Pre-Analysis Plan

Accountability Can Transform (ACT) Health: A Replication and Extension of Björkman and Svensson (2009)

Doug Parkerson*, Daniel Posner† and Pia Raffler‡

June 14, 2016

Registered before PIs have access to any outcome data.

Abstract

We conduct a scaled-up replication and extension of one of the most influential studies of social accountability, the "Power to the People" (P2P) intervention evaluated by Björkman and Svensson (2009). The P2P intervention aimed to improve local health care provision in rural Uganda by empowering community members to better monitor and sanction underperforming health care providers. Despite its limited power P2P generated striking results, including a 33% decline in under-5 mortality. Given these extremely large effects, the P2P study has received broad acclaim. The objective of this study is twofold. The first goal is to test whether P2P replicates. The second goal is to understand which part of the complex P2P intervention may have been responsible for its strong impact through a crosscutting design.

*Innovations for Poverty Action, dparkerson@poverty-action.org

†UC Los Angeles, dposner@polisci.ucla.edu

‡Yale University, pia.raffler@yale.edu

Contents

1 Introduction

This project constitutes a scaled up replication and extension of the "Power to the People" intervention (henceforth P2P) evaluated by Martina Björkman and Jakob Svensson (Björkman and Svensson, 2009). The P2P intervention aimed to improve local health care provision in rural Uganda by empowering community members to better monitor and sanction underperforming health care providers. The intervention sought to accomplish this by providing community members and providers with information about the quality of health services being provided at their local health center (HC) in a citizen report card (CRC), mobilizing them in light of this information, and then bringing community members and providers together to discuss how they might collaborate to improve health outcomes in their community. Despite its limited power – the study included just fifty health centers, with half receiving the intervention and half serving as control units – P2P generated striking results: infant weights increased in treatment communities; under-5 mortality declined by 33%; immunization rates rose; waiting times at clinics fell; staff absenteeism dropped; and communities became more engaged and monitored clinics more extensively. Given these extremely large effects, the P2P study has received broad acclaim. It has been held up as an example of the power of information to generate accountability (Björkman Nyqvist, De Walque and Svensson, 2014) and of the utility of community-based monitoring as a tool for improving health outcomes in developing country settings. There has accordingly been strong interest in testing whether the P2P findings replicate at a larger scale.

The objective of this study is twofold.[1] The first goal is to test whether P2P replicates. The

[1] In Clemens (2015) terms, our study is actually not a "replication" but a "robustness test and extension" since we are working in a different population and with a different sample. However the motivation for the study is to "replicate" the P2P study in the usually understood sense of testing whether we can reproduce the findings when we re-run the same intervention.

second goal is to understand which part of the complex P2P intervention may have been responsible for its strong impact. To do the latter, we pursue two strategies. First, we break the P2P program into two of its principal components: 1) the provision of information about health delivery quality to community members and clinic staff and the mobilization of these actors in light of this information, and 2) the holding of interface meetings between community members and clinic staff in which the action plans they had each developed could be discussed and coordinated. We then employ a facto-rial design in which we randomize which health facilities (and associated catchment areas) received which parts of the treatment (or, in some cases, both together, as in the original P2P program). Second, we augment the household and health center survey instruments used in P2P with additional questions aimed at uncovering the mechanisms at work. Apart from these small departures, the ACT Health program, implemented by GOAL Uganda, mirrors the P2P intervention strategy extremely closely. This includes questionnaire design, the format and presentation of the report cards that summarize information about health facility-level performance, the organization of the community, health center staff, and interface meetings, the creation of joint action plans, and the holding of 6-month follow-up meetings in all treated communities. The biggest difference between the P2P and ACT Health studies lies not in the contents of the intervention but in the much larger sample we utilize (379 health centers from sixteen districts, rather than 50 health centers from nine districts) and in the fact that our study was carried out ten years later.[2]

2 Theory

The theory of change underlying the original P2P intervention is that health service delivery can be improved by empowering community members to monitor service providers and hold them accountable for poor performance. To understand how P2P does this, it is useful to break it into its three key elements:

1. **Information:** The creation of CRCs, which are shared with health care providers and communities.

2. **Mobilization:** The development of action plans, separately by health care providers and by community members in light of this information.

