2 contains the results. Columns (3) and (4) contain the ITT point estimates and standard errors, outcome variables are displayed on the left.[21]

Table 3.2: Average treatment effects on intermediate outcomes

Outcome	N	Control Mean (1)	Control SD (2)	Estimates β_{ITT} **(3)**	Estimates SE (4)	
Panel A: Program Participation						
Share of clients reporting being aware of the support that FSP is providing	3895	0.355	0.479	**0.065**	0.014	****
Share of clients reporting having been offered support from FSP	1945	0.680	0.467	**0.070**	0.022	***
Share of clients reporting having participated in the program	1485	0.570	0.496	**0.141**	0.030	****
Panel B: Knowledge						
Share of marketing knowledge questions answered correctly	3960	0.382	0.311	**0.002**	0.010	
Share of financial management knowledge questions answered correctly	3966	0.647	0.316	**0.000**	0.010	
Share of clients not knowing the payment type	3303	0.114	0.318	**-0.015**	0.011	
Panel C: Business Practices						
Marketing practices index$^{n.m.}$	3975	1.448	1.592	**0.018**	0.049	
Nr. of different marketing forms used$^{n.m.}$	3975	0.245	0.511	**0.017**	0.017	
Financial management practices index$^{n.m.}$	3975	2.824	2.347	**-0.003**	0.072	
Share of clients keeping business and HH finances separately	3611	0.428	0.495	**-0.019**	0.017	
Share of clients investing profit into the business	3599	0.655	0.476	**0.002**	0.016	
Share of clients preparing a business/financial plan	3609	0.429	0.495	**0.022**	0.016	
Business plan includes cash flow	1543	0.073	0.260	**0.029**	0.014	**

- *Note*: This table shows intermediate outcome variables on the left, statistics on top. Same clients are followed over time.
- Sample: Estimation sample (N = 3,975). Sample size varies between variables due to item nonresponse and conditional questions.
- Source: Endline survey (2018 - 2019).
- Columns (1)-(2) display the control group mean and standard deviation. Columns (3)-(4) present regression results, controlling for age, gender, the twenty most imbalanced baseline covariates, dummies indicating imputation in baseline variables, enumerator and FSP fixed effects. We present robust standard errors.
- The superscript *n.m.* indicates that missing values were interpreted as zero to generate index variables.
- The statistical significance is given as follows: * indicates p < 0.1, ** p < 0.05, *** p < 0.01, **** p < 0.001.
- See Section 3-2 in the appendix for robustness checks.

As expected, clients from the treatment group were significantly more likely to be aware of the program, to have been offered the support from the FSP, and to have participated in the program. Note that relatively high control means for invitation and participation variables are due to these variables being measured conditional on awareness and do not reflect the average across the whole control group.

To test whether treatment improved knowledge on the training topics, we asked a set of three questions on marketing and financial planning each. The questions were taken directly from the original ILO training material. The main outcome variables of interest are the shares of questions answered correctly in each of the two training topics. We find no significant effects on these outcomes, with the point estimates being very close to zero. Note that in the control group, knowledge on financial management was wider-spread than on marketing: The average control group client answered two questions on financial management correctly, but only one on marketing. We further tested whether treated clients are more likely to know their basic loan

[21] To present results more concisely, we deviate from the PAP by neglecting the following outcomes: whether the business is exporting outside of Indonesia, is registered, and is paying tax. We also do not show estimates on content of business plan other than cashflow. There are no effects on any of these outcomes. We also do not show estimates on individual indicators which we aggregated as indices.

conditions, in particular their payment modalities (interest rate, profit sharing agreement), but find no effect.

To test whether treatment led to behavioral changes in marketing and financial management practices, we asked two batteries of five questions each regarding two training topics. These questions were taken from De Mel et al. (2014), who evaluate the SIYB training intervention in Sri Lanka. The main outcome variables here are the shares of best practices followed. Marketing behavior is summarized in the marketing practice index that ranges from 0 to 5 with higher values representing the use of more desirable practices (e.g., visiting competitors, asking customers, using any kind of advertisement). Questions related to financial management (e.g., frequency of reviewing the financial performance of the business) are summarized in the financial management index that ranges from 0 to 7, with higher values being more desirable. Contrary to De Mel et al. (2014) we find no significant effect of the treatment on these practices on average, which may be due to a different target group. We ask additional questions on marketing and financial management practices; specifically, the number of marketing channels used, whether the client keeps business and household finances separate, whether profits are re-invested into the business, and whether clients prepare a business/financial plan. We also fail to detect changes in these outcomes, except for a significant and positive increase in the clients whose business plan includes cash flow analysis. This effect is robust across model specification, as shown in Section 3-2 in the appendix. Given that in MSE finance, it is important to match repayment schedules with the business' cashflow, loan officers might have stressed this aspect in their training and/or counseling. This finding is hence plausible and also important for leveraging the whole potential of access to finance.

5-2 Final outcomes

The program aimed to achieve impact on three levels: Business outcomes, living standard and FSPs financials. Accordingly, we group final outcomes in these three categories. Table 3.3 contains the results. As before, Columns (3) and (4) contain the ITT point estimates and standard errors, outcome variables are displayed on the left.[22]

Given the very limited effects we find on intermediate outcomes, it is little surprising that we find no indication of business growth or increased profitability. Specifically, we find no effect on the log of revenues, costs or profits during the last thirty days or on the number of permanent workers.[23] Our quantitative measurement are heavily affected by noise, which we already expected based on similar studies (De Mel et al. 2014). To account for this, we also measured two binary outcomes: Whether earnings cover business expenses and whether profits increased in the last six months. We find no effects on these variables either. We further include

[22] Again, in the interest of a concise presentation, we deviate from the PAP by neglecting minor outcomes such as the number of casual and permanent workers, and expenses on durable assets. There are no effects on any of these outcomes.

[23] We log-transformed quantitative variables due to their screwed distribution. Results are robust when taking the original variables.

Table 3.3: Average treatment effects on final outcomes

Outcome	N	Control Mean (1)	Control SD (2)	Estimates β_{ITT} **(3)**	Estimates SE (4)
Panel A: Business Financials and Attitudes					
Log of revenue in the last 30 days in mln IDR^{99p}	2701	1.869	1.445	**0.043**	0.052
Nr. of permanent workers	3626	1.883	4.687	**-0.155**	0.144
Log of total cost of business in the last 30 days in mln IDR^{99p}	3327	1.361	1.804	**-0.030**	0.060
Log of profit generated in the last 30 days in mln IDR^{99p}	2810	1.050	1.357	**0.014**	0.050
Log of profit in the worst month mln IDR^{99p}	2803	0.612	1.365	**0.001**	0.052
Log of profit in the best month in mln IDR^{99p}	2997	1.468	1.418	**0.048**	0.050
Share of business earnings covers the exp. of this business	3592	0.909	0.288	**-0.010**	0.010
Profit increased during the last 6 months	3579	0.327	0.469	**0.010**	0.017
Business perception index (standardized score)	3933	-0.028	1.005	**0.005**	0.032
Nr. of barriers for business	3975	1.110	0.843	**0.012**	0.027
Panel B: Household Financials and Life Satisfaction					
Log of total savings from all sources in mln IDR^{90p}	1489	1.444	1.891	**-0.072**	0.090
Log of HH exp. on non-durables in the last 30 days in mln IDR^{99p}	3942	0.649	0.935	**-0.009**	0.030
Share of clients reporting increase in HH exp. in the last 6 months	3925	0.459	0.499	**0.014**	0.017
Life satisfaction (1=worst, 10=best)	3877	7.349	1.806	**-0.023**	0.058
Difference between life satisfaction now and two yrs. ago	3852	0.415	1.727	**-0.089**	0.060
Panel C: Loan Behavior					
Current/last loan used for productive purposes	3303	0.717	0.451	**-0.020**	0.016
Currently in loan default or behind with repayments for any loan	2734	0.190	0.392	**-0.018**	0.015

- *Note*: This table shows final outcome variables on the left, statistics on top. Same clients are followed over time.
- Sample: Estimation sample (N = 3,975). Sample size varies between variables due to item nonresponse and conditional questions.
- Source: Endline survey (2018 - 2019).
- Columns (1)-(2) display the control group mean and standard deviation. Columns (3)-(4) present regression results, controlling for age, gender, the twenty most imbalanced baseline covariates, dummies indicating imputation in baseline variables, enumerator and FSP fixed effects. We present robust standard errors.
- The superscript *np.* indicates the winsorizing level.
- The statistical significance is given as follows: * indicates p < 0.1, ** p < 0.05, *** p < 0.01, **** p < 0.001.
- See Section 3-2 in the appendix for robustness checks.

a number of outcomes which proxy the optimism and entrepreneurial attitudes in our sample. We asked about attitudes towards currently running a business in Indonesia, i.e., whether it is perceived as complicated, risky, financially rewarding, earns respect, satisfaction and/or gives a feeling of security. We aggregated the answers into a standardized score, which we call the business perception index. Additionally, we asked about the perceived challenges and barriers to business development and counted the number of mentioned barriers. Again, we observe no significant changes.

Furthermore, we tested a number of outcomes which relate to the clients' personal lives, i.e., total savings, log of and increase in household expenditures, life satisfaction and the change thereof during the past two years. We find no effects on these variables. Finally, we test whether treatment affected loan behavior. The large majority of clients state that they used their last loan for productive purposes and this remains unchanged by the program. Note that the share of clients being currently in loan default or behind in repayments is high in the control group (19.0%). Although the point estimate of the program effect on this variable is negative, it is insignificant.

In summary, we find no statistical evidence that the intervention has transformed the average financial situation of the MSEs or the household, business attitudes, life satisfaction or loan behavior, at least not in the time period considered in this study. This is consistent with there being no average increase in knowledge or change in business practices, except for including cash flows in the business plan.

6 Heterogeneous treatment effects

Although the program did, on average, not effect our outcomes, some subgroups might still have benefited from it. We therefore make use of our relatively large sample size and study heterogeneous treatment effects. Specifically, we analyze differential effects by entrepreneur or business characteristics, by treatment arm, and by type of FSP. In the following, we show results on intermediate outcomes only as these are more likely to have developed within the time period considered here. Results on final outcomes are available in Appendix 3.

6-1 By entrepreneur and business characteristics

Table 3.4 contains heterogeneous effects on intermediate outcomes by entrepreneur and business characteristics calculated using Equation 3.2. We present p-values for the null hypothesis that the interaction effect γ_{IA} is zero (dark grey area). Additionally, we present the p-value for the the null hypothesis that $\gamma_{IA} + \beta_{ITT}$ is zero (light gray areas).

Table 3.4: Effects on intermediate outcomes by entrepreneur and business characteristics

I. Entrepreneur characteristics	**Part** (1)				**Part** (2)				**Part** (3)				**Part** (4)				**Part** (5)			
Group:	Female				Age>45				More than secondary school education				> 5 hours p.d. spent on hh chores (E)				No-one to replace in business (E)			
	YES	NO	Diff.	Total	YES	NO	Diff.	Total	YES	NO	Diff.	Total	YES	NO	Diff.	Total	YES	NO	Diff.	Total
Panel A: Program Participation																				
Share of clients reporting being aware of the support of FSP	0.069	0.057		****	0.046	0.085		**	0.069	0.062		****	0.057	0.071		***	0.076	0.047		****
Share of clients reporting having been offered support from FSP	0.077	0.043		***	0.052	0.091		*	0.030	0.106	*		0.046	0.098			0.100	0.021	*	****
Share of clients reporting having participated in the program	0.152	0.083		****	0.204	0.072	**	****	0.176	0.103		****	0.094	0.181		**	0.115	0.210		***
Panel B: Knowledge																				
Share of marketing knowledge questions answered correctly	-0.013	0.031	**		0.012	-0.009			0.007	-0.003			0.007	0.000			-0.012	0.011		
Share of financial mgmt knowledge questions answered correctly	0.002	-0.003			0.009	-0.007			0.008	-0.007			-0.003	0.005			-0.004	-0.006		
Share of clients not knowing the payment type	-0.021	-0.003			-0.011	-0.018			-0.011	-0.018			-0.022	-0.013			-0.022	-0.007		
Panel C: Busines Practices																				
Marketing practices index$^{n.m.}$	0.085	-0.113	*		0.036	0.009			0.021	0.018			-0.034	0.065			-0.026	0.156	*	
Nr. of different marketing forms used$^{n.m.}$	0.019	0.013			0.012	0.023			0.039	-0.001			0.048	0.006		*	0.030	0.003		
Financial management practices index$^{n.m.}$	-0.000	-0.009			0.082	-0.064			0.032	-0.024			0.011	0.022			0.058	-0.050		
Share of clients keeping business and HH finances separately	-0.031	0.003			-0.012	-0.025			-0.038	-0.005			-0.039	-0.004			-0.015	-0.025		
Share of clients investing profit into the business	-0.002	0.009			0.021	-0.019			0.016	-0.011			0.024	-0.013			-0.006	0.023		
Share of clients preparing a business/financial plan	0.009	0.048			0.049	-0.005	*	**	0.016	0.026			0.032	0.028			0.017	0.046		
Business plan includes cash flow	0.025	0.037			-0.006	0.061	**		0.029	0.029			0.017	0.044			0.041	0.011		***
N	2743.000	1232.000			1960.000	2007.000			1792.000	2180.000			1588.000	2248.000			2215.000	1321.000		

II. Business characteristics	**Part** (6)				**Part** (7)				**Part** (8)				**Part** (9)				**Part** (10)			
Group:	> 9 years in business				Above average profit				Above average loan size				Above average marketing index				Preparing business plan			
	YES	NO	Diff.	Total	YES	NO	Diff.	Total	YES	NO	Diff.	Total	YES	NO	Diff.	Total	YES	NO	Diff.	Total
Panel A: Program Participation																				
Share of clients reporting being aware of the support of FSP	0.038	0.092	**	**	0.078	0.052		****	0.015	0.123	****		0.090	0.032	**	****	0.116	0.033	***	****
Share of clients reporting having been offered support from FSP	0.088	0.054		***	0.043	0.090			0.045	0.087			0.074	0.062		***	0.072	0.070		**
Share of clients reporting having participated in the program	0.153	0.130		****	0.083	0.192	*	*	0.110	0.164		**	0.136	0.154		****	0.097	0.182		**
Panel B: Knowledge																				
Share of marketing knowledge questions answered correctly	0.008	-0.004			0.015	-0.012			0.014	-0.012			0.000	0.004			-0.021	0.016	*	
Share of financial mgmt knowledge questions answered correctly	0.008	-0.008			0.017	-0.017	*		0.002	-0.002			-0.005	0.009			-0.016	0.010		
Share of clients not knowing the payment type	-0.029	-0.000		*	-0.009	-0.020			-0.003	-0.031			-0.010	-0.022			0.007	-0.028		
Panel C: Business Practices																				
Marketing practices index$^{n.m.}$	0.005	0.033			-0.005	0.037			0.039	-0.005			0.019	0.023			-0.080	0.078		
Nr. of different marketing forms used$^{n.m.}$	-0.008	0.043			0.028	0.006			0.047	-0.017	*	**	0.024	0.009			-0.004	0.030		
Financial management practices index$^{n.m.}$	0.013	-0.018			-0.047	0.028			-0.018	0.016			-0.009	0.011			-0.081	0.040		
Share of clients keeping business and HH finances separately	-0.018	-0.020			-0.015	-0.022			-0.020	-0.018			-0.010	-0.031			-0.012	-0.024		
Share of clients investing profit into the business	-0.007	0.011			-0.002	0.006			-0.001	0.005			-0.006	0.013			-0.012	0.011		
Share of clients preparing a business/financial plan	0.039	0.005		*	0.033	0.009			0.015	0.030			0.015	0.034			0.001	0.036		
Business plan includes cash flow	0.026	0.032			0.026	0.034			0.024	0.035			0.032	0.024		**	0.033	0.025		*
N	1981.000	1994.000			1987.000	1988.000			1987.000	1988.000			2350.000	1625.000			1617.000	2358.000		

- *Note*: This table shows the main outcome variables on the left. **Parts** (1)-(10) on the top indicate heterogeneous effects for different groups.
- Sample: Estimation sample (N = 3,975).
- Source: Endline survey (2018 - 2019).
- Dark grey columns show significance level of the null hypothesis which assumes the effect to be equal across the two respective groups. Light grey columns show significance level of the null hypothesis of zero effect in the respective YES-group.
- Part (4): Indicator for more than 5 hours per day spend on household chores (measured only at endline); Part (5): Indicator for respondent claiming that no-one can replace her in the business (measured only at endline); Part (9): The marketing index is the sum of dummy variables indicating whether the client has (i) a customer identification strategy, (ii) a strategy to make customer like their products, (iii) a competition strategy, (iv) any marketing practices in use; Part (10): Indicator for preparing a business plan at baseline.
- In all regressions we control for the twenty most imbalanced baseline covariates and dummies indicating imputation in baseline variables, enumerator- and FSP fixed effects, and, if applicable, for age and gender.
- The superscript *n.m.* indicates that missing values were interpreted as zero to generate index variables.
- The statistical significance is given as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, **** indicates $p < 0.001$.
- Table C.4 in Appendix 3 contains results for final outcomes.

Regarding heterogeneous effects across entrepreneur characteristics, we find no consistent significant differences in effects across gender, educational background or hours spend on household chores. In particular, we cannot confirm findings in previous literature where male participants were more likely to benefit from training (Berge et al. 2015; Gine and Mansuri 2014). The last two indicators (Parts (4) and (5)) are motivated by results from Fiala (2013), which suggest that these are important factors mitigating impact.[24] We find that an entrepreneur who claims that no-one can replace her in the business is significantly more likely to include a cashflow analysis in her business plan, but find no other differential impact beyond this. We also find that the program's effect on intermediate outcomes does not depend on business characteristics, such as age, profits, loan sizes, prior marketing strategies or business planning.[25]

Results on final outcomes are contained in Table C.4 in Appendix 3. Overall, we find no indication that individual or business characteristics affect the program's impact.

6-2 By treatment arm

We estimate effects for each of the three treatment arms separately and test for significant differences between them. Since only the cooperatives offered all three treatment arms, this part of our analysis is restricted to FSPs 1 to 6. We regress the outcomes of interest simultaneously on being assigned to one of the three treatment arms and compare the results to a control group which was assigned to not receive any treatment. Table 3.5 contains the results for intermediate outcomes, where the columns describe the sample size, control mean and standard deviation in cooperatives as well as the three ITT point estimates and standard errors of the T (Columns (3)-(4)), C (Columns (5)-(6)), and TC treatments (Columns (7)-(8)) respectively. The last three columns present the p-value testing the null-hypothesis of equal effects in two treatment arms respectively.

Among intermediate outcomes, there are some outcomes for which we estimate significant changes in one treatment arm but not in other arms. However, apart from the participation variables, none of these results are in the expected direction or robust against multiple hypotheses testing. Importantly, the null hypothesis that any two treatment coefficients are equal cannot confidently be rejected for any outcome. This holds true for final outcomes, see Table C.5 in Appendix 3. Our hypothesis that we would observe the largest effects in the highest-intensity TC treatment is hence not confirmed in our data.

[24] Note that we only measured these variables at endline survey and results might hence suffer from endogeneity.

[25] Looking first at the participation outcomes, we find that businesses with smaller loans and businesses that already prepared a business plan at baseline are more likely to be aware of the program. However, these businesses are mostly clients of cooperatives, which offered three treatment arms. As laid out above, clients who were offered training are more likely to report awareness of the treatment, and this explains the observed significant differences across subgroups.

Table 3.5: Effects on intermediate outcomes by treatment arm (cooperatives only)

Outcome	N	Control Mean (1)	Control SD (2)	Training β^{T}_{ITT} **(3)**	Training SE (4)		Counseling β^{C}_{ITT} **(5)**	Counseling SE (6)		Training&Counseling β^{TC}_{ITT} **(7)**	Training&Counseling SE (8)		P-value for H_0: $\beta^{T}_{ITT}=\beta^{C}_{ITT}$ (9)		P-value for H_0: $\beta^{T}_{ITT}=\beta^{TC}_{ITT}$ (10)		P-value for H_0: $\beta^{C}_{ITT}=\beta^{TC}_{ITT}$ (11)	
Panel A: Program Participation																		
Share of clients reporting being aware of the support that FSP is providing	2493	0.639	0.481	**0.102**	0.025	****	**0.069**	0.025	***	**0.150**	0.024	****	0.155		0.036	**	0.000	****
Share of clients reporting having been offered support from FSP	1787	0.699	0.459	**0.106**	0.026	****	**0.007**	0.028		**0.130**	0.026	****	0.000	****	0.293		0.000	****
Share of clients reporting having participated in the program	1391	0.599	0.491	**0.152**	0.037	****	**0.088**	0.038	**	**0.190**	0.035	****	0.050	**	0.200		0.001	***
Panel B:Knowledge																		
Share of marketing knowledge questions answered correctly	2541	0.389	0.309	**-0.035**	0.017	**	**-0.028**	0.017	*	**-0.019**	0.016		0.664		0.297		0.550	
Share of financial mgmt knowledge questions answered correctly	2546	0.659	0.317	**0.004**	0.016		**0.010**	0.015		**-0.006**	0.016		0.666		0.528		0.276	
Share of clients not knowing the payment type	2000	0.147	0.354	**-0.029**	0.020		**-0.004**	0.021		**-0.015**	0.020		0.206		0.450		0.598	
Panel C: Business Practices																		
Marketing practices index$^{n.m.}$	2550	1.508	1.584	**-0.015**	0.079		**0.022**	0.080		**-0.080**	0.077		0.627		0.377		0.172	
Nr. of different marketing forms used$^{n.m.}$	2550	0.284	0.527	**-0.024**	0.027		**0.017**	0.029		**-0.001**	0.027		0.142		0.376		0.534	
Financial management practices index$^{n.m.}$	2550	2.903	2.375	**-0.054**	0.120		**-0.096**	0.117		**0.085**	0.118		0.717		0.231		0.107	
Share of clients keeping business and HH finances separately	2374	0.468	0.499	**-0.044**	0.028		**-0.022**	0.027		**-0.009**	0.028		0.406		0.193		0.609	
Share of clients investing profit into the business	2360	0.646	0.479	**-0.030**	0.025		**-0.002**	0.025		**0.012**	0.025		0.255		0.087	*	0.559	
Share of clients preparing a business/financial plan	2364	0.420	0.494	**-0.010**	0.026		**-0.021**	0.026		**0.001**	0.026		0.644		0.663		0.368	
Business plan includes cash flow	967	0.093	0.291	**0.022**	0.022		**0.043**	0.024	*	**0.020**	0.023		0.330		0.925		0.321	

- *Note*: This table shows intermediate outcome variables on the left, treatment arms and test statistics on top. Same clients are followed over time.
- Sample: Batch I (cooperatives) clients (N = 2,550).
- Source: Endline survey (2018 - 2019).
- Columns: (1)-(2) control group (N = 626). (3)-(4) training only arm (N = 632). (5)-(6) counseling only arm (N = 651). (7)-(8) training and counseling arm (N = 641). Columns (9) to (11) show the p-values of testing equality of β_{ITT} for two treatment arms respectively.
- In all regressions we control for age, gender, twenty most imbalanced baseline covariates, dummies indicating imputation in baseline variables, enumerator- and FSP fixed effects, and present robust standard errors.
- The superscript *n.m.* indicates that missing values were interpreted as zero to generate index variables.
- The statistical significance is given as follows: * indicates $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$.
- Table C.5 in Appendix 3 contains results for final outcomes.