3. **Interface:** The organization of meetings that bring together members of the community with health care providers to discuss their respective action plans and generate a joint social contract to guide their future behavior and interactions.

[2] A fuller comparison of the P2P and ACT Health interventions and evaluations is provided in Appendix A.

These three components are hypothesized to generate greater monitoring and accountability via the following mechanisms:

- The receipt of information by both community members and health providers, via the CRC, will increase their knowledge about issues related to health care, such as patients' right and responsibilities and (for community members) about the services that are supposed to be offered at the local health center.

- The holding of community meetings to mobilize community members and the development of action plans in light of the information provided in the CRCs will:

  – generate a stronger sense of efficacy among community members.

  – develop a sense of responsibility for making sure that health workers provide high quality services.

  – create stronger perceptions that oversight is possible.

- The intervention as a whole, and the interface meeting in particular, will improve the relationship between community members and health providers. While an improved relationship is not strictly-speaking necessary for improved service performance – an adversarial relationship built around more intensive monitoring and sanctioning might generate higher productivity by HC staff and improved health outcomes without any improvements in the relationship between providers and community members. But insofar as a goal of P2P is to create a social contract between providers and community members – indeed, the drafting of such a social contract is, in fact, part of the intervention – it is reasonable to think that the intervention will improve community-provider relations.

Although not articulated explicitly in Björkman and Svensson (2009), the approach adopted in P2P is rooted in the logic of the principal-agent problem, which highlights the difficulty that citizens (the principals) face in trying to influence the behavior of the government workers (the agents) who are responsible for providing them with public services. As explicated in the classic theoretical treatments of Ross (1973), Arrow (1974), and Hölmstrom (1979), and more recently summarized in Besley (2007), the crux of the principal-agent problem lies in two inherent characteristics of the relationship between any actor (principal) and the agent to whom he has delegated responsibility for completing a task. The first is that the principal cannot directly observe the actions of the agent – for example, how hard he works, whether he has been wasteful with resources, etc. Since the whole motivation for P2P is to bring greater efficiency and effort out of health care workers (Björkman and Svensson, 2009), this is a critical obstacle in the context we study. The implication is that the principal must make an inference about the agent's actions from the outcome

(whether the task is completed on time and with what quality) that the principal observes. The problem – and this is the second inherent characteristic of their relationship – is that the outcome the principal observes is noisy. This matters because it undermines the principal's ability to make clear inferences about the agent's actions. These twin facts – the unobservableness of the agent's actions and the noisiness of outcomes – make it extremely difficult for the principal to hold the agent accountable.

The role played by (at least two of) the main components of P2P can be understood through this framework. The provision of information about health care outcomes contained in the CRC is designed to improve accountability by reducing the noisiness of the signal available to the principal. Indeed, this is the aspect highlighted in most informational campaigns designed to improve accountability (e.g., Banerjee et al. (2010); Humphreys and Weinstein (2012); Andrabi, Das and Khwaja (2014); Chong et al. (2015)). But information about health outcomes does nothing to solve the problem of the health workers' effort being unobservable. So if outcomes are found to be deficient, principals will have no way of knowing whether the poor performance stems from low effort by the HC staff or, as the health workers will certainly claim, circumstances outside of their control. The critical role of the interface meeting is to provide community members with the opportunity to confront HC staff directly and thus put themselves in a position to decide whether or not it is appropriate to hold the staff accountable for the outcomes they observe. Thus, while the information component of the P2P intervention speaks to the "noisiness" component of the principal-agent problem, the interface component speaks to the "unobservableness" component.[3]

3 Context

Public health services in Uganda are provided in a hierarchically organized system with National Referral Hospitals at the national level, Regional Referral Hospitals at the regional level, general hospitals and HCIVs at the district level, and smaller scale health centers at the sub-county and parish levels – the former termed HCIIIs and the latter HCIIs. Our study focuses on health care delivery at the HCIII and HCII levels, the lowest levels of the public health system.[4] HCIIIs, which are staffed by a clinical officer (a trained medical worker) and one or more nurses and lab technicians, provide preventative and out-patient care and have laboratory services to undertake basic tests. They also generally have maternity wards. HCIIs provide outpatient services and antenatal care. They are run by an enrolled nurse, sometimes working with a midwife and a nursing assistant. Both types of units are supported by a Village Health Team (VHT) comprised

[3] As Björkman and Svensson (2009) rightly point out, the mobilization aspect of the P2P intervention accomplishes a third important task – though one not directly connected to the principal-agent problem: it helps the community overcome the free riding problem inherent in community monitoring.