6-3 By FSP

The intervention was implemented by a diverse set of partner FSPs, namely six cooperatives, five rural banks and one development bank and we here analyze how effects differ across these. As before mentioned, cooperatives offered three treatment arms, whereas the banks offered only counseling. For better comparability, we restrict attention to clients assigned to the counseling treatment or control group within cooperatives as well. Table 3.6 contains the results for intermediate outcomes, where the columns describe the sample size, control mean and standard deviation, as well as three ITT point estimates and standard errors for the counseling treatment in cooperatives (Columns (3)-(4)), rural banks (Columns (5)-(6)), and the development bank (Columns (7)-(8)) respectively.

Whereas we find some weakly significant results in all subsamples, there are no highly significant estimates among cooperatives and the development bank and the direction of observed coefficients here also do not paint a clear picture. In contrast, the results restricted to the clients of rural banks give rise to some more optimism. Except for financial knowledge, the coefficient estimates for the main aggregate indices, i.e., marketing knowledge, marketing practices and financial management practices, are all positive. The sizable increase in the number of different marketing forms used is highly significant. Given the consistency of the direction, albeit not magnitude, of the impact on various outcomes we conclude the the intervention had a positive, yet small impact on the knowledge and practices for rural bank clients. Among final outcomes, which we report in Table C.6 in Appendix 3, there are no significant changes, and also the direction of the insignificant effects does not allow us to presume positive effects of the intervention on the participants' financial lives or attitudes.

To further analyze a potential impact of the intervention among rural banks, we go further into detail and run a regression model similar to equation 3.2, but interacting the treatment indicator on the FSP identifier. We present results in Table 3.7, again restricting attention to the counseling treatment arm only, with the columns containing the point estimates of the interaction of treatment and FSP indicators. For better overview, we do not show standard errors but indicate statistical significance with stars as usual. Most notably, the positive results for rural banks are mainly driven by FSP 8. Here, marketing knowledge improved decisively, with the share of questions answered correctly increasing by 9.1 percentage points. The share of clients who prepare a business/financial plan increased by 20.9 percentage points and this increase is significant with a p-value below 0.001 and therefore robust against multiple hypotheses testing. The point estimates for the changes in the marketing practices index and financial management practice index are also positive and large, but statistically insignificant. Even though intermediate outcomes improved in FSP 8, Table C.7 in Appendix 3 shows that final outcomes remain unaffected even in this FSP.

Table 3.6: Effects on intermediate outcomes by type of FSP (counseling only treatment)

Outcome	Control (All)		Cooperatives (C only)				Rural Banks				Dev. Banks		
	Mean	SD	N	$\beta_{ITT}^{Coop.}$	SE		N	β_{ITT}^{Rural}	SE		N	$\beta_{ITT}^{Dev.}$	SE
	(1)	(2)		**(3)**	(4)			**(5)**	(6)			**(7)**	(8)
Panel A: Program Participation													
Share of clients reporting being aware of the support that FSP is providing	0.639	0.481	1246	**0.073**	0.026	***	1112	**0.011**	0.018		290	**0.041**	0.045
Share of clients reporting having been offered support from FSP	0.699	0.459	835	**0.009**	0.029		109	**-0.053**	0.154		49	**-0.220**	0.364
Share of clients reporting having participated in the program	0.599	0.491	594	**0.098**	0.041	**	64	**0.075**	0.164		30	**-0.678**	
Panel B: Knowledge													
Share of marketing knowledge questions answered correctly	0.389	0.309	1272	**-0.028**	0.018		1124	**0.032**	0.018	*	295	**0.046**	0.037
Share of financial management knowledge questions answered correctly	0.659	0.317	1276	**0.015**	0.016		1126	**-0.003**	0.016		294	**0.036**	0.034
Share of clients not knowing the payment type	0.147	0.354	996	**-0.001**	0.022		1025	**-0.008**	0.014		278	**-0.017**	0.036
Panel C: Business Practices													
Marketing practices index$^{n.m.}$	1.508	1.584	1277	**0.051**	0.083		1129	**0.125**	0.081		296	**-0.288**	0.193
Nr. of different marketing forms used$^{n.m.}$	0.284	0.527	1277	**0.019**	0.030		1129	**0.076**	0.026	***	296	**-0.103**	0.079
Financial management practices index$^{n.m.}$	2.903	2.375	1277	**-0.095**	0.121		1129	**0.016**	0.121		296	**0.149**	0.255
Share of clients keeping business and HH finances separately	0.468	0.499	1189	**-0.024**	0.028		951	**-0.009**	0.032		286	**-0.050**	0.057
Share of clients investing profit into the business	0.646	0.479	1186	**0.008**	0.026		954	**0.022**	0.027		285	**-0.030**	0.054
Share of clients preparing a business/financial plan	0.420	0.494	1185	**-0.016**	0.027		958	**0.071**	0.030	**	287	**0.052**	0.055
Business plan includes cash flow	0.093	0.291	476	**0.045**	0.026	*	445	**0.016**	0.024		131	**0.055**	0.037

- *Note*: This table shows intermediate outcome variables on the left, the different samples on top. The same clients are followed over time.
- Sample: Estimation sample, counseling only treatment and control groups (N=2,692).
- Source: Endline survey (2018 - 2019).
- Columns (1)-(2) display the control group mean and standard deviation. In Columns (3)-(8) we present regression results. In all regressions we control for age, gender, the twenty most imbalanced baseline covariates, dummies indicating imputation in baseline variables, enumerator and FSP fixed effects, and present robust standard errors. In Columns (3)-(4) we present results for cooperatives, considering only the counseling treatment and control group. in Columns (5)-(6) we present results for rural banks, in Columns (7)-(8) for development banks.
- The superscript *n.m.* indicates that missing values were interpreted as zero to generate index variables.
- The statistical significance is given as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, **** indicates $p < 0.001$.
- Table C.6 in Appendix 3 contains results for final outcomes.

Table 3.7: Effects on intermediate outcomes by FSP (counseling only treatment)

Outcome	FSP 1 β_{ITT}	FSP 2 γ^{2}_{ia}	FSP 3 γ^{3}_{ia}	FSP 4 γ^{4}_{ia}	FSP 5 γ^{5}_{ia}	FSP 6 γ^{6}_{ia}	FSP 7 γ^{7}_{ia}	FSP 8 γ^{8}_{ia}	FSP 9 γ^{9}_{ia}	FSP 10 γ^{10}_{ia}	FSP 11 γ^{11}_{ia}	FSP 12 γ^{12}_{ia}
Panel A: Program Participation												
Share of clients reporting being aware of the support that FSP is providing	-0.053 ***	0.216 ***	0.092 **	0.161 ***	0.143 ***	0.012	0.033	0.030	0.169 ****	0.073	0.045	0.080
Share of clients reporting having been offered support from FSP	-0.104 ****	0.134 *	0.141 ***	0.097 *	0.070	0.078 *	0.267	0.100	-0.145	0.187	-0.207	-0.247
Share of clients reporting having participated in the program	-0.085 ***	0.248 ***	0.106	0.188 ***	0.037	0.256 ****	0.185	-0.102	0.184	0.258	-0.629 ****	0.448
Panel B: Knowledge												
Share of marketing knowledge questions answered correctly	-0.003	-0.045	-0.033	-0.031	-0.027	0.011	0.033	0.091 ***	-0.022	0.062 *	0.017	0.024
Share of financial management knowledge questions answered correctly	0.011	0.003	-0.016	-0.002	-0.008	0.010	-0.021	0.003	0.024	0.013	-0.057	-0.034
Share of clients not knowing the payment type	0.021	-0.006	-0.044	-0.045	-0.015	0.003	-0.019	-0.032	-0.051	-0.041	-0.058 *	-0.022
Panel C: Business Practices												
Marketing practices index$^{n.m.}$	0.062	-0.176	0.014	-0.081	-0.018	-0.058	0.171	0.215	-0.063	-0.344 *	-0.364 **	0.408 *
Nr. of different marketing forms used$^{n.m.}$	0.025	-0.038	-0.057 **	0.002	-0.008	0.064	0.120	0.056	0.123 *	-0.135 *	-0.043	0.006
Financial management practices index$^{n.m.}$	-0.112	-0.085	-0.034	0.027	0.118	-0.059	0.317	0.299	0.165	0.180	-0.084	-0.014
Share of clients keeping business and HH finances separately	0.007	-0.027	0.021	-0.014	-0.075	-0.095 *	0.073	-0.058	0.087	-0.058	-0.075	-0.076
Share of clients investing profit into the business	-0.001	0.080	0.045	-0.083	-0.030	0.016	-0.019	0.086	0.073	-0.041	-0.062	0.058
Share of clients preparing a business/financial plan	-0.019	-0.031	0.030	0.038	0.006	-0.029	0.076	0.208 ****	0.108	0.060	-0.059	-0.036
Business plan includes cash flow	0.024	0.007	0.078	0.007	0.015	0.043	-0.022	-0.027	-0.006	0.034	0.033	0.002
N	128	122	288	212	273	244	228	314	220	296	195	172

- *Note*: This table shows intermediate outcome variables on the left, the different FSPs on top. The same clients are followed over time.
- Sample: Estimation sample, counseling only treatment and control groups (N=2,692).
- Source: Endline survey (2018 - 2019).
- The columns contain the point estimate of the interaction of the treatment indicator with the FSP identifier for the twelve FSPs. In all regressions we control for age, gender, twenty most imbalanced baseline covariates, dummies indicating imputation in baseline variables, enumerator and FSP fixed effects, and present robust standard errors.
- The superscript *n.m.* indicates that missing values were interpreted as zero to generate index variables.
- The statistical significance is given as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, **** indicates $p < 0.001$.
- Table C.7 in Appendix 3 contains results for final outcomes.

6-4 Discussion

We find no indication that entrepreneur or business characteristics influence the program's impact, nor does treatment intensity (training, counseling or both) seem to make a difference on average. The only notable factor we find driving result is the respective FSP. This is an important result as organizations aiming to upscale their programs will likely need to rely on a variety of implementing partners. In principle, differential impacts across partner institutions can be due to (1) the treatment reception, or (2) its delivery. That is, either the institutions serve different clients, or the institutions implemented the treatment differently.

Regarding the first possible explanation, the largest differences in the client base likely exist between the three types of institutions as groups. For example, cooperatives have a larger share of female clients, and the development bank serves larger MSEs, see Table C.2 in Appendix 2. Our analysis of heterogeneous effects however suggests that impact does not depend on observable characteristics to any measurable extent. It is however possible that unobservable characteristics remain, such as the willingness to learn and innovate, and that rural banks are, on average, more successful than credit cooperatives in selecting entrepreneurs along unobservable dimensions which predict impact heterogeneity. This could for example be because rural banks employ more experienced loan officers or have more thorough lending procedures in place.

Regarding the second possible explanation, differential treatment implementation, three explanatory factors might play a role. First, the different types of FSPs were exposed to the treatment for different periods of time and this phase-in was not random. To be precise, banks offered the services at a later point of time than cooperatives, so we might detect short-term effects here which are no longer present at cooperatives. Second, loan officers at banks might be higher qualified than at cooperatives and hence better suited as trainers/counselors. This seems likely as salaries of loan officers in rural banks exceed those of loan officers in credit cooperatives by an approximate factor 1.5. Third, the intensity of treatment implementation varies across FSPs, as shown in Section 3-4. Whereas the first two arguments can plausibly explain differences between the three types of FSPs, only the third is able to explain the differences between individual rural banks. The consistent and relatively large effects we estimate for FSP 8 are in line with the fact that this FSP showed the highest participation rate in the monitoring system among rural banks, see Figure 3.4. This suggests that this institution showed a higher implementation fidelity.

7 Conclusion

Using an RCT with 3,975 microfinance clients, we evaluate a pilot training and counseling intervention, which was initiated by the ILO as one intervention within the broader PROMISE-IMPACT program in Indonesia. The intervention built the capacities of loan officers of twelve participating FSPs to provide classroom training and individual counseling to their MSE clients. The appeal of this approach is apparent: Classroom training is cost-efficient and easily replicated, and even individual counseling can be cost-efficient if delivered through loan officers of FSPs that themselves benefit from improved loan behavior of their clients. On the downside, the intervention is of relatively low intensity and the cooperation with twelve FSPs implies a longer impact chain and hence a loss of control compared to direct training delivery.

Indeed, we find no significant effects of the training and/or counseling intervention on final outcomes averaged over all FSPs, such as profits, household spending or business attitudes. Among intermediate outcomes, i.e., knowledge and business practices, the only robustly significant and positive outcome is a 2.9 percentage point increase in the share of clients whose business plan includes a cashflow analysis. The control group mean for this variable is 7.3%, suggesting that the effect is sizable in magnitude. Matching loan repayment schedules to actual business cashflow is important to leverage the whole potential of access to finance for both, lenders and borrowers. Hence, loan officers in their training/counseling might either have focused on what they knew best or what was most beneficial to their FSP.

Exploiting our large sample size, we take a closer look at a variety of potential heterogeneous effects. We find that entrepreneur or business characteristics have little or no influence on the intervention's impact. Notably, we cannot replicate the effect that men benefit more than women, which similar studies have found. Furthermore, a comparison of the treatment effects across the three different treatment arms offered by the participating cooperatives also does not support the hypothesis that the largest treatment intensity, i.e. classroom training with subsequent individual counseling visits, achieved greatest impact.

A novel finding of our study is, however, that effects vary across the partner institutions. Selecting these is the first step in any intervention implementation, yet its importance has been little discussed in the literature. We exploit the fact that ILO works with twelve independent FSPs spanning three different institutional set-ups and find consistent, albeit small, improvements in knowledge and practice outcomes among clients from rural banks. This could be due to heterogeneous effects on client characteristics which are unobservable to us if rural banks are more successful in selecting high potential entrepreneurs as their clients. It could also be due to higher quality of training if loan officers in rural banks are better educated and/or more experienced. Within rural banks, one FSP achieved particularly large impact on knowledge and practice outcomes. In fact, this FSP also showed the highest implementation fidelity among rural banks.

Whereas we conclude that the program was not successful in helping entrepreneurs grow their businesses within the considered timeframe, our research has highlighted the need to carefully choose implementing partners in multi-stakeholder settings. Especially when scaling up training interventions while maintaining cost-efficiency, the need to cooperate with various partners is inevitable. Further research should explore in more detail the institutional characteristics of successful intervention partners.

Bibliography

Abadie, A., & Imbens, G. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, *76*(6), 1537–1557.

ADB. (2019). *Khyber Pakhtunkhwa Health Sector Review - Hospital Care* (Tech. Rep.). Asian Development Bank.

Aldianto, L., Rudito, B., Mirzanti, I., Situmorang, B., & Larso, D. (2010). The development of center of entrepreneurship and business incubator in pangalengan, west java - indonesia. *Technology Management for Global Economic Growth (PICMET), 2010 Proceedings of PICMET*.

Ali, N. (2010). Sectarian imaginaries. the micropolitics of sectarianism and state-making in nothern pakistan. *Current Sociology*, *58*(5), 738-754.

Allcott, H. (2015). Site selection bias in program evaluation. *The Quarterly Journal of Economics*, *130*(3), 1117–1165.

Asif, A. F. (2017). Healthcare challenges in Gilgit Baltistan: The way forward. *Pakistan Journal of Public Health*, *7*(2), 113-116.

Asuming, P. O. (2013). Getting the poor to enroll in health insurance, and its effects on their health: Evidence from a field experiment in Ghana. *Job Market Paper–Columbia University*.

Atmadja, A. S., Su, J.-J., & Sharma, P. (2016). Examining the impact of microfinance on microenterprise performance (implications for women-owned microenterprises in indonesia). *International Journal of Social Economics*, *43*(10), 962-981.

Ayyagari, M., Beck, T., & Demirguc-Kunt, A. (2007). Small and medium enterprises across the globe. *Small Business Economics*, *29*(4), 415–434.

Banerjee, A., Duflo, E., Glennerster, R., & Kinnan, C. (2015). The miracle of microfinance? evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, *7*(1), 22–53.

Banerjee, A., Finkelstein, A., Hanna, R., Olken, B. A., Ornaghi, A., & Sumarto, S. (2019). *The challenges of universal health insurance in developing countries: Evidence from a large-scale randomized experiment in Indonesia* (Tech. Rep.). National Bureau of Economic Research.

Benz, A. (2016). Framing modernization interventions:reassessing the role of migration and translocality in sustainable mountain development in gilgit-baltistan, pakistan. *Mountain Research and Development*, *36*, 141-152. doi: 10.1659/MRD-JOURNAL-D-15-00055.1

Berge, L., Bjorvatn, K., & Tungodden, B. (2015). Human and financial capital fro microenterprise development: Evidence from a field and lab experiment. *Management Science*, *61*(4), 707-722.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., & Robert, J. (2013). Does management matter? evidence from india. *Quarterly Journal of Economics*, *128*(1).

Bruhn, M., Karlan, D., & Schoar, A. (2010). What capital is missing in developing countries? *American Economic Review*, *100*(2), 629–33.

Bruhn, M., Karlan, D. S., & Schoar, A. (2013). The impact of consulting services on small and medium enterprises: Evidence from a randomized trial in mexico. *World Bank Policy Research Working Paper*.

Calderon, G., Cunha, J., & De Giorgi, G. (2013). Business literacy and developmetn: Evidence from a randomized controlled trial in rural mexico. *Working Papers, Banco de Mexico*(2013-24).

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). rdrobust: Software for regression-discontinuity designs. *The Stata Journal*, *17*(2), 372–404.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, *82*(6), 2295–2326.

Cole, S., Sampson, T., & Zia, B. (2009). Financial literacy, financial decisions, and the demand for financial services: Evidence from india and indonesia. *Harvard Business School Working Paper*.

Contech International Health Consultants. (2012). *Health Facility Assessment Provincial Report* (Tech. Rep.). Technical Resource Facility.

De Mel, S., McKenzie, D., & Woodruff, C. (2008). Returns to capital in microenterprises: Evidence from a field experiment. *The Quarterly Journal of Economics*, *123*(4), 1329-1372.

De Mel, S., McKenzie, D., & Woodruff, C. (2014). Business training and female enterprise start-up, growth, and dynamics: Experimental evidence from sri lanka. *Journal of Development Economics*, *106*, 199–210.

Drexler, A., Fischer, G., & Schoar, A. (2010). *Keeping it simple: Financial literacy and rules of thumb.*

Fiala, N. (2013). *Stimulating microenterprise growth: Results from a loans, grants and training experiment in uganda.* doi: http://dx.doi.org/10.2139/ssrn.2358086

Field, E., Jayachandran, S., & Pande, R. (2010). Do traditional institutions constrain female entrepreneurship? a field experiment on business training in india. *The American Economic Review*, *100*(2), 125-129.

Finkelstein, A. (2007). The aggregate effects of health insurance: Evidence from the introduction of medicare. *The Quarterly Journal of Economics*, *122*(1), 1–37.

Frölich, M., & Landmann, A. (2018). Effects of insurance on child labour: Ex-ante and ex-post behavioural changes. *The Journal of Development Studies*, *54*(6), 1002–1018.

Frölich, M., & Huber, M. (2019). Including covariates in the regression discontinuity design. *Journal of Business & Economic Statistics*, *37*(4), 736–748.

Gine, X., & Mansuri, G. (2014). Money or ideas? a field experiment on constraints to entrepreneurship in rural pakistan. *The World Bank Policy Research Working Paper*(6959).

Government of Pakistan. (2016). *Pakistan Social and Living Standards Measurement Survey 2014/ 2015* (Tech. Rep.). Statistics Devision, Pakistan Bureau of Statistics.

Gugerty, M. K., & Karlan, D. (2018). Ten reasons not to measure impact—and what to do instead. *Stanf. Soc. Innov. Rev.*

Habib, S. S., Perveen, S., & Khuwaja, H. M. A. (2016). The role of micro health insurance in providing financial risk protection in developing countries a systematic review. *BMC Public Health*, *16*(1), 281.

Helmsmueller, S., & Landmann, A. (2020). *Does free hospitalization insurance change health care consumption of the poor? short-term evidence from pakistan.*

Holden, L. (2019). Law, governance, and culture in gilgit-baltistan: introduction. *South Asian History and Culture*, *10*(1), 1–13.