[4] As discussed below, the P2P study only included units at the HCIII level.

of volunteer community health workers who undertake health education outreach, provide simple curative services, and refer patients to health centers for care for more complicated conditions. Generally speaking, patients seek care at the facility closest to their residence and are then referred on to higher-level facilities as the nature of their medical condition requires.

Although health care in Uganda is poor by international standards, it has improved over the past decade. Hence, a challenge in comparing the results of the P2P study (which was conducted in 2004) with the ACT Health replication (which was conducted in 2014) is that baseline conditions may have changed so much as to make further improvements more difficult to achieve. A comparison of baseline conditions in P2P and ACT Health is provided in Table 1.

Table 1: P2P and ACT Health Baselines Compared

Outcome	P2P	ACT Health Baseline (2014)		
	HCIII	HCII	HCIII	All HCs
Under-5 Mortality	144	111.2	111.0	111.1
Share of visits to the HC of all visits	30%◇	31.2%	32.8%	31.9%
Share of visits to traditional healers/self-treatment of all visits	34%◇	34.8%	35.3%	35.0%
Waiting time	131 minutes◇	98	129	111
Staff absence rate	47%◇	43.5%	42.3%	42.7%
Equipment used during examination	41%◇	64.3%	69.9%	66.7%

◇ indicates measures for which baseline values are not included in Björkman and Svensson (2009). For such cases, we use endline measures in the control group data instead. Our calculation deviates slightly from P2P to minimize measurement error – we define it as the sum of the waiting time declared before the initial consultation and the time between the initial consultation and the examination; whereas the P2P approach takes the time between arrival at the clinic and departure from the clinic, minus the duration of the examination.

The greatest change is in under-5 mortality rates. According to UNICEF's figures, national under-five mortality rates declined from around 133 per 1,000 in 2005 to 56 per 1,000 in 2014.[5] These national trends suggest that comparisons in treatment outcomes across the two studies may be challenging. That said, using the relatively naive estimation approach described in Björkman and Svensson (2009)[6], we have a more similar baseline level of under-five mortality rates than national trends would suggest. In particular, we have a baseline level of 111.1 deaths under five per 1,000 live births, compared to 144 per 1,000 live births in P2P. The divergence from the UNICEF figures could be explained by the fact that we are working in relatively remote, rural areas, where improvements in child health take longer to reach.[7]

[5]Figures are from http://data.unicef.org/child-mortality/under-five.html.

[6]Where the numerator is the number of children under five who died in the past 12 months, and the denominator the number of children who were born alive in the past 12 months.

[7]We are also revising our measure of under-five mortality rates for data analysis. See Appendix B for a discussion of different methods of calculating child mortality rates.

4 Experimental Design

4.1 Factorial Design

We employ a factorial design to break the complex treatment contained in the P2P study into two of the three principal components described above. We combine the information and mobilization components into one treatment arm and cross it with the interface treatment, as depicted in Figure 1. Health centers (and their associated catchment areas) are randomized into these four treatment combinations. The comparison of average outcomes in the units assigned to the upper left and bottom right cells constitutes our replication of P2P. The comparison of average outcomes in the other cells (and in the rows and columns) gives us leverage in ascertaining which aspects of the P2P intervention are "doing the work" in generating the dramatic effects reported in Björkman and Svensson (2009).

		T2: Interface meetings are held between community and health facility staff	
		No (T2=0)	Yes (T2=1)
T1: Report card information is reported separately to community and health facility staff and separate action plans are developed	No (T1=0)	Control	Interface Only
	Yes (T1=1)	Information and Mobilization Only	**P2P Replication (Full)** Information and Mobilization + Interface

Figure 1: ACT Health Treatment Groups

4.2 Sampling

As in P2P, our unit of analysis is the health center and its associated catchment area (i.e., the area surrounding the health facility from which it draws its patient base). Our sample includes nearly the entire universe of 379 functioning government-funded HCII and HCIII-level health centers in our sixteen study districts.[8] Our inclusion of both HCII and HCIII-level facilities is a departure from