Hunzai, I. (2013). *Conflict dynamics in gilgit-baltistan.* United States Institute of Peace Washington, DC.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

International Labour Organization. (2013). *Start and improve your business programme.* (http://www.ilo.org/empent/areas/start-and-improve-your-business/lang–en/index.htm ; accessed on 2018.02.04)

International Labour Organization. (2018). *Start and improve your business programme.* (http://www.ilo.org/jakarta/whatwedo/projects/WCMS_444105/lang–en/index.htm ; accessed on 2018.02.04)

Jalal, S., & Haq, I. (2014). Revisiting the three different tiers of the health system of pakistan and their implications for the achievement of mdgs by pakistan. *J Pak Med Assoc*, *64*(2), 195–200.

Karlan, D., Knight, R., & Udry, C. (2015). Consulting and capital experiments with microenterprise tailors in ghana. *Journal of Economic Behavior and Organization*, *118*, 281-302.

Karlan, D., & Valdivia, M. (2011). Teaching entrepreneurship: impact of business training on microfinance clients and institutions. *The Review of Economics and Statistics*, *93*(2), 510-527.

Khan, S. A. (2016). *Comparison of Federal and Khyber Pakhtunkhwa health insurance schemes, qualitative analysis of their proximity to universal health coverage* (Unpublished master's thesis). Khyber Medical University.

Lafortune, J., Riutort, J., & Tessada, J. (2017). Are micro-entrepreneurs constrained by their lack of knowledge or motivation? lessons from a randomized experiment in chile. (working paper)

Lagomarsino, G., Garabrant, A., Adyas, A., Muga, R., & Otoo, N. (2012). Moving towards universal health coverage: health insurance reforms in nine developing countries in africa and asia. *The Lancet*, *380*(9845), 933–943.

Landmann, A., & Frölich, M. (2015). Can health-insurance help prevent child labor? An

impact evaluation from Pakistan. *Journal of Health Economics*, *39*, 51–59.
Levine, D., Polimeni, R., & Ramage, I. (2016). Insuring health or insuring wealth? an experimental evaluation of health insurance in rural cambodia. *Journal of Development Economics*, *119*, 1–15.
Mac Ginty, R. (2014). Everyday peace: Bottom-up and local agency in conflict-affected societies. *Security Dialogue*, *45*(6), 548–564.
Mangham, L. J., & Hanson, K. (2010). Scaling up in international health: what are the key issues? *Health policy and planning*, *25*(2), 85–96.
McCrary, J. (2008). Manipulation of the running variable in the regression disconinuity design: A density test. *Journal of Econometrics*, *142*, 698-714.
McKenzie, D., & Woodruff, C. (2013). What are we learning from business training and entrepreneurship evaluations around the developing world? *The World Bank Research Observer*, *29*(1), 48–82.
McKenzie, D., & Woodruff, C. (2015). *Business practices in small firms in developing countries.* The World Bank.
MICS. (2017). *Gilgit Baltistan Multiple Indicator Cluster Survey 2016-2017, Final Report* (Tech. Rep.). United Nationals Children Fund and Planning & Development Department, Government of Gilgit-Baltistan.
Miller, K. J. L. (2015). *A spiritual development: Islam, volunteerism and international development in the hunza valley, northern pakistan* (Unpublished doctoral dissertation). UC San Diego.
Nishtar, S., Boerma, T., Amjad, S., Alam, A. Y., Khalid, F., ul Haq, I., & Mirza, Y. A. (2013). Pakistan's health system: Performance and prospects after the 18th Constitutional Amendment. *The Lancet*, *381*(9884), 2193–2206.
OECD. (2018). *Sme and entrepreneurship policy in indonesia 2018.* OECD Publishing. (http://read.oecdilibrary.org/employment/smeandentrepreneurshippolicyin-indonesia2018_9789264306264en; accessed on 2019.05.31)
OLeary, S., Cheema, I., Hunt, S., Carraro, L., & Pellerano, L. (2011). *Benazir Income Support Programme impact evaluation: Baseline survey report.* Oxford Policy Management Oxford.
Oxford Policy Management. (2016). *Social Health Protection Initiative, Government of Khyber Pakhtunkwa - inception report* (Tech. Rep.). Oxford Policy Management.
Oxford Policy Management. (2017). *Social Health Protection Health Insurace Programme - consulting services for programme implementation, Pakistan - progress report (january to june 2017)* (Tech. Rep.). Oxford Policy Management.
Prinja, S., Chauhan, A. S., Karan, A., Kaur, G., & Kumar, R. (2017). Impact of publicly financed health insurance schemes on healthcare utilization and financial risk protection in India: A systematic review. *PLoS ONE*, *12*(2). doi: 10.1371/journal.pone.0170996
Reich, M. R., Harris, J., Ikegami, N., Maeda, A., Cashin, C., Araujo, E. C., … Evans, T. G. (2016). Moving towards universal health coverage: lessons from 11 country studies. *The Lancet*, *387*(10020), 811–816.
Rosengard, J., & Prasetyantoko, A. (2011). If the banks are doing so well, why can't i get a

loan? regulatory constraints to financial inclusion in indonesia. *Asian Economic Policy Review*, *6*(2), 273–296.

Shabbir, A., & Malik, S. A. (2016). Measuring patients' healthcare service quality perceptions, satsifaction, and loyalty in public and private sector hospitals in pakistan. *International Journal of Qualtiy and Relability Management*, *33*(5), 538-557.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, *125*(1-2), 305–353.

Sökefeld, M. (2015). *Spaces of conflict in everyday life: Perspectives across asia.* transcript Verlag.

Statistics Indonesia. (2018). *Percentage of poor people 2007 to 2018.* (accessed on 2018.07.28)

Swain, R. B., & Varghese, A. (2013). Delivery mechanisms and impact of microfinance training in indian self-help groups. *Journal of International Development*, *25*, 11-21.

Thornton, R. L., Hatt, L. E., Field, E. M., Islam, M., Solis Diaz, F., & Gonzalez, M. A. (2010). Social security health insurance for the informal sector in nicaragua: a randomized evaluation. *Health economics*, *19*(S1), 181–206.

Todd, P. E. (2010). Matching estimators. In *Microeconometrics* (pp. 108–121).

TRC. (2012). *Health facility assessment, Khyber Pakhtunkhwa provincial report* (Tech. Rep.).

Uddin, B. R., Rafi, S., Uddin, H., & Khurshid, A. (2013). *Targeting survey spot check* (Tech. Rep.). Benazir Income Support Programme. (Innovative Development Strategies)

UNPO. (2017). *Member profil Gilgit Baltistan* (Tech. Rep.). The Unrepresented Nations and Peoples Organization.

Valdivia, M. (2015). Business training plus for female entrepreneurship? short and medium-term experimental evidence from peru. *Journal of Development Economics*, *113*, 33-51.

Varley, E. (2010). Targeted doctors, missing patients: obstetric health services and sectarian conflict in northern pakistan. *Social science & medicine*, *70*(1), 61–70.

Varley, E. (2015). Exclusionary infrastructures: Crisis and the rise of sectarian hospitals in northern pakistan. In M. Sökefeld (Ed.), *Spaces of conflict in everyday life: Perspectives across asia* (pp. 187–210). transcript Verlag.

Vidyattama, Y., Miranti, R., & Resosudarmo, B. P. (2014). The role of health insurance membership in health service utilisation in Indonesia. *Bulletin of Indonesian Economic Studies*, *50*(3), 393–413. doi: 10.1080/00074918.2014.980380

Wagstaff, A. (2002). Poverty and health sector inequalities. *Bulletin of the World Health Organization*, *80*, 97–105.

Wagstaff, A., Lindelow, M., Jun, G., Ling, X., & Juncheng, Q. (2009). Extending health insurance to the rural population: An impact evaluation of China's new cooperative medical scheme. *Journal of Health Economics*, *28*(1), 1–19.

WHO. (2013). *Pakistan: Health system profile* (Tech. Rep.). World Health Organization Regional Office for the Eastern Mediterranean.

WHO. (2017). *Annual report 2017* (Tech. Rep.). World Health Organization. (http://www.healthkp.gov.pk/wp-content/uploads/2018/03/annualreport2017.pdf)

Appendix A

Appendix to Chapter 1

1 Additional information on the program and survey

1-1 Features of the two program phases

Table A.1: Features of the two program phases

Salient features	**SHPI Phase-1**	**SHPI Phase-2**
Area	**Four (4) districts**	**Entire province, i.e., twenty-six (26) districts**
Total funding	PKR 1.4 billion	PKR 5.4 Billion
Source of funding	Kreditanstalt für Wiederaufbau (KfW) + Government of KP	Government of KP
Funding (PKR)	Total cost: PKR 1.4 billion, KfW share; PKR 1,233 million (88%) KP share; PKR 1,66 million (12%)	Total cost: PKR 5.4 billion, all through Government's general revenue
Premium	PKR 1,661/- household	PKR 1,549/- household
Project launch	**15th December 2015**	**31st August 2016**
Duration of project	ADP scheme for 5 years	ADP schemes for 2 years
Who is covered		
Percentage of population	**21% poorest population of target districts**	**51% poorest population of entire KP**
Enrolment criteria	**Families with poverty score of 16.17 or less**	**Families with poverty score of 26.75 or less**[1]
Family size	**7 persons per household**	**8 persons per household**

continued

[1] Officially, the cut-off score is 24.51. However, the administrative data reveals that households up to 26.75 were enrolled in our surveyed districts.

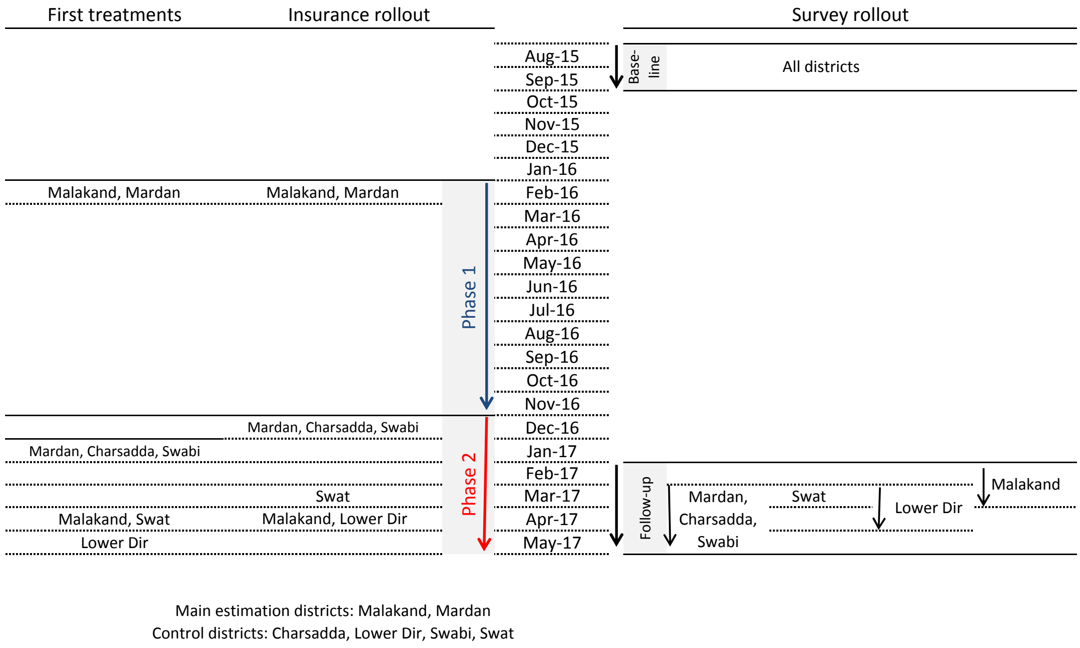
Table A.1: Features of the two program phases

Salient features	**SHPI Phase-1**	**SHPI Phase-2**
Total population covered	0.1 Million households, comprised of 7 members, hence 0.7 million people are covered	1.8 million households, comprised of 8 members, hence 14.4 million people are covered
What is covered		
Type of services	**Inpatient services only**	**Predominantly inpatient services**
Outpatient cover	Maternity services only	Maternity services and cancer care
Secondary diseases	Almost all needing admission	All needing admission
Tertiary conditions	None	Yes, limited tertiary cover
What amount of expenditure is covered		
Mode of payment	Beneficiary-provider interaction is cashless. The insurer pays the providers.	Beneficiary-provider interaction is cashless. The insurer pays the providers.
Premium (Government paid)	PKR 1,661/- household	PKR 1,549/- household, it includes PKR 50 for stop loss coverage
Upper limit (secondary care)	**PKR 25,000/- per person per year, PKR 175,000/- per household per year**	**PKR 30,000/- per person per year, PKR 240,000/- per household per year**
Upper limit (tertiary care)	None	PKR 3,000/- per household per year
Wage replacement	None	PKR 250/- per day for 3 days
Tertiary transportation	None	PKR 2,000/- after discharge from a tertiary care
Maternity transportation	None	PKR 1,000/- post-delivery at hospital
Burial allowance	None	PKR 10,000/- at death of insured household member, during a hospital admission
OPD voucher	None	One OPD visit after discharge from hospital

- *Note*: This table illustrates the features of the two program phases. In bold those features which are most relevant to our study.
- Source: Khan (2016)

1-2 Timeline of program and survey

Figure A.1: Timeline of program implementation and survey



► *Note*: This figure illustrates the months of baseline and endline survey, enrollment and first reported card usage in hospitals for each of the six districts considered here.

► We estimate effects for households in the districts Malakand and Mardan. During the time of our study, Malakand saw roll-out of Phase 1 only, whereas Phase 2 was initiated three months prior to the endline survey in Mardan.

► For robustness checks, we use the districts Charsadda, Lower Dir, Swabi, and Swat. Om Charsadda and Swat, Phase 2 started three months prior to the endline survey. In Swat, enrollment for Phase 2 was taking place in parallel to our follow-up survey, but hospitals had not started offering treatment yet. In Lower Dir, enrollment for Phase 2 started after the endline survey.

1-3 Baseline characteristics for random sub-sample

Table A.2: Baseline characteristics of random full sample and subsample of eligible population (selected variables)

	Full sample				Eligible sample			
	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Household-level variables								
Insurance status at endline	0.38	0.49	0.00	1.00	0.66	0.48	0.00	1.00
Poverty score	24.31	12.56	0.00	79.00	11.17	3.41	0.00	16.17
P.c. monthly income (sqr. root equiv.)	607.71	901.10	0.00	7500.00	373.72	535.92	0.00	4375.00
Wealth index	-0.08	2.02	-3.40	12.23	-0.64	1.54	-3.21	5.79
HH size	7.12	2.74	1.00	23.00	8.02	2.46	3.00	21.00
Electricity in HH	0.97	0.18	0.00	1.00	0.94	0.23	0.00	1.00
Tab water supply in residence	0.11	0.32	0.00	1.00	0.11	0.31	0.00	1.00
Private flush toilet	0.38	0.49	0.00	1.00	0.26	0.44	0.00	1.00
Reported dist. to next hosp. (minutes, win99)	43.76	26.21	0.00	150.00	45.66	27.51	0.00	150.00
Use of prof. assist. during childbirth	0.90	0.31	0.00	1.00	0.86	0.35	0.00	1.00
Case of neglected health care	0.15	0.35	0.00	1.00	0.18	0.38	0.00	1.00
Case of outpatient care	0.78	0.42	0.00	1.00	0.79	0.41	0.00	1.00
Citing health shock as a risk	0.94	0.23	0.00	1.00	0.95	0.22	0.00	1.00
Dif'ty finding money for health care ¿= 8 (scale 1/10)	0.47	0.50	0.00	1.00	0.51	0.50	0.00	1.00
Having heard of insurance	0.02	0.15	0.00	1.00	0.01	0.12	0.00	1.00
Observations	1,146				566			
Panel B: Member-level variables								
Age (win99)	23.97	18.70	1.00	90.00	21.99	17.41	1.00	90.00
School-aged (6 to 16)	0.30	0.46	0.00	1.00	0.40	0.49	0.00	1.00
Female	0.48	0.50	0.00	1.00	0.49	0.50	0.00	1.00
Prim. school not comp'd.	0.57	0.50	0.00	1.00	0.66	0.47	0.00	1.00
Comp'd sec. educ. or higher	0.13	0.34	0.00	1.00	0.07	0.26	0.00	1.00
Worked for salary in previous month	0.20	0.40	0.00	1.00	0.18	0.39	0.00	1.00
Usage of inpatient care	0.05	0.22	0.00	1.00	0.04	0.20	0.00	1.00
Cost of last treatment (PKR, win99)	24,613	42,664	0	300,000	25,587	47,216	500	300,000
More than one admittance to hospital	0.23	0.42	0.00	1.00	0.24	0.43	0.00	1.00
Use of private hospital	0.28	0.45	0.00	1.00	0.27	0.45	0.00	1.00
Observations	7,687				4,238			

- *Note*: This table shows the baseline characteristics of the random full sample and the eligible subsample (households (members) with a poverty score below 16.17). Selected variables on the left, statistics on top.
- Samples: Households and their members in full or eligible sample (varying N). To obtain representative statistics, the full sample contains only randomly selected households, i.e., excluding oversampling below and around cut-off. Eligible sample contains randomly selected households below cut-off, i.e., excluding oversampling around cut-off.
- Source: Baseline survey (2015), insurance status from endline (2017).
- Column (1) displays the mean for continuous/shares for binary variables in the full sample, Column (2) the standard deviation, Columns (3) the minimal and (4) the maximal value in the full sample. Columns (5) to (8) display the same statistics for the subsample of eligible households and their members.
- The suffix *win99* indicates that we winsorized the variable at the 99th percentile level. Monetary variables in PKR (100 PKR = 0.953 USD on December 31, 2015).

1-4 Non-childbirth hospitalization

Table A.3: Logit regression of non-childbirth related hospitalization on individual and household characteristics, baseline

Admission to inpatient care	Logit 1 Coef. (1)	Logit 1 Std. Error (2)	Logit 1 P-value (3)	Logit 2 Coef. (4)	Logit 2 Std. Error (5)	Logit 2 P-value (6)	Logit 3 Coef. (7)	Logit 3 Std. Error (8)	Logit 3 P-value (9)
Poverty score	-0.001	0.006	0.862						
pcear				-0.008	0.006	0.159			
Wealth index							-0.113	0.028	0.000
Female	0.071	0.087	0.416	0.069	0.087	0.429	0.062	0.086	0.469
Age	0.029	0.002	0.000	0.029	0.002	0.000	0.030	0.002	0.000
Household size	-0.032	0.022	0.149	-0.032	0.021	0.135	-0.006	0.022	0.785
Hygiene index	0.017	0.043	0.697	0.004	0.040	0.912	-0.042	0.045	0.344
Dist. to next hospital (min.)	-0.002	0.003	0.425	-0.002	0.003	0.440	-0.002	0.003	0.504
Const.	-3.523	0.308	0.000	-3.487	0.225	0.000	-3.820	0.261	0.000

- *Note*: This table shows the coefficients of logit regressions of a dummy indicating non-childbirth related admission to hospital on individual and household covariates. Covariates on the left, statistics on top.
- Sample: Member-level sample (panel, N = 12,852).
- Source: Baseline survey (2015).
- Columns (1), (4), (7) display the coefficient estimates from logit regressions, Columns (2), (5), (8) the standard errors and Columns (3), (6), (9) the p-values of testing that the coefficient is equal to zero, with one of three different proxies for poverty respectively. Standard errors are adjusted for 24 clusters in union councils.

2 Additional information on econometric approach

2-1 Estimation of propensity scores

We here describe the estimation of propensity scores for our PSM sample, based on the methodology suggested in Imbens and Rubin (2015).

First, we select a set of base variables, which we believe important for the selection model, guided by economic theory and program knowledge. As described in the main text, we believe the union council to be of fundamental importance as it may capture information on otherwise unobservable characteristics. Furthermore, we linearly include baseline values of the following variables: poverty score, the average monthly household income (winsorized), the household size, household-level usage of inpatient care (extensive margin and number of household members treated), as well as the minimum of the reported health status over all household members. We denote this set of variables the base variables K_B. For estimation of propensity scores on household member level, the set additionally includes age, gender, subjective health status, admittance to hospital (whether admitted at all and whether admitted more than once), and dummy variables for whether the member completed primary school and whether the member completed senior or higher education.

Second, we search for further baseline variables to be included linearly into the selection model in an iterative process. To start this process, we estimate logit models of assignment to treatment on the base variables in addition to one more variable of a choice set. Subsequently, we calculate likelihood ratio test statistics and test the null hypothesis that the additional variable has a zero coefficient. Of all variables in the choice set, we include the variable with the largest likelihood ratio statistic in the base model. Iteratively, we repeat this step, each time testing another variable and including the one with the largest likelihood ratio statistic. We stop when all these statistics are smaller than one. Our choice set includes all the baseline variables which are not already in K_B. The iterative process leads us to include ten variables on household and four variables on member level. We denote the union of K_B and selected additional variables as K_L, which will enter the selection model linearly.

Third, we select higher order terms to be included into the selection model. In line with Imbens and Rubin 2015, we restrict ourselves to quadratic and interaction terms of the variables in the set K_L and refrain from including higher order terms. We create interaction terms for all but the union-council-dummy variables and run the same iterative process as before. We refrain from including interactions with the union-council-dummies, as this would lead to overfitting, violating the common support assumption. This time, we stop when all likelihood ratio statistics are smaller than 2.7, as in Imbens and Rubin 2015. This leads us to include another 20 interaction terms on household and 23 terms on member level.

In conclusion, in addition to union council dummies, we match households on 36 and individuals on 40 linear and square baseline variables. We repeat the same variable selection algorithm for any subsample analysis.

2-2 Balancing of baseline variables in unmatched and matched sample

On average, insured households are slightly smaller and composed of slightly better educated members. They are also a little less poor and live in relatively larger villages, with a smaller distance to the next empanelled hospital as measured by GPS data. None of these differences are large in magnitude or highly significant, and only few are marginally significant. Notably, prior insurance knowledge is virtually non-existing in any of the two sub-population. Even more so, indicators generated at household member level, for which we can draw on a much larger sample size, are very well balanced in all variables, e.g., usage of inpatient care and cost thereof. We also compare the prevalence of different causes for inpatient care among household members (results not shown in the tables). Abdominal pain and diarrhea was cited more often among insured households (7% versus 2% for each disease, p-value 0.09), but there are no relevant differences in the incidence rates of the most common problems such as appendicitis (8-10%), heart attack (6-10%), diabetes (5-9%), or malaria (6%).

Table A.4: Balancing before and after matching - Household level

Variable	Unmatched	Mean			% reduct.	t-test	
	Matched	Insured	Uninsured	%bias	—bias—	t	p> $\|t\|$
		(1)	(2)	(3)	(4)	(5)	(6)
UC1	U	0.0344	0.0361	-0.900	.	-0.120	0.904
	M	0.0347	0.0444	-5.200	-469.5	-0.750	0.453
		.	.	.	.	.	.
UC2	U	0.0258	0.0614	-17.50	.	-2.420	0.016
	M	0.0260	0.0259	0	99.80	0.010	0.993
		.	.	.	.	.	.
UC3	U	0.0516	0.0325	9.500	.	1.220	0.222
	M	0.0521	0.0447	3.700	61.30	0.520	0.601
		.	.	.	.	.	.
UC4	U	0.0624	0.0253	18.20	.	2.280	0.023
	M	0.0607	0.0507	4.900	73	0.660	0.508
		.	.	.	.	.	.
UC5	U	0.0280	0.0397	-6.500	.	-0.870	0.382
	M	0.0282	0.0244	2.100	67.20	0.370	0.715
		.	.	.	.	.	.
UC6	U	0.0473	0.0433	1.900	.	0.250	0.802
	M	0.0477	0.0516	-1.900	2.800	-0.270	0.787
		.	.	.	.	.	.
UC7	U	0.0344	0.0325	1.100	.	0.140	0.889
	M	0.0347	0.0340	0.400	64.20	0.060	0.954
		.	.	.	.	.	.
UC8	U	0.0559	0.0433	5.800	.	0.750	0.452
	M	0.0564	0.0816	-11.60	-99.80	-1.510	0.132
		.	.	.	.	.	.
UC9	U	0.0366	0.0541	-8.400	.	-1.140	0.254
	M	0.0369	0.0428	-2.900	66	-0.460	0.643
		.	.	.	.	.	.
UC10	U	0.0516	0.0253	13.70	.	1.740	0.083
	M	0.0521	0.0458	3.300	76.10	0.440	0.658
		.	.	.	.	.	.
UC11	U	0.0538	0.0253	14.60	.	1.850	0.065
	M	0.0521	0.0489	1.600	88.90	0.220	0.827
		.	.	.	.	.	.
UC12	U	0.0409	0.0289	6.500	.	0.840	0.400

continued

Table A.4: Balancing before and after matching - Household level

Variable	Unmatched Matched	Mean Insured (1)	Uninsured (2)	%bias (3)	% reduct. —bias— (4)	t-test t (5)	p> $\|t\|$ (6)
	M	0.0391	0.0268	6.700	-2.400	1.040	0.297
		.	.	.	.	.	.
UC13	U	0.0323	0.0217	6.500	.	0.840	0.400
	M	0.0325	0.0300	1.500	76.40	0.220	0.827
		.	.	.	.	.	.
UC14	U	0.0473	0.0469	0.200	.	0.020	0.981
	M	0.0477	0.0415	2.900	-1532	0.460	0.649
		.	.	.	.	.	.
UC15	U	0.0495	0.0614	-5.200	.	-0.690	0.488
	M	0.0499	0.0457	1.800	64.70	0.300	0.765
		.	.	.	.	.	.
UC16	U	0.0237	0.0433	-10.90	.	-1.500	0.135
	M	0.0239	0.0254	-0.900	92.20	-0.150	0.880
		.	.	.	.	.	.
UC17	U	0.0495	0.0397	4.700	.	0.610	0.540
	M	0.0499	0.0682	-8.800	-87.50	-1.180	0.239
		.	.	.	.	.	.
UC18	U	0.0602	0.0505	4.200	.	0.550	0.582
	M	0.0607	0.0507	4.400	-3.500	0.660	0.508
		.	.	.	.	.	.
UC19	U	0.0301	0.0758	-20.50	.	-2.850	0.004
	M	0.0304	0.0201	4.600	77.50	0.990	0.320
		.	.	.	.	.	.
UC20	U	0.0452	0.0433	0.900	.	0.120	0.907
	M	0.0456	0.0410	2.200	-147.6	0.340	0.734
		.	.	.	.	.	.
UC21	U	0.0387	0.0397	-0.500	.	-0.070	0.946
	M	0.0369	0.0445	-3.900	-665.1	-0.590	0.557
		.	.	.	.	.	.
UC22	U	0.0215	0.0505	-15.60	.	-2.170	0.031
	M	0.0217	0.0269	-2.800	82.10	-0.510	0.609
		.	.	.	.	.	.
UC23	U	0.0366	0.0397	-1.600	.	-0.220	0.828
	M	0.0369	0.0345	1.200	24.60	0.190	0.846
		.	.	.	.	.	.
HH size	U	8.021	8.285	-10.10	.	-1.350	0.177
	M	8.022	7.941	3.100	69.40	0.500	0.617
		.	.	.	.	.	.
% mem. w/o education in HH	U	0.384	0.407	-10	.	-1.330	0.184
	M	0.385	0.375	4.200	57.70	0.660	0.511
		.	.	.	.	.	.
HH with >=1 mem. w.sec. education	U	0.357	0.325	6.800	.	0.890	0.375
	M	0.360	0.331	6.200	8.100	0.940	0.347
		.	.	.	.	.	.
HH w/o mem. having completed primary school	U	0.226	0.267	-9.600	.	-1.270	0.203
	M	0.228	0.237	-2.100	77.80	-0.330	0.742
		.	.	.	.	.	.
% children under 6 years within HH	U	0.110	0.114	-3.100	.	-0.410	0.685
	M	0.111	0.114	-2.700	11	-0.410	0.684
		.	.	.	.	.	.
% elderly over 65 years within HH	U	0.0232	0.0250	-2.500	.	-0.320	0.746
	M	0.0232	0.0159	10.20	-304.9	1.680	0.093
		.	.	.	.	.	.
% female mem. within HH	U	0.482	0.471	6.400	.	0.850	0.398
	M	0.482	0.478	2.100	67.60	0.320	0.750
		.	.	.	.	.	.
% wage earning mem. within HH	U	0.224	0.235	-8.200	.	-1.100	0.271
	M	0.224	0.232	-6	26.60	-0.970	0.331
		.	.	.	.	.	.
poverty score	U	12.52	12.30	6	.	0.790	0.428
	M	12.54	12.49	1.500	75.10	0.220	0.823
		.	.	.	.	.	.
avg. monthly HH income (PKR)	U	20188	19613	2.700	.	0.340	0.732
	M	20265	19900	1.700	36.50	0.260	0.798
		.	.	.	.	.	.
receiving transfers	U	0.688	0.635	11.20	.	1.480	0.140
	M	0.688	0.672	3.400	69.70	0.520	0.603
		.	.	.	.	.	.

continued

Table A.4: Balancing before and after matching - Household level

Variable	Unmatched Matched	Mean Insured (1)	Mean Uninsured (2)	%bias (3)	% reduct. —bias— (4)	t-test t (5)	t-test p> $\|t\|$ (6)
receiving remittances	U	0.110	0.108	0.400	.	0.060	0.954
	M	0.111	0.127	-5.200	-1087	-0.760	0.445
		.	.	.	.	.	.
loans (PKR, win95)	U	65561	62069	3	.	0.390	0.696
	M	65605	51631	11.90	-300.1	1.920	0.055
		.	.	.	.	.	.
savings (PKR, win95)	U	503.1	515.5	-0.500	.	-0.070	0.947
	M	507.5	470.0	1.600	-202.8	0.240	0.809
		.	.	.	.	.	.
wealth index	U	-0.595	-0.545	-3.100	.	-0.420	0.676
	M	-0.597	-0.566	-2	36.60	-0.310	0.755
		.	.	.	.	.	.
hygiene index	U	0.180	0.141	2.700	.	0.350	0.723
	M	0.179	0.205	-1.800	33.20	-0.280	0.780
		.	.	.	.	.	.
number of HH in village	U	1361	1273	8.300	.	1.090	0.277
	M	1363	1312	4.800	41.70	0.740	0.457
		.	.	.	.	.	.
% eligible HH in village	U	0.322	0.302	17.90	.	2.360	0.018
	M	0.321	0.324	-2.500	86.20	-0.390	0.700
		.	.	.	.	.	.
reported distance to next hosp. (minutes)	U	45.27	42.47	10.70	.	1.390	0.166
	M	45.40	44.67	2.800	74	0.410	0.681
		.	.	.	.	.	.
linear distance to next empanelled hosp. (km)	U	10.26	10.59	-4.100	.	-0.580	0.560
	M	10.26	10.27	-0.100	96.50	-0.030	0.977
		.	.	.	.	.	.
Strongest disagreement with: would never go to govt. hosp.	U	0.477	0.487	-2	.	-0.260	0.793
	M	0.477	0.413	12.80	-545.4	1.960	0.050
		.	.	.	.	.	.
willingness to take fin. risk <= 5 (scale 1 to 10)	U	0.516	0.394	24.80	.	3.260	0.001
	M	0.514	0.519	-0.900	96.30	-0.140	0.892
		.	.	.	.	.	.
having heard of insurance	U	0.0129	0.0144	-1.300	.	-0.180	0.861
	M	0.0130	0.00468	7.200	-442.3	1.350	0.177
		.	.	.	.	.	.
negl. health care due to fin. reasons	U	0.932	0.913	6.800	.	0.370	0.714
	M	0.932	0.918	5.200	24.70	0.300	0.762
		.	.	.	.	.	.
citing health shock as a risk	U	0.961	0.942	8.900	.	1.200	0.231
	M	0.961	0.967	-2.700	69.50	-0.470	0.637
		.	.	.	.	.	.
citing health shock as main risk	U	0.682	0.686	-0.900	.	-0.120	0.905
	M	0.681	0.706	-5.400	-503.4	-0.830	0.404
		.	.	.	.	.	.
ladder of life rating <= 5 (scale 1/10)	U	0.733	0.751	-4	.	-0.530	0.598
	M	0.735	0.739	-0.800	79.20	-0.130	0.900
		.	.	.	.	.	.
fin. satisfaction <= 5 (scale 1/10)	U	0.705	0.675	6.500	.	0.870	0.387
	M	0.705	0.704	0.200	97.30	0.030	0.978
		.	.	.	.	.	.
dif'ty finding money for health care >=8 (scale 1/10)	U	0.488	0.498	-2	.	-0.260	0.792
	M	0.486	0.453	6.600	-230	1.010	0.315
		.	.	.	.	.	.
use of inp. care (ext. margin, HH-level)	U	0.252	0.300	-10.70	.	-1.430	0.154
	M	0.249	0.203	10.50	2.600	1.700	0.090
		.	.	.	.	.	.
inp. care (no. of mem. per HH)	U	0.312	0.379	-10.60	.	-1.420	0.155
	M	0.310	0.254	8.900	16.60	1.480	0.140
		.	.	.	.	.	.
case of neglected health care	U	0.157	0.166	-2.500	.	-0.330	0.745
	M	0.158	0.134	6.600	-166.5	1.040	0.299
		.	.	.	.	.	.
case of outpatient care	U	0.817	0.769	11.90	.	1.590	0.113

continued

Table A.4: Balancing before and after matching - Household level

Variable	Unmatched / Matched	Mean: Insured (1)	Mean: Uninsured (2)	%bias (3)	% reduct. |bias| (4)	t-test: t (5)	t-test: p>|t| (6)
	M	0.816	0.861	-11.20	5.800	-1.880	0.061
		.	.	.	.	.	.
Min. of health status within hh (scale of 1/ 5)	U	3.112	2.978	11.80	.	1.550	0.121
	M	3.111	3.263	-13.40	-14	-2.060	0.039

- *Note*: This table shows the balancing of UCs and baseline covariates in the unmatched (U) and matched (M) samples. Variables on the left, statistics on top.
- Sample: Household-level PSM sample (N=795).
- Source: Baseline survey (2015).
- Columns (1) and (2) display the mean of the insured and uninsured households in the matched and unmatched sample respectively. Column (3) and (4) show the percentage bias before and after matching, together with the achieved reduction in the absolute value of the bias. The standardized percentage bias is the percent-difference of the sample means in the insured and the uninsured subsamples as a percentage of the square root of the average of the sample variances in the insured and uninsured groups (formula from Rosenbau and Rubin, 1985). Columns (5) and (6) display the t-static and the p-value testing for equality of means in the two samples. T-tests are based on regressions of the variable on a treatment indicator.
- The suffix *win99* indicates that we winsorized the variable at the 99th percentile level.

Table A.5: Balancing before and after matching - Member level

Variable	Unmatched / Matched	Mean: Insured (1)	Mean: Uninsured (2)	%bias (3)	% reduct. |bias| (4)	t-test: t (5)	t-test: p>|t| (6)
age (win99)	U	22.02	22.24	-1.300	.	-0.470	0.637
	M	21.97	21.88	0.500	61.70	0.210	0.830
		.	.	.	.	.	.
school-aged (6 to 16)	U	0.391	0.375	3.200	.	1.160	0.246
	M	0.392	0.395	-0.600	80	-0.270	0.788
		.	.	.	.	.	.
female	U	0.483	0.482	0.200	.	0.060	0.952
	M	0.483	0.479	0.600	-289.9	0.270	0.785
		.	.	.	.	.	.
prim. school not comp'd	U	0.633	0.644	-2.400	.	-0.890	0.373
	M	0.632	0.636	-0.800	66.30	-0.350	0.727
		.	.	.	.	.	.
comp'd sec. educ. or higher	U	0.0836	0.077	2.200	.	0.790	0.432
	M	0.0819	0.0855	-1.300	39.20	-0.550	0.582
		.	.	.	.	.	.
worked for salary in previous month	U	0.179	0.188	-2.300	.	-0.860	0.391
	M	0.177	0.185	-1.900	18.50	-0.820	0.414
		.	.	.	.	.	.
working in agriculture	U	0.021	0.027	-4	.	-1.470	0.141
	M	0.0206	0.0215	-0.600	85.80	-0.250	0.799
		.	.	.	.	.	.
in schooling	U	0.296	0.282	3.100	.	1.130	0.259
	M	0.296	0.293	0.700	77.40	0.300	0.767
		.	.	.	.	.	.
main occupation: HH/ child care	U	0.228	0.222	1.400	.	0.490	0.621
	M	0.228	0.233	-1.200	11	-0.510	0.611
		.	.	.	.	.	.
resp'ble for fin. decisions	U	0.384	0.392	-1.800	.	-0.660	0.512
	M	0.382	0.383	-0.200	88.50	-0.090	0.930
		.	.	.	.	.	.
resp'ble for health decisions	U	0.271	0.274	-0.500	.	-0.200	0.841
	M	0.270	0.276	-1.300	-141.4	-0.560	0.573
		.	.	.	.	.	.
total cost of last treatment (PKR, win99)	U	24990	29651	-9.900	.	-0.750	0.456
	M	25252	22309	6.200	36.80	0.560	0.576

continued

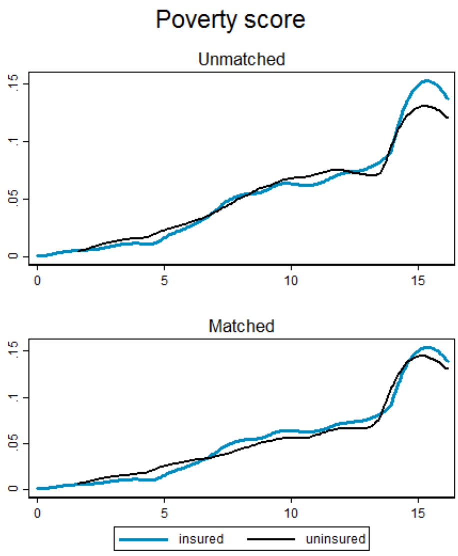
Table A.5: Balancing before and after matching - Member level

Variable	Unmatched Matched	Mean Insured (1)	Mean Uninsured (2)	%bias (3)	% reduct. —bias— (4)	t-test t (5)	t-test p> $\|t\|$ (6)
		.	.	.	.	.	.
usage of inpatient care	U	0.0343	0.0474	-6.600	.	-2.470	0.013
	M	0.0335	0.0360	-1.200	81.20	-0.570	0.567
		.	.	.	.	.	.
use of pub. hos. for inp. care	U	0.0231	0.0303	-4.500	.	-1.660	0.097
	M	0.0225	0.0249	-1.500	66.50	-0.670	0.501
		.	.	.	.	.	.
health step of HH mem. (scale: 1/5)	U	4.385	4.315	7.300	.	2.680	0.007
	M	4.383	4.386	-0.300	95.80	-0.130	0.895

► *Note*: This table shows the balancing of UCs and baseline covariates in the unmatched (U) and matched (M) samples. Variables on the left, statistics on top.
► Sample: Member-level PSM sample (panel, N=6,007).
► Source: Baseline survey (2015).
► Columns (1) and (2) display the mean of the insured and uninsured households in the matched and unmatched sample respectively. Column (3) and (4) show the percentage bias before and after matching, together with the achieved reduction in the absolute value of the bias. The standardized percentage bias is the percent-difference of the sample means in the insured and the uninsured subsamples as a percentage of the square root of the average of the sample variances in the insured and uninsured groups (formula from Rosenbau and Rubin, 1985). Columns (5) and (6) display the t-static and the p-value testing for equality of means in the two samples. T-tests are based on regressions of the variable on a treatment indicator.
► The suffix *win99* indicates that we winsorized the variable at the 99th percentile level.

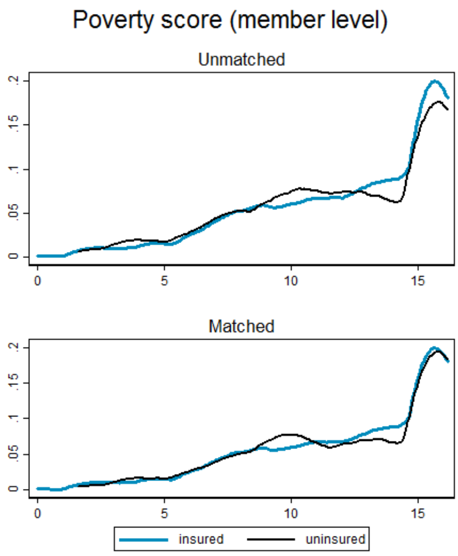
2-3 Distribution of poverty score among insured and matched uninsured sample

Figure A.2: Distribution of poverty score in households of PSM sample by insurance status



► *Note*: This figure shows the distribution of the poverty score among insured and uninsured households in the unmatched (top) and matched (bottom) samples.
► Sample: Household-level PSM sample (panel, N=795).
► Source: Endline survey (2017).

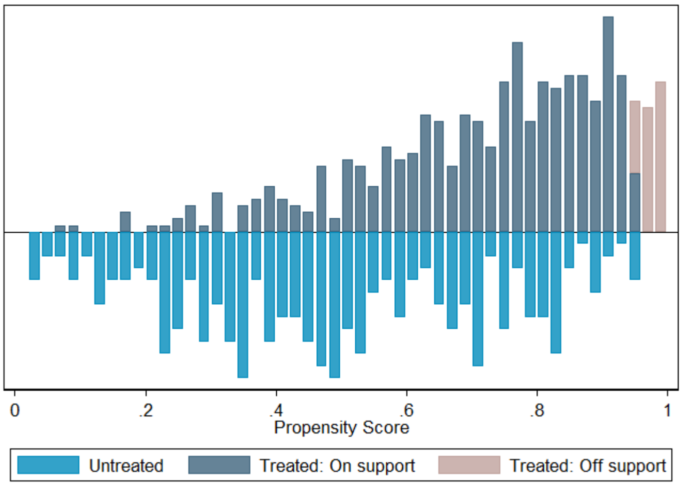
Figure A.3: Distribution of poverty score in household members of PSM sample by insurance status



► *Note*: This figure shows the distribution of the poverty score among members of insured and uninsured households in the unmatched (top) and matched (bottom) samples.
► Sample: Member-level PSM sample (panel, N=6,007).
► Source: Endline survey (2017).

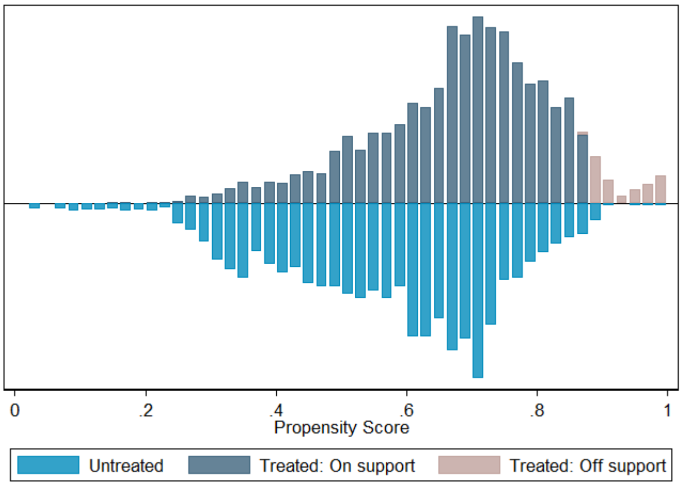
2-4 Common support

Figure A.4: Distribution of propensity scores in households of PSM sample by insurance status



► *Note*: This figure shows the distribution of the propensity scores among insured (top) and uninsured (bottom) households. The blue bars indicate common support, whereas rose bars are off the common support.
► Sample: Household-level PSM sample (panel, N=795).
► Source: Propensity scores calculated from baseline survey (2015) and insurance status in endline survey (2017).