[8]The sixteen districts are: Lira, Apac, Pader, Gulu, Lamwo, Kitgum, Agago, Katakwi, Bukedea, Manafwa, Tororo, Kabarole, Mubende, Nakaseke, Kibaale, and Bundibugyo. These districts were selected because of GOAL's strong pre-existing relationships with local partner organisations in these areas. Initially, the study was to include Kyegegwa and Otuke districts rather than Mubende and Apac. After mapping health centers' catchment areas as described below, it became evident that 20 health centers had to be excluded from the sample because the study design requires a one-to-one mapping of catchment areas onto health facilities, but the proximity of health centers in Gulu and Tororo districts generated a number of overlapping catchment areas. To preserve the sample size, after consultation with our partners, it was decided to drop two small districts from the study sample and add the two bigger districts of Mubende and Apac. For similar reasons, we also dropped, within the sixteen districts in our sample, health centers that were less than 2.5 km apart or that shared a village among their three closest villages (which is our definition of a health center's catchment area, as described below). We also did not include government health centers funded by the military or prison departments because of the unique communities they serve.

P2P, which only included HCIII-level clinics in their study. Our decision to include these lower level units stemmed from our desire to test whether community mobilization operated differently at this lower level of the health system and also to ensure that our study included the tier of the system that provides primary care for the largest number of poor citizens.[9] Our sample includes 227 HCIIs and 152 HCIIIs. Randomization to our four treatments was blocked by district and health center level. In our analyses below, we break out our results for HCIIs and HCIIIs so as to test for these differential effects (and also to provide comparisons with P2P at the same level of the health system). The sample size for each treatment cell, broken down by HC level, is provided in Table 2.

		T2: Interface meetings		
		No	Yes	
T1: Information and mobilization	No	95 HCIII: 41 HCII: 54	97 HCIII: 37 HCII: 60	192
	Yes	94 HCIII: 38 HCII: 56	92 HCIII: 36 HCII: 56	186
		189	189	

Table 2: Number of HCs, by Treatment Group

For the purposes of both data collection and community mobilization, we define the catchment area as the three villages that are closest in proximity to the health center in question (including the village in which the health center is located), as measured by straight-line distance from the health center to the village centroid.[10] In identifying these villages, we only include villages located in the same parish (for HCIIs) or sub-county (for HCIIIs) as the health center. If only two villages were located within a parish or sub-county, then only these two villages were included in the catchment area.[11]

This definition of the catchment area differs from the one used in P2P, in which it is defined as all people living within a 5 km radius of the health center. We decided to alter the definition for three reasons. First, defining the catchment area in terms of all households within a 5 km

[9] Björkman and Svensson describe their sample as focusing on "the lowest tier of the health system where a professional interaction between users and providers takes place" (2009: 738). Although this characterization is incorrect, it suggests the importance they attached to testing for the impact of their intervention at the level of the health system that was closest to the citizens it served.

[10] Catchment areas were determined according to this algorithm using QGIS mapping software, village-level shape files provided by the Uganda Bureau of Statistics (UBOS), and health center GPS coordinates either collected by GOAL or provided by the Ministry of Health.

[11] In addition, if a village was split into smaller subunits (typically the village subunits would be named "A" and "B" or "1" and "2") and if field teams confirmed that this had occurred within the last 12 months (or had not been formally recognized by the appointment of a new local council), then both of these villages were included and considered as a single village.

radius (which defines an area of 78 square kms) would result in significant overlap between the catchment areas of neighboring health centers, which suggests that the actual catchment areas are significantly smaller. Second, it is logistically much easier to create a sampling frame for collecting representative household data if we define the catchment area in terms of the three closest villages. Finally, the three most proximate villages provide a more natural community (or communities) for mobilization than the set of people who happen to live within a 5 km radius of the clinic.

Within each catchment area, we created a sampling frame of households containing either at least one child under five years old or a pregnant woman, based on village household lists and consultations with the village chairperson (LC1 Chairperson), VHT members, Health Unit Management Committee (HUMC) members and other knowledgeable persons.[12] We then randomly sampled the households to which we administered our household surveys (described below) from these lists.