Figure A.5: Distribution of propensity scores in household members of PSM sample by insurance status

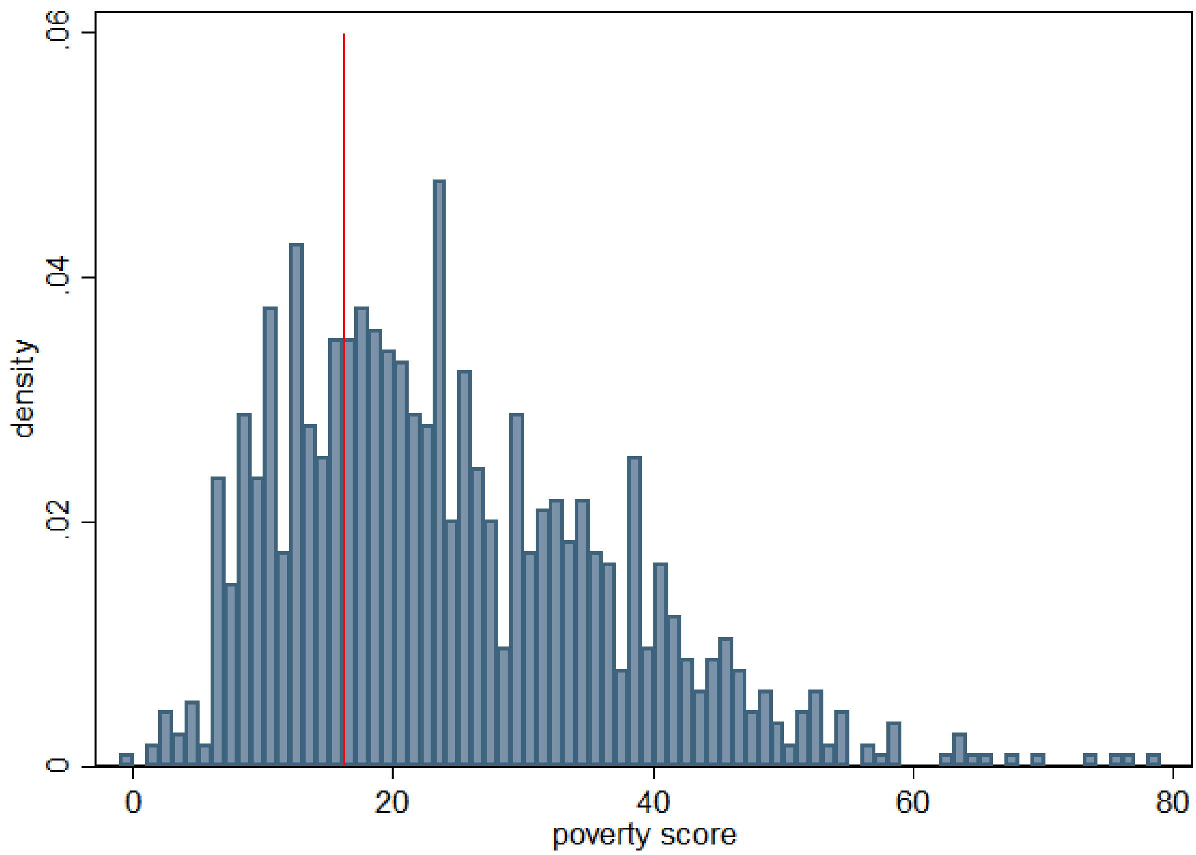


► *Note*: This figure shows the distribution of the propensity scores among individuals in insured (top) and uninsured (bottom) households. The blue bars indicate common support, whereas rose bars are off the common support.
► Sample: Member-level PSM sample (panel, N=6,007).
► Source: Propensity scores calculated from baseline survey (2015) and insurance status in endline survey (2017).

2-5 McCrary density test

To assess manipulative sorting more formally, Figure A.6 depicts the distribution of the poverty score in our sample, restricted to those households which were sampled at random and not for oversampling below or around the cut-off score. One indication of precise sorting would be a bunching of households just below the cut-off poverty score, implying a discontinuity in the density of the poverty score at this point (McCrary 2008). Figure A.6 suggests that comparable proportions of households in the respective districts fall on both sides. This visual evidence is confirmed by a density test as proposed in McCrary (2008) which fails to detect a significant discontinuity, see Figure A.7 below.

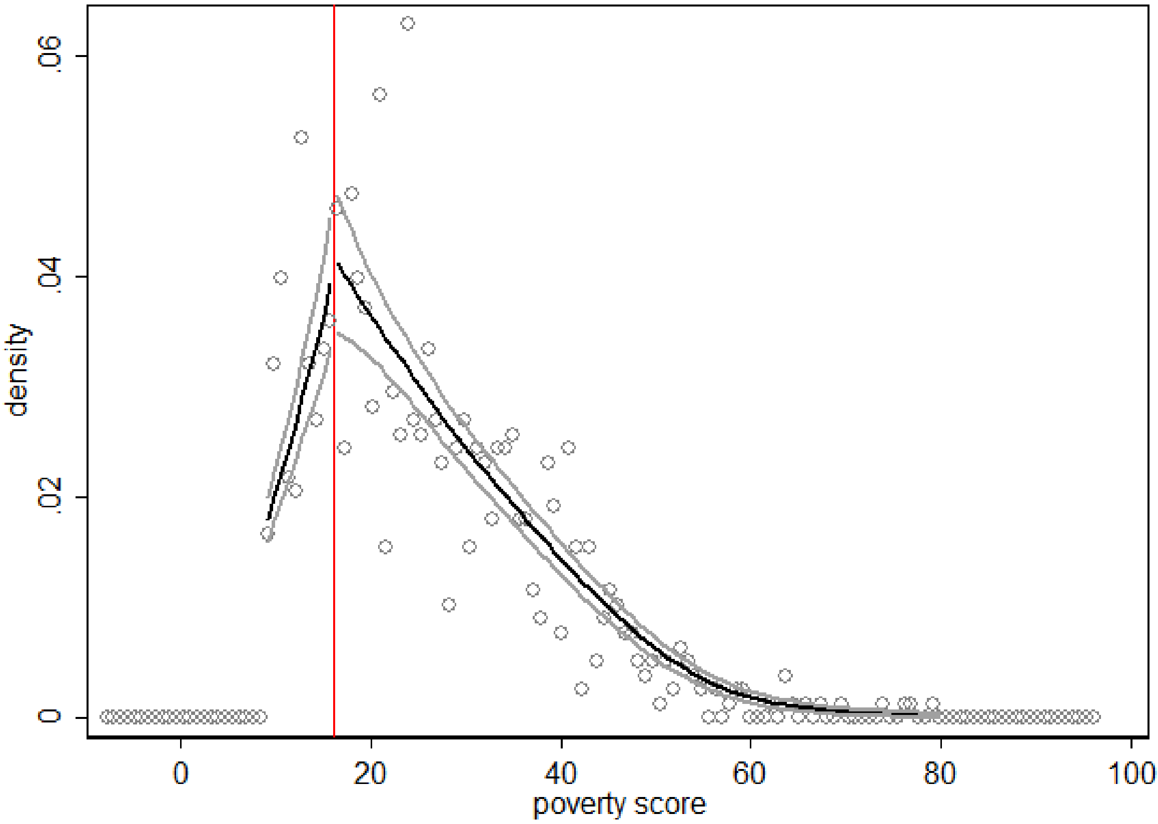
Figure A.6: Distribution of poverty score in random sample



► *Notes:* This figure shows the distribution of the poverty score in the random panel sample. The red vertical line indicates the cut-off score of 16.17.
► Sample: Household-level random sample (N=1,152).
► Source: Endline survey (2017).

Additionally, note that it is also highly unlikely that households could influence their poverty score. This score was assigned based on a PMT in a targeting survey in 2010, which used 23 household-level indicators (OLeary, Cheema, Hunt, Carraro, and Pellerano 2011). Households may have seized the opportunity to manipulate their poverty score by exaggerating their poverty status in order to obtain social benefits in the future. However, the fear of perfect manipulation is mitigated by the fact that each variable is weighted differently in the PMT (Uddin, Rafi, Uddin, and Khurshid 2013). Moreover, to the best of our knowledge, there is no official household questionnaire available online, and no list with all variables including their

Figure A.7: McCrary density test



► *Notes:* This figure illustrates the result of the McCrary density test around the cut-off 16.17. See McCrary (2008) for details.
► Sample: Household-level random sample (N=1,152).
► Source: Endline survey (2017).

weighting can be found. Pakistani households and the enumerating survey staff could not have known the exact method of how the poverty score was computed and therefore failed to select into treatment. This result is also supported by the report on the initial targeting survey of the BISP program, which also finds no sharp break in the density of the poverty score and no significant jump at the threshold for baseline covariates and outcome variables (OLeary et al. 2011).

3 Robustness checks

3-1 Specification tests for PSM

Table A.6: Robustness of effects on inpatient care consumption - PSM estimates

Main outcome (see Table 5 in main text)	Main spec.		Random sample		Cluster unit HH		Different PS	
	β_{ATT}	S.E.	β^1_{ATT}	S.E.	β^2_{ATT}	S.E.	β^3_{ATT}	S.E.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Individual outcomes								
Usage of inpatient care	**-0.002**	0.011	0.006	0.013	-0.002	0.010	-0.004	0.011
N (uninsured/ insured)	*2,526/ 3,638*		*1,480/ 2,608*		*2,526/ 3,638*		*2,111/ 3,751*	
conditional on usage of inpatient care								
More than one admittance	**0.022**	0.077	0.051	0.112	0.022	0.079	0.026	0.065
N (uninsured/ insured)	*107/ 212*		*71/ 158*		*107/ 212*		*107/ 202*	
Usage of private versus public hospitals	**0.068**	0.075	-0.027	0.104	0.068	0.086	0.075	0.079
N (uninsured/ insured)	*101/ 202*		*65/ 152*		*101/ 202*		*101/ 193*	
Household outcomes								
Neglected health care	**0.007**	0.026	-0.002	0.040	0.007	–	-0.000	0.026
N (uninsured/ insured)	*277/ 461*		*194/ 335*		*277/ 461*		*277/ 497*	

- *Note*: This table illustrates the robustness of our main PSM estimation results against alternative specifications. Outcome variables on the left, different econometric models/sample on top.
- Samples: Member-level and household-level PSM samples (panel, varying sample size depending on common support restriction, overall: 795 households with 6,007 members).
- Source: Endline survey (2017).
- Columns (1) and (2) show the coefficient and standard error for the average treatment effect on the treated, estimated using propensity score kernel matching in our main specification (corresponding to Table 1.6 in main paper). Subsequent columns show results from various specifications:
  - Column (3) and (4): Coefficient estimated on only randomly selected sample (excluding over-sampled households around the cut-off).
  - Column (5) and (6): Coefficient estimates correspond to main specification, standard error are clustered on household-level (9,999 bootstraps).
  - Column (7) and (8): Propensity scores estimated from probit regression on union council dummies, poverty score, household size, dummy indicating risk aversion, and instance of neglected health care at baseline (household level), and on case of inpatient care at baseline, age and gender (member level). Matching and inference as in main specification, i.e., kernel matching and standard errors are clustered on UC-level (9,999 bootstraps).
- The statistical significance is given as follows: * indicates $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$, with the null hypothesis of the two-sided test being a zero effect size.

3-2 Specification tests for RDD

Table A.7: Robustness of effects on inpatient care consumption - RDD estimates

Main outcome (see Table 5 in main text)	Main spec. β_{ITT}	S.E.	With UC-cluster β^1_{ITT}	S.E.	With covariates β^2_{ITT}	S.E.	Fuzzy design β^3_{LATE}	S.E.	Local constant β^4_{ITT}	S.E.	IK bandwidth β^5_{ITT}	S.E.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Individual outcomes												
Usage of inpatient care	**-0.002**	0.007	-0.002	0.010	0.002	0.006	-0.003	0.011	0.000	0.006	-0.003	0.009
N (left/ right of cut-off)	*4,336/ 3,227*		*4,503/ 3,321*		*4,322/ 3,227*		*4,336/ 3,227*		*3,544/ 2,740*		*2,591/ 2,214*	
conditional on usage of inpatient care												
More than one admittance	**0.062**	0.054	0.057	0.090	0.062	0.055	0.094	0.082	0.046	0.046	0.087	0.064
N (left/ right of cut-off)	*255/ 174*		*284/ 201*		*248/ 172*		*255/ 174*		*205/ 151*		*168/ 134*	
Usage of private versus public hospitals	**0.237*** **	0.073	0.226***	0.081	0.286***	0.047	0.363***	0.112	0.177***	0.060	0.279***	0.081
N (left/ right of cut-off)	*186/ 131*		*192/ 134*		*160/ 121*		*186/ 131*		*170/ 124*		*143/ 128*	
Household outcomes												
Neglected health care	**0.001**	0.022	0.002	0.023	0.015	0.019	0.002	0.034	0.005	0.018	0.011	0.026
N (left/ right of cut-off)	*581/ 466*		*557/ 460*		*584/ 467*		*584/ 468*		*577/ 464*		*362/ 332*	

- *Note*: This table illustrates the robustness of our main RDD estimation results against alternative specifications. Outcome variables on the left, different econometric models on top.
- Samples: Member-level and household-level RDD samples (panel, varying sample size depending on selected bandwidth, overall: 1,849 households with 12,864 members).
- Source: Endline survey (2017).
- Columns (1) and (2) show the coefficient and standard error for the intention-to-treat effect around the cut-off, estimated using robust local linear RDD estimations of our main specification (corresponding to Table 1.6 in main paper). Subsequent columns show results from various robustness checks:
  - Column (3) and (4): Cluster-robust plug-in residuals variance estimator with clustering on union council level.
  - Column (5) and (6): Estimations include additional covariates (gender, age and wealth index for individual level variables, only wealth index for household level variables), estimated with method proposed by Frölich and Huber (2019).
  - Column (7) and (8): Left-sided fuzzy design with insurance status used as first stage below the cut-off, all above the cut-off considered treated; bandwidth selection procedure for the sharp RD model.
  - Column (9) and (10): Coefficient estimated using local constant smoothing.
  - Column (11) and (12): Estimation using local linear regression models on both sides of the cutoff, with bandwidth from Imbens and Kalyanaraman (2009); analytical standard errors based on the regressions.
- The statistical significance is given as follows: * indicates $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$, with the null hypothesis of the two-sided test being a zero effect size.

3-3 Effects on non-log financial outcomes

Table A.8: Effects on non-log financial outcomes

Outcome	Control Mean (1)	PSM β_{ATT} (2)	PSM S.E. (3)	RDD β_{ITT} (4)	RDD S.E. (5)
Effects on financial outcomes conditional on inpatient usage (member-level)					
Out-of-pocket expenditures (PKR, win99)	16,275	**-1,304**	5,210	**4,560**	3,853
Cost for diagnosis and treatment (PKR, win99)	4,261	**-1,110**	2,003	**1,591**	1,222
Cost for medicines (PKR, win99)	8,775	**-565,5**	2,626	**3,662**	1,896
N (uninsured/ insured, left/ right of cut-off)		*107/ 212*		*255/ 172*	

- *Note*: This table shows estimates of non-logarithmic financial outcomes, compare to Table 1.8. Different outcomes variables on the left, statistics on top.
- Samples: PSM sample and RDD sample (panel).
- Source: Endline survey (2017).
- Column (1) displays the mean for the matched controls of uninsured, but eligible households (poverty score below 16.17). Columns (2) and (3) show the coefficient and standard error for the average treatment effect on the treated, estimated using propensity score kernel matching. Columns (4) and (5) show the coefficient and standard error for the intention-to-treat effect for households just below the cut-off score of 16.17, estimated using a sharp regression discontinuity design.
- *Note on PSM:* S.E. are derived by bootstrapping the whole process of estimation of propensity scores, restricting the sample to common support, matching and ATT estimation. Unit of clustering is the union council. Number of bootstraps: 9,999. Reported sample size refers to area of common support.
- *Note on RDD:* Estimated using local linear regression models with a triangular kernel and bandwidth estimated to minimize the mean squared error; S.E. as proposed in Calonico et al. (2014). Reported sample size refers to observations within selected bandwidth.
- The statistical significance is given as follows: * indicates $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$, with the null hypothesis of the two-sided test being a zero effect size.
- The suffix *win99* indicates that we winsorized the variable at the 99th percentile level.

3-4 Falsification tests

Table A.9: Pseudo effects on baseline and invariant endline outcomes

Outcome	Control Mean (1)	PSM β_{ATT} (2)	PSM S.E. (3)	RDD β_{ITT} (4)	RDD S.E. (5)
Individual outcomes					
Female (endline)	0.473	**0.008**	0.007	**-0.010**	0.016
N (uninsured/ insured, left/ right of cut-off)		*2,111/ 3,638*		*6,008/ 6,856*	
conditional on usage of inpatient care					
Cause of hospitalization: Accident/ injury (endline)	0.061	**0.029**	0.047	**0.001**	0.037
N (uninsured/ insured, left/ right of cut-off)		*107/ 212*		*333/ 409*	
Household outcomes					
Childbirth (endline)	0.106	**0.043**	0.034	**0.005**	0.033
N (uninsured/ insured, left/ right of cut-off)		*277/ 461*		*796/ 1,053*	
Hygiene index (endline)	0.070	**-0.098**	0.165	**-0.000**	0.133
N (uninsured/ insured, left/ right of cut-off)		*277/ 461*		*796/ 1,053*	
Childbirth (baseline)	0.145	**-0.050**	0.035	**-0.041**	0.029
N (uninsured/ insured, left/ right of cut-off)		*277/ 461*		*796/ 1,053*	
Hygiene index (baseline)	0.205	**-0.026**	0.169	**0.131**	0.130
N (uninsured/ insured, left/ right of cut-off)		*277/ 461*		*796/ 1,053*	

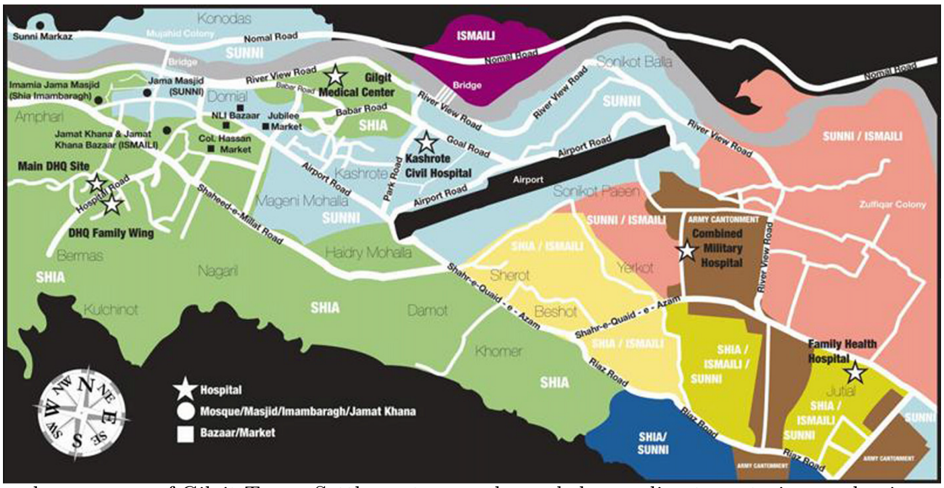
- *Note*: This table shows estimates for baseline variables or invariant endline variables (pseudo effects), using our main identification specification. Outcome variables on the left, different econometric models and statistics on top. Note that we do not use baseline variables for member-level outcomes as these were used for matching.
- Samples: Member-level and household-level PSM and RDD samples (panel, varying N).
- Source: Baseline (2015) and endline survey (2017).
- Column (1) displays the mean for the matched controls of uninsured, but eligible households (poverty score below 16.17). Columns (2) and (3) show the coefficient and standard error for the average treatment effect on the treated, estimated using propensity score kernel matching. Columns (4) and (5) show the coefficient and standard errors for the intention-to-treat effect for households just below the cut-off score of 16.17, estimated using a sharp regression discontinuity design.
- *Note on PSM:* S.E. are derived by bootstrapping the whole process of estimation of propensity scores, restricting the sample to common support, matching, and ATT estimation. Unit of clustering is the union council. Number of bootstraps: 9,999. Reported sample size refers to area of common support (overall sample size: 795 households with 6,007 members).
- *Note on RDD:* Estimated using local linear regression models with a triangular kernel and bandwidth estimated to minimize the mean squared error; S.E. as proposed in Calonico et al. (2014). Reported sample size refers to observations within selected bandwidth (overall sample size: 1,842 households with 12,862 members).
- The statistical significance is given as follows: * indicates $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$, with the null hypothesis of the two-sided test being a zero effect size.

Appendix B

Appendix to Chapter 2

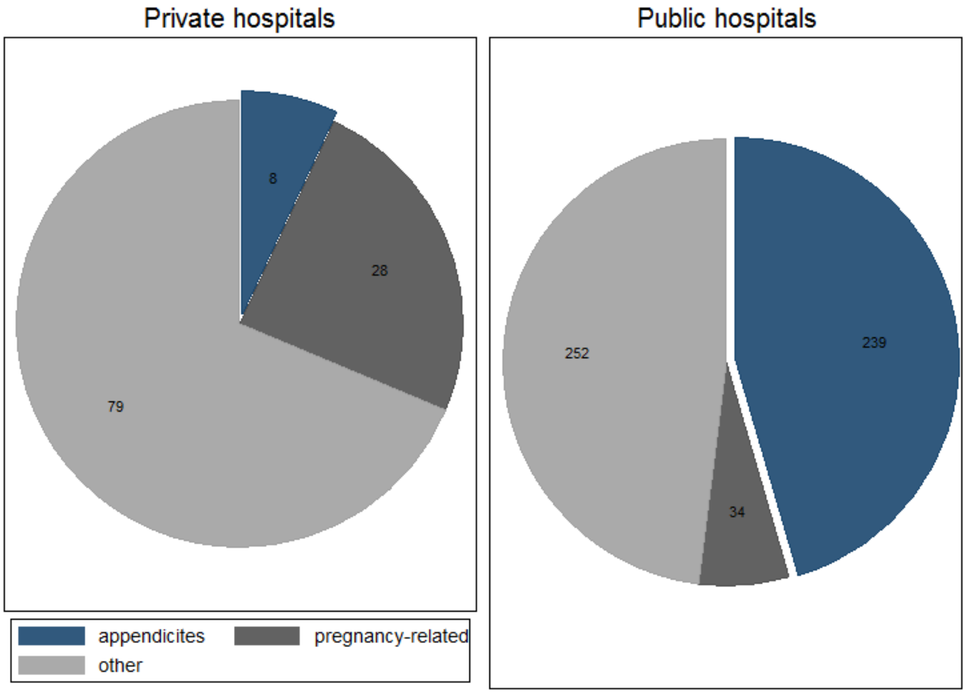
1 Additional figures

Figure B.1: Sectarian geography of Gilgit Town



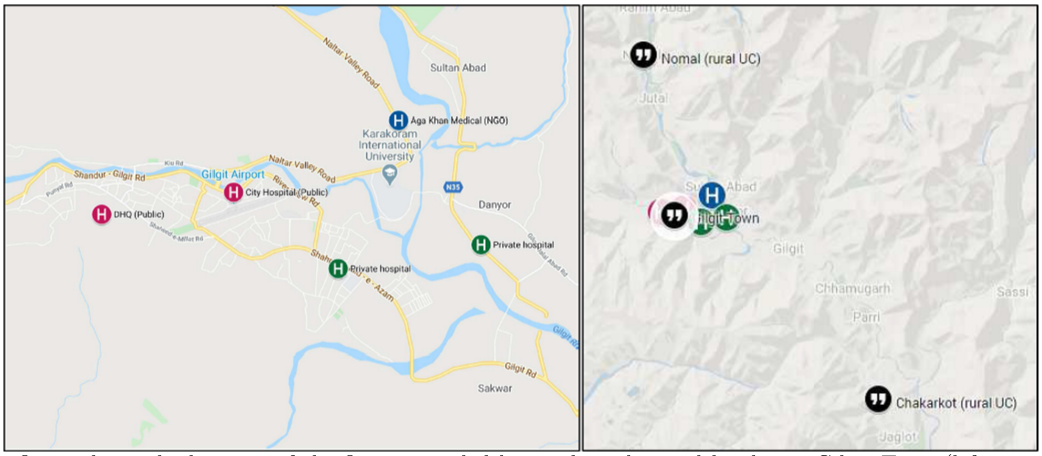
► *Note:* The figure shows a map of Gilgit Town. Settlements are color-coded according to sectarian predominance. Some hospitals marked by a star. The DHQ is located in the left (green Shia area).
► Source: Map by Carly Murray and Duane Jones, Dalhousie University MedIT Computing + Media Services, as published in Varley (2010).
► The map is intended to give a general imprecision of the mangitude of the sectarian divide in Gilgit Town. It likely does not exactly reflect the current situation and be partially outdated due to in-town migration following conflicts in more recent years.

Figure B.2: Reported causes of admission in public and private hospitals



► *Note:* The figure shows the causes of admission in private and public hospitals respectively from January to December 2017.
► Source: Own illustration of official claim data, as reported in the Annual Report (2017), available at http://sehathifazat.gog.pk/index.php/downloads/, retrieved on May 21, 2020.

Figure B.3: Location of hospitals within Gilgit Town and of rural areas



► *Note:* The figure shows the location of the five empaneled hospitals within and bordering Gilgit Town (left picture), and the location of the two rural areas considered in this study relative to Gilgit Town (right picture).
► Left picture: Gilgit Town lies to the South and East of the Gilgit River. Public hospitals are marked in red, private hospitals green, and the NGO-run hospital in blue. Right picture: Gilgit Town is in the center, the union council of Nomal to the North, and the union council of Charkakot to the South. There is only one road connecting the rural areas with Gilgit Town.
► Source: Google Maps.

2 Additional tables

Table B.1: Continuity of covariates - sharp design

Variable	Household sample			Member sample			Conditional sample		
	β^H_{RDD}	S.E.	p-val.	β^M_{RDD}	S.E.	p-val.	β^C_{RDD}	S.E.	p-val.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Urban	-0.068	0.157	0.660	-0.050	0.159	0.752	-0.167	0.211	0.428
Household size	-1.324	0.733	0.071	-2.017	1.096	0.066	-2.441	0.845	0.004
Avg. Monthly HH income (PKR, win99)	-4,653	3,476	0.181	-6,112	4,145	0.140	-11,905	5,524	0.031
Wealth index	-1.073	0.782	0.170	-1.141	0.740	0.123	-2.789	1.264	0.027
Reported distance to next hosp. (minutes, win99)	143.9	187.4	0.442	74.476	165.440	0.653	93.291	221.730	0.674
No HH member completed primary school	0.145	0.085	0.089	0.114	0.088	0.200	0.181	0.126	0.150
Some HH member with higher educ.	-0.162	0.117	0.164	-0.155	0.126	0.218	-0.422	0.165	0.011
Age				-0.136	1.367	0.921	-3.906	3.640	0.283
Female				-0.006	0.038	0.866	-0.106	0.117	0.362
N (left/right of cut-off	*170/231*			*1,408/1,822*			*182/326*		

- *Note*: This table contains rdd estimation results using covariates as outcomes. Variables on the left, statistics on top.
- Sample: Columns (1) to (3): Household sample (panel, N=401); Columns (4) to (6): Member sample (panel, N=3,230); Columns (7) to (9): Conditional sample (panel, N=508)
- Source: Baseline survey (2016)
- Columns (1), (4), and (7) display the coefficient estimates, Columns (2), (5), and (8) the associated standard errors, and Columns (3), (6), and (9) the p-value of the null hypothesis of zero effect size, for the three samples respectively. All estimates are obtained by running local linear regression models to the left and right of the cut-off, using the bias-corrected bandwidth that minimizes the mean squared error, as implemented in the Stata command rdrobust by Calonico et al. (2017). For the member and conditional sample, standard errors are clustered on household level.

Table B.2: Specification test: Control district Ghizer

Variable	β_{RDD} (1)	S.E. (2)	p-val. (3)
I. Health care financing			
Fin. Satisfaction $\leq$ 5(scale 1 to 10)	0.508	0.501	0.310
Citing health shock as main risk	-0.617	0.328	0.060
Dif.ty finding money for health care $\leq$ 8 (scale 1 to 10)	0.211	0.259	0.415
N (left / right of cut-off)		*41 / 89*	
II. Health care usage			
Case of inpatient care	0.055	0.063	0.385
N (left / right of cut-off)		*310 / 655*	
Neglected health care	-0.099	0.156	0.524
Case of outpatient care	-0.068	0.403	0.865
N (left / right of cut-off)		*41 / 89*	

- *Note*: This table contains rdd estimation results on the control sample in the district of Ghizer, where the program was not implemented. Variables on the left, statistics on top.
- Sample: Control sample (panel, varying N)
- Source: Endline survey (2018)
- Columns (1) display the coefficient estimates, Columns (2) the associated standard errors, and Columns (3) the p-value of the null hypothesis of zero effect size. All estimates are obtained by running local linear regression models to the left and right of the cut-off, using the bias-corrected bandwidth that minimizes the mean squared error, as implemented in the Stata command rdrobust by Calonico et al. (2017). For the member, standard errors are clustered on household level.

Table B.3: SPECIFICATION TEST: INVARIANT ENDLINE OUTCOMES

Variable		Fuzzy design			Sharp design		
		β_{LATE}	S.E.	p-val.	β^{M}_{ITT}	S.E.	p-val.
		(1)	(2)	(3)	(4)	(5)	(6)
Childbirth		-0.563	0.305	0.065	-0.173	0.114	0.128
	N (left/right of cut-off)			*170/231*			
Age		-1.601	3.177	0.614	-0.734	1.248	0.557
	N (left/right of cut-off)			*1,409/1,822*			

- *Note*: This table contains rdd estimation results on two invariant endline variables. Variables on the left, statistics on top.
- Sample: Household and member sample (panel, varying N)
- Source: Endline survey (2018)
- Columns (1) to (3) contain results applying a fuzzy design, where insurance status is instrumented via the poverty score. Columns (4) to (6) contain results applying a sharp design, where treatment is strictly determined by poverty score. Columns (1) and (4) display the coefficient estimates, Columns (2) and (5) associated standard errors, and Columns (3) and (6) the p-value of testing the null hypothesis of zero effect size. All estimates are obtained by running local linear regression models to the left and right of the cut-off, using the bias-corrected bandwidth that minimizes the mean squared error, as implemented in the Stata command rdrobust by Calonico et al. (2017). For the variable age, standard errors are clustered on household level.

Table B.4: SPECIFICATION TEST: PSEUDO CUT-OFF SCORES

Variable	cut-off=17.19			cut-off=19.19		
	β_{pseudo}	S.E.	p-val.	β_{pseudo}	S.E.	p-val.
	(1)	(2)	(3)	(4)	(5)	(6)
I. Health care financing						
Fin. Satisfaction $\leq$ 5(scale 1 to 10)	-0.223	0.154	0.147	-0.100	0.183	0.585
Citing health shock as main risk	0.011	0.152	0.938	-0.015	0.177	0.931
Dif.ty finding money for health care $\leq$ 8 (scale 1 to 10)	0.061	0.131	0.641	0.021	0.143	0.881
N (left / right of cut-off)	*211 / 190*			*288 / 113*		
Total cost of last treatment (PKR, win99)	8,038	17,576	0.647	3,576	12,569	0.776
Very worried re. cost of inp. Care	-0.154	0.158	0.330	-0.0470	0.249	0.850
N (left/right of cut-off)	*246/262*			*348/160*		
II. Health care usage						
Case of inpatient care	0.011	0.053	0.833	-0.042	0.060	0.487
N (left/right of cut-off)	*1,785/1,446*			*2,376/855*		
Neglected health care	0.147	0.106	0.163	0.010	0.125	0.934
Case of outpatient care	-0.085	0.148	0.566	-0.066	0.170	0.699
N (left/right of cut-off)	*211/190*			*288/113*		
Use of private hospital	-0.059	0.108	0.580	-0.024	0.128	0.851
More than one admittance	0.165	0.215	0.442	-0.137	0.205	0.502
N (left/right of cut-off)	*246/262*			*348/160*		

- *Note*: This table contains rdd estimation results at two pseudo cut-off points. Variables on the left, statistics on top.
- Sample: Household and member sample (panel, varying N)
- Source: Endline survey (2018)
- Columns (1) to (3) contain results for a cut-off score 17.19. Columns (4) to (6) contain results for a cut-off score at 19.19. Columns (1) and (4) display the coefficient estimates, Columns (2) and (5) associated standard errors, and Columns (3) and (6) the p-value of testing the null hypothesis of zero effect size. All estimates are obtained by running local linear regression models to the left and right of the cut-off, using the bias-corrected bandwidth that minimizes the mean squared error, as implemented in the Stata command rdrobust by Calonico et al. (2017). For the member and conditional sample, standard errors are clustered on household level.

Table B.5: Robustness checks

Variable	Fuzzy design						Sharp design					
	Main specification		With covariates		Unclustered		Main specification		With covariates		Unclustered	
	β_{LATE}	S.E.	$\beta_{LATE,cov}$	S.E.	$\beta_{LATE,uncl.}$	S.E.	β_{ITT}	S.E.	$\beta_{ITT,cov}$	S.E.	$\beta_{ITT,uncl.}$	S.E.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
I. Health care financing												
Fin. Satisfaction $\leq$ 5(scale 1 to 10)	0.320	0.344	0.352	0.343			0.133	0.124	0.149	0.123		
Citing health shock as main risk	0.075	0.298	0.031	0.289			-0.006	0.115	-0.028	0.115		
Dif.ty finding money for health care $\leq$ 8 (scale 1 to 10)	-0.237	0.282	-0.236	0.266			-0.094	0.112	-0.110	0.112		
N (left/right of cut-off)						*170/231*						
Total cost of last treatment (PKR, win99)	32,775	110,000	24,244	84,330	20,634	91,769	-289	12,407	3,568	12,092	-159	11,773
Very worried re. cost of inp. care	-0.150	1.287	-1.181	1.736	-0.281	0.660	-0.033	0.152	-0.149	0.132	-0.0430	0.098
N (left/right of cut-off)						*182/326*						
II. Health care usage												
Case of inpatient care	-0.110	0.106	-0.161	0.106	-0.153*	0.086	-0.053	0.038	-0.075**	0.037	-0.048	0.032
N (left/right of cut-off)						*1,408/1,822*						
Neglected health care	-0.093	0.208	-0.091	0.204			-0.044	0.079	-0.043	0.079		
Case of outpatient care	0.249	0.288	0.179	0.286			0.069	0.123	0.047	0.124		
N (left/right of cut-off)						*170/231*						
Use of private hospital	0.856	1.453	1.051	1.564	0.792	0.793	0.098	0.097	0.137	0.099	0.101	0.073
More than one admittance	1.123	1.673	1.436	1.786	0.895	0.882	0.186	0.134	0.255*	0.140	0.217	0.137
N (left/right of cut-off)						*182/326*						

- *Note*: This table contains rdd estimation results for different model specification as robustness checks. Variables on the left, statistics on top.
- Sample: Household and member sample (panel, varying N)
- Source: Endline survey (2018)
- Columns (1) to (6) contain results for the fuzzy design, where insurance status is instrumented by the poverty score. Columns (7) to (12) contain results for the sharp design.All estimates are obtained by running local linear regression models to the left and right of the cut-off, using the bias-corrected bandwidth that minimizes the mean squared error, as implemented in the Stata command rdrobust by Calonico et al. (2017). The following model specification were used:
  - Columns (1)-(2) and (7)-(8): Estimates and standard errors of main specification with clustering on household level in the member and conditional sample.
  - Columns (3)-(4) and (9)-(10): Estimates and standard errors of main specification but with covariates household size and wealth index (household sample) and additionally gender and age (member and conditional sample).
  - Columne (5)-(6) and (11)-(12): Estimates and standard errors of main specification but without clustering of standard errors.
- Statistical significance is given as follows: *** indicates $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.6: HETEROGENEOUS EFFECTS - FUZZY DESIGN

Variable	Urban		Rural		Near DHQ		Near City Hosp.	
	β_u	S.E.	β_r^M	S.E.	β_{DHQ}	S.E.	β_{CH}	S.E.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
I. Health care financing								
Fin. Satisfaction ≤ 5 (scale 1 to 10)	0.271	0.399	0.206	0.674	-0.026	0.469	0.645	0.965
Citing health shock as main risk	-0.230	0.306	0.481	0.804	-0.286	0.322	0.497	1.001
Dif.ty finding money for health care ≤ 8 (scale 1 to 10)	0.052	0.288	-0.858	0.888	0.185	0.346	0.787	1.163
N (left/right of cut-off)	*107/171*		*63/60*		*61/83*		*46/88*	
II. Health care usage								
Case of inpatient care	-0.098	0.088	-0.214	0.311	-0.060	0.070	-0.154	0.238
N (left/right of cut-off)	*875/1,347*		*534/475*		*520/651*		*355/696*	
Neglected health care	-0.032	0.207	-0.071	0.333	-0.017	0.135	0.649	1.092
Case of outpatient care	0.815	0.405	-0.878	1.117	-0.056	0.346	3.041	4.850
N (left/right of cut-off)	*107/171*		*63/60*		*61/83*		*46/88*	

- *Note*: This table contains heterogeneous effect estimation on main outcome variables. Variables on the left, statistics on top.
- Sample: Subsamples of household and member sample (panel, varying N)
- Source: Endline survey (2018)
- Columns contain the effect estimate and standard errors for the following subsamples:
  - Column (1)-(2): Households/members living in Gilgit Town
  - Column (3)-(4): Households/members living outside of Gilgit Town
  - Column (5)-(6): Households/members in Gilgit Town living closer to the DHQ than to the City Hospital, implying a higher propensity to belong to the Shia sect
  - Column (7)-(8): Households/members in Gilgit Town living closer to the City Hospital than to the DHQ, implying having a higher propensity to belong to the Sunni sect

  All estimates are obtained by running local linear regression models to the left and right of the cut-off, using the bias-corrected bandwidth that minimizes the mean squared error, as implemented in the Stata command rdrobust by Calonico et al. (2017), using the fuzzy design where insurance status is instrumented by the poverty score. For the variable *Case of inpatient care*, standard errors are clustered on household level.
- Statistical significance is given as follows: *** indicates $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Appendix C

Appendix to Chapter 3

1 Additional Information on the FSPs and the program

1-1 Course outline Training of Trainers (ToT)

Day 1:

- Introduction to SIYB program
- Introduction to GYB training package
- SIYB game module 1
- Self-learning GYB training materials

Day 2:

- Presentation skills part 1
- Introduction to SYB training packages
- Principles of adult learning
- Participatory training methods
- Adapting SIYB training for women entrepreneurs
- Self-learning SYB training materials

Day 3:

- Effective use of training tools
- IYB training package
- SIYB game module 2
- Self-learning IYB training materials

Day 4:

- SIYB training cycle: Marketing SIYB program
- Selection and training needs analysis
- SIYB game module 3

Day 5:

- Developing SIYB training session plan
- Presentation skills, part 2
- Manage the logistics of SIYB seminars
- Evaluation of seminar participants

Day 6:

- Entrepreneurial competence
- Generate business ideas
- Choosing a business idea
- SIYB game module 1

Day 7:

- Determining costs of products and services
- Forms of businesses
- Perform follow-up interventions of SIYB
- SIYB game module 2

Day 8:

- Stock control
- Marketing I and II
- Bookkeeping

Day 9:

- Purchasing/buying
- Monitoring and evaluation of SIYB trainings
- ToT SIYB test

Day 10:

- Financial planning I and II
- Partner organization action plan
- Seminar evaluation and closing session

1-2 Course outline Training of Counselors (ToC)

Day 1:

- Introduction: Objectives of the training, agenda
- Pre-test: Understanding characteristics of MSEs
- Business counseling
- Adult learning principles and participatory methods, use of visual aids

Day 2:

- Technique of facilitation
- Role play: Individual consulting
- Business game

Day 3:

- Business improvement
- Positioning of products
- 7 P's marketing mix: Product, price, place, promotion, people, process

Day 4:

- Physical evidence
- Planning your business for future: Financial planning
- Planning process, developing business plans

Day 5:

- Using business plans
- Post test
- Action plan and end of seminar evaluation

1-3 Key characteristics of FSPs

Table C.1: Key characteristics of FSPs

	Batch I (Cooperatives)					
FSP	1	2	3	4	5	6
General						
Established	2007	2007	2008	2001	2004	1959
Branches	8	3	14		24	5
# employees	76	49	226	110	161	174
Total value loans disbursed in last 12 months (mln IDR)	22,386	22	77,107	9,235	138,738	326,607
# active borrowers ('000)	10	3	32	1	19	41
% female borrowers	94	40	100	97	35	100
% clients in rural areas	85	65	100	75	95	75
# active depositors ('000)	10	4	39	2	46	20
Treatment						
# loan officers in ToT	6	4	10	7	7	7
# loan officers in ToC	4	5	4	4	4	3
# client treatment group(s)/ control group	172/ 68	223/ 50	440/ 141	298/ 112	395/ 142	396/ 113
	Batch II (Banks)					
FSP	7	8	9	10	11	12
General						
Established	1978	1952	2002	1961	2006	2015
Branches	21	1	32	41	12	15
# employees	153	250	1,142	4,554	299	237
Total value loans disbursed in last 12 months (mln IDR)	25,291	285	108,467	14,263,823	293,372	147,337
# active borrowers ('000)	9	14	55	290		17 10
% female borrowers		31			43	42
% clients in rural areas	70	70	70	70	70	50
# active depositors ('000)		140	76	5,062	65	43
Treatment						
# loan officers in ToT						
# loan officers in ToC	10	10	10	11	15	10
# client treatment group(s)/ control group	123/ 105	176/ 138	107/ 113	160/ 136	80/ 115	80/ 92

▸ *Note*: This table shows key characteristics of the FSPs. Characteristics on the left, FSP ID on top.
▸ Sample: 12 FSPs.
▸ Source: ILO.

2 Additional baseline statistics and balance

Table C.2: Baseline characteristics of estimation sample by FSP type

	Cooperatives		Rural Banks		Development Banks	
	Mean	Stand. Dev.	Mean	Stand. Dev.	Mean	Stand. Dev.
Panel A: Client Characteristics						
Share of female clients	0.829	0.377	0.459	0.499	0.261	0.440
Average age of client	43.675	9.552	44.484	8.565	45.419	7.878
Total nr. of people in the HH btw 18-65yrs. (incl. the client)^{99p*}	2.805	1.077	2.774	1.005	2.835	1.028
Share of clients having no education	0.038	0.191	0.024	0.153	0.011	0.106
Share of clients having primary school education	0.349	0.477	0.229	0.420	0.148	0.356
Share of clients having secondary education	0.184	0.388	0.222	0.416	0.204	0.403
Share of clients having vocational education	0.295	0.456	0.421	0.494	0.511	0.500
Share of clients having no add. income	0.558	0.497	0.603	0.489	0.677	0.468
Share of clients having add. income from another business	0.285	0.451	0.198	0.399	0.197	0.398
Panel B: MSE Characteristics						
Nr. of yrs. the business exists^{99p*}	11.156	8.733	12.183	9.418	12.701	9.173
Share of MSEs not registered	0.839	0.367	0.846	0.361	0.473	0.500
Share of MSEs selling directly at the market	0.529	0.499	0.548	0.498	0.483	0.500
Share of MSEs selling through agents	0.176	0.381	0.134	0.341	0.227	0.419
Share of MSEs that are part of the HH	0.790	0.407	0.548	0.498	0.434	0.496
Share of clients preparing a business/ financial plan	0.502	0.500	0.209	0.407	0.307	0.461
Share of clients keeping record of all transactions	0.327	0.469	0.345	0.476	0.595	0.491
Share of clients keeping business and HH finances separately	0.561	0.496	0.422	0.494	0.591	0.492
Share of clients investing profit into the business	0.631	0.483	0.668	0.471	0.703	0.457
Total nr. of non-family, permanent workers^{99p*}	0.992	2.280	0.922	2.379	2.468	4.291
Panel C: Loan Behavior						
Average number of loans per year	0.951	0.679	0.631	0.450	0.577	0.336
Share of clients reporting being late with the loan payment at the FSP	0.134	0.340	0.288	0.453	0.157	0.364
Share of clients want to borrow more for business	0.798	0.401	0.566	0.496	0.648	0.478

- *Note*: This table contains selected summary statistics from the baseline survey for the three types of FSPs. The table shows baseline variables on the left, the different samples considered and descriptive statistics on top.
- Sample: Estimation sample, i.e., net of attrition (N=2,550 in cooperatives, 1,129 in rural banks, 296 in development banks).
- Source: Baseline survey (2017 - 2018).
- Columns (1) - (2) display the mean and standard deviation for cooperatives. Columns (3) - (4) display the mean and standard deviation for rural banks. Columns (5) - (6) display the mean and standard deviation for the development bank.
- The superscript *np* indicates the winsorizing level. We winsorized variables per batch prior to randomization and following an automated rule to define percentiles. * indicates that we re-defined the winsorizing level manually.

Table C.3: Balance tests (12 FSPs)

Randomization Variables	Baseline Summary				Estimation Sample [- *without attrited households* -]	
	All				All	Treatment Arms
Variable	(1) Treatment Mean/SE	(2) Control Mean/SE	t-test Difference (1)-(2)	Normalized difference (1)-(2)	t-test Difference (1)-(2)	F-test for joint orthogonality
Share of female clients	0.722	0.604	0.118	0.253	0.122	N/A
	[0.008]	[0.012]				
Average age of client	44.093	44.213	-0.121	-0.013	-0.057	N/A
	[0.168]	[0.225]				
Total nr. of people in the HH btw 18-65yrs. (incl. the client)[99p*]	2.803	2.789	0.014	0.013	0.018	0.029
	[0.019]	[0.025]				
Share of clients having no education	0.032	0.027	0.005	0.029	0.006	1.623
	[0.003]	[0.004]				
Share of clients having primary school education	0.291	0.271	0.020	0.045	0.030	1.004
	[0.008]	[0.011]				
Share of clients having secondary education	0.195	0.195	-0.001	-0.002	0.004	1.540
	[0.007]	[0.010]				
Share of clients having vocational education	0.341	0.386	-0.045	-0.094	-0.060	0.614
	[0.009]	[0.012]				
Share of clients having university education	0.141	0.121	0.021	0.060	0.020	1.760
	[0.006]	[0.008]				
Total nr. of people in the hh that additionally earn income	1.263	1.216	0.048	0.050	0.029	0.311
	[0.017]	[0.023]				
Share of clients having no add. income	0.575	0.567	0.009	0.017	0.006	1.415

Continued on next page

Table C.3 – *Continued from previous page*

Randomization Variables	**Baseline Summary**				**Estimation Sample** [- *without attrited households* -]	
		All			**All**	**Treatment Arms**
	[0.009]	[0.012]				
Share of clients having add. income from another business	0.262	0.252	0.010	0.022	0.015	2.187*
	[0.008]	[0.011]				
Share of clients having add. income from full time job	0.054	0.077	-0.023*	-0.095	-0.023*	1.367
	[0.004]	[0.007]				
Share of clients having add. income from oth. sources	0.041	0.037	0.004	0.019	0.004	0.538
	[0.004]	[0.005]				
Share of clients covering the HH against unforeseen exp.	0.879	0.894	-0.015	-0.046	-0.019	0.519
	[0.006]	[0.008]				
Nr. of yrs. the business exists^{99p*}	11.431	11.468	-0.037	-0.004	0.101	0.520
	[0.162]	[0.228]				
Nr. of yrs. with FSP^{99p*}	4.494	4.407	0.088	0.018	0.055	0.555
	[0.089]	[0.116]				
Share of MSEs not registered	0.826	0.812	0.014	0.036	0.014	0.159
	[0.007]	[0.010]				
Share of MSEs registered as single merchant	0.109	0.114	-0.005	-0.015	-0.006	0.234
	[0.006]	[0.008]				
Share of MSEs selling directly at the market	0.542	0.536	0.006	0.012	0.012	0.935
	[0.009]	[0.012]				
Share of MSEs selling through agents	0.162	0.153	0.009	0.025	0.011	0.186
	[0.007]	[0.009]				
Share of MSEs that are part of the HH	0.711	0.641	0.071	0.152	0.054	0.265
	[0.008]	[0.012]				
Share of MSEs active throughout the year	0.918	0.902	0.015	0.054	0.017	1.160

Continued on next page

Table C.3 – *Continued from previous page*

Randomization Variables	**Baseline Summary**				**Estimation Sample** [- *without attrited households* -]	
	All				**All**	**Treatment Arms**
	[0.005]	[0.007]				
Nr. of family worker/s at the start of the business[99.5p*]	1.108	1.167	-0.060	-0.051	-0.087	1.745
	[0.021]	[0.030]				
Share of MSEs having a written contract with workers	0.031	0.035	-0.005	-0.027	-0.006	0.829
	[0.003]	[0.005]				
Share of MSEs paying workers regularly	0.559	0.542	0.018	0.035	0.022	0.494
	[0.009]	[0.012]				
Share of clients understanding the 4Ps	0.358	0.313	0.046	0.096	0.046	0.098
	[0.009]	[0.012]				
Share of clients stating that marketing is important	0.873	0.819	0.054	0.154	0.051	0.047
	[0.006]	[0.010]				
Share of clients having a customer identification strategy	0.740	0.700	0.040	0.090	0.030	0.848
	[0.008]	[0.011]				
Share of clients having a strategy to make customers like their product	0.812	0.760	0.052	0.129	0.040	2.251*
	[0.007]	[0.011]				
Share of clients having a competition strategy	0.766	0.720	0.046	0.107	0.035	0.451
	[0.008]	[0.011]				
Share of clients preparing a business/ financial plan	0.445	0.369	0.076	0.155	0.078	1.454
	[0.009]	[0.012]				
Share of clients keeping record of all transactions	0.364	0.364	0.000	0.000	-0.003	1.389
	[0.009]	[0.012]				
Share of clients keeping business and HH finances separately	0.532	0.506	0.026	0.051	0.018	1.326
	[0.009]	[0.012]				

Continued on next page

Table C.3 – *Continued from previous page*

Randomization Variables	**Baseline Summary**				**Estimation Sample** [- *without attrited households* -]	
	All				**All**	**Treatment Arms**
Share of clients investing profit into the business	0.646 [0.009]	0.652 [0.012]	-0.006	-0.012	-0.006	0.806
Profit generated in the last 30 days in mln IDR^{99p**}	5.003 [0.124]	5.505 [0.197]	-0.501	-0.069	-0.333	0.840
Revenue in the last 30 days in mln IDR^{99p**}	16.932 [0.538]	19.997 [0.920]	-3.065	-0.094	-2.107	0.100
Cost/day for raw material in mln IDR^{90p**}	1.014 [0.028]	1.032 [0.040]	-0.018	-0.011	0.006	0.177
Cost/day for workers salaries in mln IDR^{90p**}	0.189 [0.006]	0.191 [0.008]	-0.002	-0.006	0.002	0.557
Cost/day for transport in mln IDR^{95p**}	0.036 [0.001]	0.035 [0.002]	0.001	0.022	0.002	0.015
Cost/day for equipment or leasing in ths. IDR^{90p}	0.076 [0.005]	0.072 [0.006]	0.004	0.017	0.006	0.623
Total nr. of non-family, permanent workers^{99p*}	1.079 [0.044]	1.172 [0.065]	-0.093	-0.037	-0.066	0.142
HH exp. on non-durables in the last 30 days in mln IDR^{99p**}	2.557 [0.034]	2.643 [0.048]	-0.087	-0.046	-0.050	1.010
HH exp. on educ. for kids in the last 30 days in mln IDR^{99p*}	0.492 [0.013]	0.499 [0.019]	-0.007	-0.010	-0.012	0.165
HH exp. on food in the last 30 days in mln IDR^{99p*}	1.247 [0.015]	1.270 [0.022]	-0.024	-0.028	-0.005	0.853
Share of clients having positive spending on durables	0.356 [0.009]	0.426 [0.012]	-0.070	-0.143	-0.076*	1.431

Continued on next page

Table C.3 – *Continued from previous page*

Randomization Variables	**Baseline Summary**				**Estimation Sample** [- *without attrited households* -]	
	All				**All**	**Treatment Arms**
Average number of loans per year	0.879	0.767	0.112*	0.182	0.120**	2.209*
	[0.012]	[0.012]				
Last loan amount in mln IDR^{99p**}	16.965	28.286	-11.321*	-0.257	-10.003	0.721
	[0.689]	[1.305]				
Share of clients having their last loan as business/ individual loan	0.644	0.752	-0.108	-0.231	-0.106	0.597
	[0.009]	[0.011]				
Share of clients reporting being late with the loan payment at the FSP	0.163	0.187	-0.025	-0.066	-0.033	0.305
	[0.007]	[0.010]				
Share of clients want to borrow more for business	0.738	0.701	0.036	0.082	0.034	0.810
	[0.008]	[0.011]				
Share of clients stating finding market as the main barrier	0.262	0.219	0.043	0.101	0.044	1.320
	[0.008]	[0.010]				
Share of clients stating expensive raw materials as the main barrier	0.367	0.304	0.063	0.133	0.065	0.674
	[0.009]	[0.011]				
Share of clients stating little experience as the main barrier	0.048	0.039	0.009	0.043	0.009	0.065
	[0.004]	[0.005]				
Share of clients stating tough competition as the main barrier^{NR}	0.280	0.241	0.039**	0.088	0.035	1.113
	[0.008]	[0.011]				
Share of clients not stating any business barrier^{NR}	0.176	0.221	-0.045	-0.115	-0.045	1.403
	[0.007]	[0.010]				

Continued on next page

Table C.3 – *Continued from previous page*

Randomization Variables	**Baseline Summary**				**Estimation Sample** [- *without attrited households* -]	
		All			**All**	**Treatment Arms**
Share of clients stating that business definitely brings high income	0.321	0.310	0.012	0.025	0.011	2.589*
	[0.008]	[0.012]				
Share of clients stating that business gives security	0.559	0.488	0.071	0.143	0.076	0.877
	[0.009]	[0.012]				
Share of clients stating that business is definitely complicated	0.213	0.207	0.006	0.015	0.008	0.122
	[0.007]	[0.010]				
Share of clients stating that business is definitely risky	0.261	0.257	0.005	0.010	0.004	0.234
	[0.008]	[0.011]				
Share of clients stating that business definitely brings respect	0.489	0.414	0.075	0.150	0.079	1.272
	[0.009]	[0.012]				
Share of clients stating that business definitely brings satisfaction	0.592	0.540	0.052	0.105	0.056	0.795
	[0.009]	[0.012]				
Attrition between baseline and endline	0.144	0.176	-0.032*	-0.089	N/A	N/A
	[0.006]	[0.009]				
N	3095	1608				

- *Note*: This table contains the summary and balance statistics on all variables from the baseline survey. The table shows variables on the left, the different samples considered and statistics on top. Variables are those which we used for randomization, in addition to other variables of interest (marked with the superscript *NR*).
- Sample: Baseline sample, i.e. with attrited households (N=4,703), and estimation sample, i.e. net of attrition (N=3,975).
- Source: Baseline survey (2017 - 2018).
- Columns (1)-(2) display the mean and standard error for the treatment and control group for the full baseline sample. Column (3) displays the difference in means across treatment and control group, and Column (4) the normalized difference, both columns refer to the full baseline sample. Columns (5) and (6) consider estimation samples, i.e., net of attrition. Column (5) shows the difference in means for the estimation sample, Column (6) displays the p-value of an F-test for joint significance for the three treatment arms, i.e., the treatment arms regressed on the balance variables.
- All regressions include enumerator- and FSP fixed effects and the covariates variables age and gender. This is in line with the regressions for estimating treatment effects.
- The superscript *np* indicates the winsorizing level. We winsorized variables per batch prior to randomization and following an automated rule to define percentiles. The stars indicate deviation from this rule: * indicates that we defined the winsorizing level manually, ** indicates that we re-ran the winsorizing routine for all batches combined after the randomization.
- We test the null hypothesis of equality of means. The statistical significance is given as follows: * indicates $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$.

3 Additional results

3-1 Heterogeneous effects on final outcomes

Table C.4: Effects on final outcomes by entrepreneur and business characteristics

I. Entrepreneur characteristics	**Part** (1)				**Part** (2)				**Part** (3)				**Part** (4)				**Part** (5)			
Group:	Female				Age>45				More than secondary school education				> 5 hours p.d. spent on hh chores (E)				No-one to replace in business (E)			
	YES	NO	Diff.	Total	YES	NO	Diff.	Total	YES	NO	Diff.	Total	YES	NO	Diff.	Total	YES	NO	Diff.	Total
Panel A: Business Financials and Attitudes																				
Log of revenue in the last 30 days in mln IDR^{99p}	0.039	0.043			0.152	-0.059	**	*	0.042	0.045			0.041	0.046			0.090	-0.018		
Nr. of permanent workers	0.044	-0.542	*		-0.246	-0.057			-0.341	-0.005			0.075	-0.189			-0.095	-0.222		
Log of total cost of business in the last 30 days in mln IDR^{99p}	-0.026	-0.039			0.060	-0.117			-0.050	-0.015			-0.041	0.016			0.015	-0.134		
Log of profit generated in the last 30 days in mln IDR^{99p}	0.004	0.030			0.106	-0.075	*		-0.052	0.064			-0.028	0.052			0.053	-0.047		
Log of profit in the worst month mln IDR^{99p}	0.009	-0.011			0.071	-0.059			-0.048	0.038			-0.065	0.063			0.069	-0.131	*	
Log of profit in the best month in mln IDR^{99p}	0.104	-0.062		*	0.131	-0.029	*	*	0.029	0.063			-0.036	0.097			0.098	-0.078	*	
Share of business earnings covers the exp. of this business	-0.020	0.008			-0.003	-0.018			0.007	-0.024			-0.013	-0.007			-0.016	0.002		
Profit increased during the last 6 months	0.005	0.020			0.011	0.012			-0.005	0.023			0.015	0.007			-0.002	0.029		
Business perception index (standardized score)	0.031	-0.046			0.000	0.005			0.061	-0.043	*		0.030	-0.018			0.018	-0.017		
Nr. of barriers for business	0.018	-0.000			0.023	0.003			0.007	0.016			0.031	0.017			-0.014	0.056		
Panel B: Household Financials and Life Satisfaction																				
Log of total savings from all sources in mln IDR^{90p}	-0.066	-0.076			-0.086	-0.057			-0.162	0.014			0.061	-0.170			-0.152	0.021		
Log of HH exp. on non-durables in the last 30 days in mln IDR^{99p}	-0.004	-0.018			0.006	-0.025			-0.024	0.005			0.032	-0.029			0.020	-0.030		
Share of clients reporting increase in HH exp. in the last 6 months	0.005	0.031			0.020	0.012			0.007	0.021			0.014	0.016			0.021	0.002		
Life satisfaction (1=worst, 10=best)	0.028	-0.116			-0.026	-0.022			-0.214	0.135	***	***	-0.171	0.087	**	**	-0.043	0.040		
Difference between life satisfaction now and two yrs. ago	-0.113	-0.044			-0.089	-0.084			-0.217	0.020	**	***	-0.017	-0.109			-0.109	-0.005		
Panel C: Loan Behaviour																				
Current/last loan used for productive purposes	-0.014	-0.030			-0.008	-0.031			-0.008	-0.029			-0.057	0.002	*	**	-0.040	0.024	*	**
Currently in loan default or behind with repayments for any loan	-0.018	-0.018			-0.035	-0.003		*	-0.023	-0.015			-0.050	0.004	*	**	-0.018	-0.028		
N	2743.000	1232.000			1960.000	2007.000			1792.000	2180.000			1588.000	2248.000			2215.000	1321.000		

II. Business characteristics	**Part** (6)				**Part** (7)				**Part** (8)				**Part** (9)				**Part** (10)			
Group:	> 9 years in business				Above average profit				Above average loan size				Above average marketing index				Preparing business plan			
	YES	NO	Diff.	Total	YES	NO	Diff.	Total	YES	NO	Diff.	Total	YES	NO	Diff.	Total	YES	NO	Diff.	Total
Panel A: Business Financials and Attitudes																				
Log of revenue in the last 30 days in mln IDR^{99p}	0.060	0.024			0.037	0.043			0.061	0.028			0.005	0.099			0.061	0.029		
Nr. of permanent workers	-0.034	-0.285			-0.298	-0.038			-0.307	0.016			-0.194	-0.082			-0.521	0.080		*
Log of total cost of business in the last 30 days in mln IDR^{99p}	-0.011	-0.052			-0.060	-0.021			0.023	-0.089			-0.042	0.000			-0.099	0.012		
Log of profit generated in the last 30 days in mln IDR^{99p}	-0.002	0.029			-0.003	0.022			0.065	-0.037			0.004	0.032			0.082	-0.037		
Log of profit in the worst month mln IDR^{99p}	-0.006	0.004			0.022	-0.033			-0.003	0.007			0.033	-0.033			0.137	-0.094	**	
Log of profit in the best month in mln IDR^{99p}	0.038	0.057			0.036	0.053			0.061	0.034			0.094	-0.008			0.125	-0.007		
Share of business earnings covers the exp. of this business	-0.012	-0.009			-0.014	-0.008			0.003	-0.025			0.004	-0.029			-0.004	-0.014		
Profit increased during the last 6 months	0.007	0.014			0.021	-0.002			0.026	-0.007			0.001	0.024			0.017	0.006		
Business perception index (standardized score)	-0.096	0.102	***	**	-0.049	0.059	*		-0.007	0.018			-0.010	0.023			-0.000	0.008		
Nr. of barriers for business	0.020	0.003			-0.014	0.033			0.025	-0.004			-0.026	0.065	*		-0.008	0.024		
Panel B: Household Financials and Life Satisfaction																				
Log of total savings from all sources in mln IDR^{90p}	-0.182	0.037			-0.113	-0.036			-0.044	-0.088			-0.078	-0.042			-0.161	-0.010		
Log of HH exp. on non-durables in the last 30 days in mln IDR^{99p}	-0.009	-0.009			-0.018	-0.004			-0.019	0.004			-0.058	0.061	**		-0.069	0.028		
Share of clients reporting increase in HH exp. in the last 6 months	0.009	0.019			0.007	0.022			0.023	0.005			0.026	-0.001			0.005	0.021		
Life satisfaction (1=worst, 10=best)	0.031	-0.076			-0.074	0.024			-0.000	-0.047			-0.082	0.060			-0.097	0.022		
Difference between life satisfaction now and two yrs. ago	-0.063	-0.115			-0.067	-0.111			-0.059	-0.123			-0.162	0.010		**	-0.194	-0.027		**
Panel C: Loan Behaviour																				
Current/last loan used for productive purposes	0.005	-0.045			-0.003	-0.039			-0.013	-0.029			-0.016	-0.025			-0.008	-0.028		
Currently in loan default or behind with repayments for any loan	-0.023	-0.013			-0.035	-0.000			-0.015	-0.023			-0.003	-0.040			-0.041	-0.005		*
N	1981.000	1994.000			1987.000	1988.000			1987.000	1988.000			2350.000	1625.000			1617.000	2358.000		

- *Note*: This table shows final outcome variables on the left. **Parts** (1)-(10) on the top indicate heterogeneous effects for different groups.
- Sample: Estimation sample (N = 3,975).
- Source: Endline survey (2018 - 2019).
- Dark grey columns show significance level of the null hypothesis which assumes the effect to be equal across the two respective groups. Light grey columns show significance level of the null hypothesis of zero effect in the respective YES-group.
- Part (4): Indicator for more than 5 hours per day spend on household chores (measured only at endline); Part (5): Indicator for respondent claiming that no-one can replace her in the business (measured only at endline); Part (9): The marketing index is the sum of dummy variables indicating whether the client has (i) a customer identification strategy, (ii) a strategy to make customer like their products, (iii) a competition strategy, (iv) any marketing practices in use; Part (10): Indicator for preparing a business plan at baseline.
- In all regressions we control for the twenty most imbalanced baseline covariates and dummies indicating imputation in baseline variables, enumerator- and FSP fixed effects, and, if applicable, for age and gender.
- The superscript *np*. indicates the winsorizing level.
- The statistical significance is given as follows: * p < 0.1, ** p < 0.05, *** p < 0.01, **** indicates p < 0.001.

Table C.5: Effects on final outcomes by treatment arm (cooperatives only)

Outcome	N	Control Mean (1)	Control SD (2)	Training β^{T}_{ITT} (3)	Training SE (4)		Counseling β^{C}_{ITT} (5)	Counseling SE (6)		Training&Counseling β^{TC}_{ITT} (7)	Training&Counseling SE (8)		P-value for H_0: $\beta^{T}_{ITT} = \beta^{C}_{ITT}$ (9)		$\beta^{T}_{ITT} = \beta^{TC}_{ITT}$ (10)	$\beta^{C}_{ITT} = \beta^{TC}_{ITT}$ (11)
Panel A: Business Financials and Attitudes																
Log of revenue in the last 30 days in mln IDR^{99p}	1748	1.512	1.295	**0.003**	0.080		**0.038**	0.079		**0.001**	0.079		0.656		0.974	0.621
Nr. of permanent workers	2377	1.504	3.383	**0.144**	0.265		**-0.073**	0.197		**-0.219**	0.166		0.433		0.176	0.429
Log of total cost of business in the last 30 days in mln IDR^{99p}	2167	0.995	1.703	**-0.095**	0.097		**-0.015**	0.093		**-0.085**	0.093		0.388		0.917	0.431
Log of profit generated in the last 30 days in mln IDR^{99p}	1863	0.797	1.311	**-0.067**	0.083		**0.060**	0.075		**0.021**	0.078		0.091	*	0.253	0.576
Log of profit in the worst month mln IDR^{99p}	1865	0.273	1.419	**-0.082**	0.099		**0.036**	0.087		**-0.057**	0.085		0.206		0.788	0.257
Log of profit in the best month in mln IDR^{99p}	1984	1.096	1.541	**0.144**	0.082	*	**0.076**	0.086		**0.089**	0.085		0.342		0.430	0.866
Share of business earnings covers the exp. of this business	2358	0.917	0.276	**-0.019**	0.017		**-0.013**	0.016		**-0.029**	0.017	*	0.727		0.567	0.340
Profit increased during the last 6 months	2345	0.349	0.477	**-0.022**	0.027		**0.022**	0.027		**-0.005**	0.027		0.103		0.533	0.302
Business perception index (standardized score)	2526	-0.096	1.026	**0.046**	0.053		**0.044**	0.053		**0.024**	0.055		0.970		0.671	0.697
Nr. of barriers for business	2550	1.112	0.800	**-0.028**	0.043		**-0.028**	0.040		**-0.014**	0.041		0.997		0.743	0.736
Panel B: Household Financials and Life Satisfaction																
Log of total savings from all sources in mln IDR^{90p}	971	0.818	1.735	**0.040**	0.147		**-0.097**	0.145		**-0.006**	0.154		0.348		0.767	0.551
Log of HH exp. on non-durables in the last 30 days in mln IDR^{99p}	2524	0.504	0.956	**0.027**	0.050		**-0.086**	0.051	*	**-0.023**	0.050		0.020	**	0.302	0.207
Share of clients reporting increase in HH exp. in the last 6 months	2520	0.454	0.498	**0.026**	0.028		**-0.011**	0.027		**0.011**	0.027		0.163		0.573	0.403
Life satisfaction (1=worst, 10=best)	2480	7.550	1.762	**-0.024**	0.095		**-0.010**	0.094		**0.032**	0.093		0.881		0.532	0.631
Difference between life satisfaction now and two yrs. ago	2462	0.640	1.669	**0.043**	0.099		**-0.136**	0.101		**-0.096**	0.100		0.069	*	0.156	0.684
Panel C: Loan Behaviour																
Current/last loan used for productive purposes	2000	0.725	0.447	**-0.007**	0.027		**-0.011**	0.027		**0.008**	0.027		0.893		0.585	0.495
Currently in loan default or behind with repayments for any loan	1632	0.116	0.321	**-0.006**	0.020		**0.006**	0.020		**0.025**	0.021		0.568		0.131	0.329

- *Note*: This table shows final outcome variables on the left, treatment arms and test statistics on top. Same clients are followed over time.
- Sample: Batch I (cooperatives) clients (N = 2,550).
- Source: Endline survey (2018 - 2019).
- Columns: (1)-(2) control group (N = 626). (3)-(4) training only arm (N = 632). (5)-(6) counseling only arm (N = 651). (7)-(8) training and counseling arm (N = 641). Columns (9) to (11) show the p-values of testing equality of β_{ITT} for two treatment arms respectively.
- In all regressions we control for age, gender, twenty most imbalanced baseline covariates, dummies indicating imputation in baseline variables, enumerator- and FSP fixed effects, and present robust standard errors.
- The superscript *np.* indicates the winsorizing level.
- The statistical significance is given as follows: * indicates p < 0.1, ** p < 0.05, *** p < 0.01, **** p < 0.001.

Table C.6: Effects on final outcomes by type of FSP (counseling only treatment)

Outcome	Control (All) Mean (1)	SD (2)	Cooperatives (C only) N	$\beta^{Coop.}_{ITT}$ **(3)**	SE (4)		Rural Banks N	β^{Rural}_{ITT} **(5)**	SE (6)		Dev. Banks N	$\beta^{Dev.}_{ITT}$ **(7)**	SE (8)	
Panel A: Business Financials and Attitudes														
Log of revenue in the last 30 days in mln IDR^{99p}	1.512	1.295	885	**0.019**	0.081		757	**0.036**	0.086		196	**0.428**	0.314	
Nr. of permanent workers	1.504	3.383	1193	**-0.031**	0.193		962	**-0.441**	0.285		287	**-0.339**	0.670	
Log of total cost of business in the last 30 days in mln IDR^{99p}	0.995	1.703	1085	**-0.014**	0.094		891	**0.010**	0.111		269	**0.035**	0.202	
Log of profit generated in the last 30 days in mln IDR^{99p}	0.797	1.311	926	**0.052**	0.075		770	**0.063**	0.085		177	**0.002**	0.286	
Log of profit in the worst month mln IDR^{99p}	0.273	1.419	918	**0.057**	0.086		765	**0.034**	0.077		173	**0.298**	0.211	
Log of profit in the best month in mln IDR^{99p}	1.096	1.541	982	**0.079**	0.085		818	**0.010**	0.072		195	**-0.007**	0.158	
Share of business earnings covers the exp. of this business	0.917	0.276	1184	**-0.010**	0.017		953	**0.007**	0.019		281	**-0.007**	0.030	
Profit increased during the last 6 months	0.349	0.477	1177	**0.025**	0.028		950	**0.049**	0.030		284	**-0.028**	0.054	
Business perception index (standardized score)	-0.096	1.026	1262	**0.050**	0.055		1112	**-0.051**	0.053		295	**0.038**	0.133	
Nr. of barriers for business	1.112	0.800	1277	**-0.027**	0.041		1129	**0.060**	0.049		296	**-0.018**	0.098	
Panel B: Household Financials and Life Satisfaction														
Log of total savings from all sources in mln IDR^{90p}	0.818	1.735	495	**-0.062**	0.156		436	**-0.039**	0.158		82	**-0.641**	0.535	
Log of HH exp. on non-durables in the last 30 days in mln IDR^{99p}	0.504	0.956	1265	**-0.090**	0.052	*	1125	**0.042**	0.048		293	**-0.026**	0.107	
Share of clients reporting increase in HH exp. in the last 6 months	0.454	0.498	1262	**-0.010**	0.028		1113	**0.018**	0.030		292	**0.048**	0.060	
Life satisfaction (1=worst, 10=best)	7.550	1.762	1237	**-0.052**	0.097		1110	**-0.071**	0.097		287	**0.213**	0.231	
Difference between life satisfaction now and two yrs. ago	0.640	1.669	1227	**-0.151**	0.105		1104	**-0.223**	0.105	**	286	**0.325**	0.187	*
Panel C: Loan Behaviour														
Current/last loan used for productive purposes	0.725	0.447	996	**-0.024**	0.028		1025	**-0.048**	0.028	*	278	**0.033**	0.057	
Currently in loan default or behind with repayments for any loan	0.116	0.321	836	**0.014**	0.020		842	**-0.069**	0.031	**	260	**-0.003**	0.050	

- *Note*: This table shows final outcome variables on the left, the different samples on top. The same clients are followed over time.
- Sample: Estimation sample, counseling only treatment and control groups (N=2,692).
- Source: Endline survey (2018 - 2019).
- Columns (1)-(2) display the control group mean and standard deviation. In Columns (3)-(8) we present regression results. In all regressions we control for age, gender, the twenty most imbalanced baseline covariates, dummies indicating imputation in baseline variables, enumerator and FSP fixed effects, and present robust standard errors. In Columns (3)-(4) we present results for cooperatives, considering only the counseling treatment and control group. in Columns (5)-(6) we present results for rural banks, in Columns (7)-(8) for development banks.
- The superscript $np.$ indicates the winsorizing level.
- The statistical significance is given as follows: * indicates p < 0.1, ** p < 0.05, *** p < 0.01, **** p < 0.001.

Table C.7: Effects on final outcomes by FSP (counseling only treatment)

Outcome	FSP 1 β_{ITT}	FSP 2 γ^2_{ia}	FSP 3 γ^3_{ia}	FSP 4 γ^4_{ia}	FSP 5 γ^5_{ia}	FSP 6 γ^6_{ia}	FSP 7 γ^7_{ia}	FSP 8 γ^8_{ia}	FSP 9 γ^9_{ia}	FSP 10 γ^{10}_{ia}	FSP 11 γ^{11}_{ia}	FSP 12 γ^{12}_{ia}
Panel A: Business Financials and Attitudes												
Log of revenue in the last 30 days in mln IDR^{99p}	0.044	-0.379 *	-0.108	0.112	0.096	0.100	-0.003	-0.234	0.133	0.308	0.085	0.105
Nr. of permanent workers	-0.023	-0.478	0.083	0.166	0.567	-0.420	-0.021	-0.953	-0.123	-0.329	-0.280	-0.309
Log of total cost of business in the last 30 days in mln IDR^{99p}	0.077	-0.257	-0.089	-0.046	0.056	-0.065	-0.189	-0.142	0.235	-0.043	-0.017	-0.054
Log of profit generated in the last 30 days in mln IDR^{99p}	0.067	-0.416 **	-0.038	-0.205	0.205	0.190	0.162	-0.187	-0.149	-0.020	-0.079	0.124
Log of profit in the worst month mln IDR^{99p}	0.101	-0.233	-0.075	-0.103	0.277	-0.218	0.012	-0.186	-0.147	0.004	-0.168	0.141
Log of profit in the best month in mln IDR^{99p}	-0.036	-0.254	0.197	0.001	0.211	0.261	0.090	-0.142	0.019	0.071	-0.082	0.237
Share of business earnings covers the exp. of this business	0.008	-0.001	-0.045	-0.002	-0.009	-0.004	0.003	-0.018	-0.004	-0.009	0.012	0.002
Profit increased during the last 6 months	0.035	-0.000	-0.005	0.008	-0.033	-0.009	-0.114 *	0.040	0.128 **	-0.064	-0.062	-0.011
Business perception index (standardized score)	0.007	-0.020	0.055	0.229 *	-0.094	0.110	-0.103	-0.095	-0.266 **	-0.024	0.173	-0.020
Nr. of barriers for business	-0.007	0.013	-0.005	0.053	-0.082	-0.024	0.324 ***	-0.051	-0.029	-0.031	0.158	0.041
Panel B: Household Financials and Life Satisfaction												
Log of total savings from all sources in mln IDR^{90p}	-0.074	-0.437	0.114	0.158	-0.128	-0.202	0.011	0.564	0.262	0.017	-0.279	-0.161
Log of HH exp. on non-durables in the last 30 days in mln IDR^{99p}	-0.084 **	0.042	0.139 *	0.115	-0.193 *	-0.073	0.130	0.001	0.254 *	0.047	0.183 *	0.042
Share of clients reporting increase in HH exp. in the last 6 months	-0.023	0.121 *	-0.037	0.016	0.000	-0.014	0.039	0.061	-0.008	0.071	-0.053	0.115
Life satisfaction (1=worst, 10=best)	-0.005	-0.617 **	0.024	-0.140	0.181	-0.040	-0.416 **	0.168	0.449 **	0.135	-0.652 **	0.059
Difference between life satisfaction now and two yrs. ago	-0.098	-0.378	0.166	-0.142	-0.220	0.045	-0.214	0.133	0.268	0.395 **	-0.969 ***	0.120
Panel C: Loan Behaviour												
Current/last loan used for productive purposes	-0.010	-0.056	-0.028	0.026	-0.028	0.018	-0.026	-0.018	0.006	0.034	-0.039	-0.165 **
Currently in loan default or behind with repayments for any loan	0.003	-0.138	-0.002	0.071	-0.056	0.030	-0.027	-0.121 **	0.004	-0.004	-0.009	-0.166 *

- *Note*: This table shows final outcome variables on the left, the different FSPs on top. The same clients are followed over time.
- Sample: Estimation sample, counseling only treatment and control groups (N=2,692).
- Source: Endline survey (2018 - 2019).
- The columns contain the point estimate of the interaction of the treatment indicator with the FSP identifier for the twelve FSPs. In all regressions we control for age, gender, twenty most imbalanced baseline covariates, dummies indicating imputation in baseline variables, enumerator and FSP fixed effects, and present robust standard errors.
- The superscript *np.* indicates the winsorizing level.
- The statistical significance is given as follows: * p < 0.1, ** p < 0.05, *** p < 0.01, **** indicates p < 0.001.

3-2 Robustness checks for average treatment effects

As in any empirical study, bias and imprecision might arise during the processes of data collection (which comprises the steps defining the constructs underlying outcomes, determining the sample composition and randomization, as well as taking measurement) as well as during the process of data analysis (which comprises the steps cleaning the data, selecting the econometric model, estimation and inference). Hence, our robustness checks address all stages as follows:

Definition of Constructs. For profits, we included two different constructs in our questionnaire: firstly, we measure profit directly, and secondly, as the difference between revenues and costs. Cross-checking these different measurements for profit helps judging their validity. However, a training in financial planning might have changed the reporting of these financial outcomes without changing the underlying true values. Therefore, we also tested whether treatment had an impact on the reporting bias, which we define as the difference of the two constructs as percent of the direct profit measurement.

Sample Composition and Randomization. In our main specification we include observations for which some baseline covariates were missing by imputing the corresponding value. As a robustness check we ran regressions on the subsample of observations for which no randomization variable was missing. This excludes 10.3% of our sample. Note that in our endline survey we allowed for baseline respondents to be replaced by other household members if these share business responsibilities, including in financial decision making. In the PAP, we suggested to constrain our analysis to the subsample of panel individuals as robustness check. Since only 2.03% of baseline respondents were replaced by other household members, we refrained from doing so. Instead, 12.34% of our sample reported having changed the FSP between baseline and endline survey. As a robustness check, we restricted our sample to those clients who did not switch FSPs.

Taking Measurements. We included enumerator fixed effects in our main specification to account for differences in asking questions and reporting answers. We checked robustness against excluding these. As a further robustness test, we excluded all observations from enumerators who reported at least one outlier in more than 40% of their observations. This excludes five enumerators and 7.3% of observations from our sample. Furthermore, we ran all analyses restricted to the subset of clients without obvious reporting bias. More precisely, we excluded clients who reported higher profits than revenues, or who reported higher profits in the worst month than in the best month. This excludes 14.64% of our sample.

Cleaning. For key quantitative variables, such as revenues, cost and profits, we conducted the same analyses at different levels of winsorizing: non-winsorized, 1%-winsorized and 5%-winsorized (as in Bruhn et al. 2013). We also ran regressions excluding all clients with any outlier (18.89% of the sample).[1] Since we deviated from the PAP by using the log value of quantitative outcomes, we also estimated results for quantitative outcomes without this transformation.

Econometric Models. Our estimation is based on the econometric model described in section 4. In particular, our main model uses the twenty most imbalanced baseline variables as covariates. We also ran two regressions using the ten most imbalanced, and the thirty most imbalanced variables as covariates respectively. Furthermore, we estimated two models with an additional control variable: First, we controlled for the distance of the client and the FSP using GPS data. Second, we included a dummy variable which takes value one for branches/cashpoints which did not report implementation in the monitoring system.

Inference. In our main specification, we present heteroscedasticity-robust, unclustered standard errors. As a robustness check, we cluster standard errors at the branch/cashpoint level. We also estimated pseudo-effects, using variables as outcomes which are plausibly unaffected by treatment, namely age, having completed primary education, marital status, number of adults in the household and number of people in the household who additionally earn an income.

Results. Table C.8 contains results from the main robustness checks on selected variables. Further results are available upon request. The individual estimates differ slightly across the different specifications as would be expected. The important thing to note is that there are hardly any changes in the significance of the estimates however. Whereas some models yield additional weak significance for some outcomes, none of these are robust against changes in the models. Overall, the robustness checks confirms the results from the main specification presented above.

[1] The winsorizing we use in the main specification identifies 14.24% of our observations to contain exactly one outlier, 4.23% to contain exactly two outliers, and 0.43% to contain more than two outliers.

Table C.8: Robustness checks

H	Main	Alternative Specifications										
	(0)	(1) No miss.	(2) Same FSP	(3) -En.FE	(4) No enum. outlier	(5) No incons. profit	(6) No outlier	(7) 10C	(8) 30C	(9) +Distance	(10) +No Impl. Cl	(11) Cl(CashP)
Aware of support	0.065 ****	0.063 ****	0.069 ****	0.064 ****	0.087 ****	0.066 ****	0.085 ****	0.065 ****	0.065 ****	0.065 ****	0.066 ****	0.065 ****
Offered support	0.070 ***	0.066 ***	0.059 **	0.091 ****	0.121 ****	0.069 ***	0.082 ***	0.072 ***	0.072 ****	0.071 ***	0.070 ***	0.070 ****
Participated	0.141 ****	0.144 ****	0.142 ****	0.144 ****	0.212 ****	0.137 ****	0.133 ***	0.141 ****	0.142 ****	0.140 ****	0.143 ****	0.141 ****
Marketing knowl.	0.002	-0.004	0.003	0.008	-0.018	0.001	-0.008	0.003	0.001	0.002	0.002	0.002
Fin.mgmt. knowl.	0.000	-0.001	-0.001	0.008	0.015	-0.002	0.006	0.002	0.001	0.001	0.000	0.000
Marketing index	0.018	-0.017	0.006	0.022	0.082	0.029	-0.005	0.022	0.016	0.018	0.017	0.018
Fin mgmt. index	-0.003	-0.005	0.003	0.045	0.010	-0.011	-0.084	0.015	-0.004	0.000	-0.003	-0.003
Business plan	0.022	0.026	0.030 *	0.025	0.001	0.016	0.013	0.025	0.023	0.023	0.022	0.022
Cash flow	0.029 **	0.031 **	0.024 *	0.004	0.026	0.025 *	0.043 **	0.030 **	0.029 **	0.028 **	0.029 **	0.029 *
Log revenue	0.043	0.048	0.034	0.047	0.043	0.020	0.004	0.037	0.051	0.048	0.043	0.043
Log profit	0.014	0.035	0.024	0.028	0.025	0.009	-0.027	0.006	0.016	0.015	0.012	0.014
Worst month	0.001	0.009	0.013	0.020	-0.005	-0.006	-0.036	-0.007	-0.000	0.002	0.000	0.001
Best month	0.048	0.076	0.084	0.070	0.082	0.026	0.011	0.040	0.055	0.048	0.048	0.048
Earn. cover exp.	-0.010	-0.014	-0.006	-0.010	-0.019	-0.013	-0.022 *	-0.009	-0.010	-0.011	-0.011	-0.010
Profit inc.	0.010	0.001	-0.004	0.010	0.013	0.003	0.002	0.010	0.011	0.011	0.011	0.010
Bus. perc. index	0.005	0.012	-0.002	0.012	-0.012	0.009	0.009	0.009	0.005	0.006	0.006	0.005
Log savings	-0.072	-0.025	-0.076	0.022	-0.087	-0.060	-0.147	-0.066	-0.059	-0.072	-0.072	-0.072
Loan default	-0.018	-0.012	-0.024	-0.026 *	-0.025	-0.020	-0.030	-0.018	-0.019	-0.018	-0.018	-0.018

- *Note*: The Table contains results from below mentioned robustness checks. Different models on top, outcomes on the left. Results for further outcomes available upon request.
- Sample: All FSPs. Sample size varying.
- Column (0) includes the main specification.
  Column (1) relative to main specification in column (0): Excludes observations with missing values at baseline.
  Column (2) relative to main specification in column (0): Excludes clients who changed their FSP since baseline.
  Column (3) relative to main specification in column (0): Removes enumerator fixed effects.
  Column (4) relative to main specification in column (0): Excludes observations collected by enumerators who entered outlier values in more than 40% of observations.
  Column (5) relative to main specification in column (0): Excludes observations where profit is larger than revenue, and the best-month profit is larger than the worst-month profit.
  Column (6) relative to main specification in column (0): Excludes observations with at least one outlier in quantitative variables.
  Column (7) relative to main specification in column (0): Adds ten most imbalanced control variables instead of twenty.
  Column (8) relative to main specification in column (0): Adds thirty most imbalanced control variables instead of twenty.
  Column (9) relative to main specification in column (0): Additionally controls for the distance to the FSP.
  Column (10) relative to main specification in column (0): Additionally adds a dummy for zero implementation in a cluster.
  Column (11) relative to main specification in column (0): Clusters the standard errors at the cash point.
- The statistical significance is given as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, **** indicates $p < 0.001$.

Curriculum Vitae, Simona Helmsmüller

since 2019	Lecturer at University of Applied Sciences Bonn-Rhein-Sieg, Germany.
2019 – 2020	Project Director Agriculture and Finance Consultants (AFC) International GmbH, Bonn, Germany.
2013 – 2020	PhD Studies in Economics, Center for Doctoral Studies in Economics, University of Mannheim, Germany.
2010 – 2016	Consultant and Project Manager, AFC International GmbH, Bonn, Germany.
2006 – 2012	B.Sc. in Economics, Distance University Hagen, Germany.
2002 – 2009	Diplom in Mathematics, Saarland University, Germany.
2006 – 2007	Visiting student, Université de Montréal, Canada.
2002	Abitur, Gymnasium Syke, Germany.

Eidesstattliche Erklärung

Hiermit erkläre ich, die vorliegende Dissertation selbstständig angefertigt und mich keiner anderen als der in ihr angegebenen Hilfsmittel bedient zu haben. Insbesondere sind sämtliche Zitate aus anderen Quellen als solche gekennzeichnet und mit Quellenangaben versehen.

Mannheim, October 12, 2020:

Simona Helmsmüller