4.3 Statistical Power

The original P2P study had weak power, with just 50 units of randomization. This meant that results had to be very large in magnitude to be statistically significant and that even large results might only be weakly significant. For example, the headline health outcome result of a 33 percent drop in child mortality is not in fact significant at the conventional 5 percent level. ACT Health is much more highly powered – both for the investigation of the impact of our main treatments and for our replication of the P2P intervention itself.

We conducted a series of power calculations using Optimal Design. Due to the fully factorial experimental design, we assess power for two types of questions. First, we assess whether the study is suitably calibrated to detect *main* effects – that is, the effect of information and mobilization alone or the effect of holding an interface meeting alone. These inferences are made by comparing the entire rows or columns in Figure 1 above. When assessing main effects we are able to use all 379 of our units. Second, we assess whether the study is calibrated to detect *partial* effects – the interactive effects of T1 and T2. These inferences are made by comparing individual cells in Figure 1 above – the key comparison for our purposes being between the bottom right cell (the full P2P package) and the upper left cell (control). When conducting tests of partial effects we use only half of the full data, which decreases the sensitivity of the test.

We make a number of standard assumptions in the power calculations. First, for all tests we set alpha (significance level) at the conventional 0.05, and power at 0.8, implying an 80 percent chance of detecting an effect of the specified magnitude if all assumptions are met. Second, we set the proportion of variance explained by covariates or blocking variables to 0, a very conservative

[12] In instances in which our informants were unsure about the ages of children in a particular household, we verified this information by visiting the household with a knowledgeable person from the village.

approach. Third, in assessing the power of individual tests we assume the intra-cluster correlation (rho) for each clinic-catchment area is 0.06, which we took from the P2P control data for child weight. We set the number of sampled households per health center to 40, which is the number of households we actually sampled. Finally, for all power calculations we report the standardized minimum detectable effect, which represents the minimum detectable effect in terms of standard deviations in the outcome variable. Our estimates are reported in Table 3.

Table 3: Minimum Detectable Effects (MDEs), in standard deviations

Individual level:	
Main effect	0.084 (178 v 178)
Partial effect	0.119 (94 v. 94)

To interpret the MDEs reported above, consider the case of under-5 mortality. We use under-5 mortality rates as the benchmark for our power calculations for two reasons: first, the under-5 mortality rates is one of the key dependent variables in the original P2P study and the reduction in this measure is probably that study's most discussed finding. Second, and more importantly, changes in mortality rates are among the hardest to detect since mortality is a binary outcome that – thankfully! – occurs rarely. Thus, our ability to detect changes in under-5 mortality rates is a binding constraint in our power analysis. Björkman and Svensson (2009) find a 33 percent decline in under-5 mortality. Our baseline data suggests that under-5 mortality in our study areas, pre-intervention, is around 11 percent. We use a conservative estimate of 9 percent for the calculation, accounting for the natural decline in child mortality. This suggests that in any random sample of under-5 children, a one standard deviation decline would be around 29 percent.

From this we then know that a 0.10 MDE is equivalent to a 2.9 percentage point drop in under-5 mortality (from 9 percent to 6.1 percent). This represents, roughly, a 30 percent decline, akin to the point estimates from Björkman and Svensson (2009). Thus the target MDEs for the individual-level effects is around 0.10, which implies that the study is sufficiently powered to detect the target effect at conventional levels.

5 Data

5.1 Data Collection Instruments and Methodologies

As in P2P, our major data collection instruments consist of a household survey and a health center survey. Households in each catchment area were selected for inclusion in the household survey by

randomly selecting 40 of them from the sampling frame lists described above.[13] If a household head could not be found after two attempts, the household was replaced with one of the randomly drawn replacement households. The surveys collected information about household health seeking behavior, experience at the health center that serves the catchment area, and maternal and child health. The primary respondent was the female household head. All household surveys included an anthropometric survey component in which we recorded the weight, height and middle-upper arm circumference (MUAC) of each child under the age of five in the household. The ages of the children, and their immunization status, were also gathered using immunization cards, if available. The households that were interviewed at baseline were re-interviewed at endline to create a panel. We were successful in reinterviewing 94.4 percent of our study households.

During the baseline only, an additional abbreviated ("short") survey was administered to another 15 households in units assigned to the information and mobilization treatments (i.e., units along the bottom row in Figure 1). These additional households were included to reduce noise in the measures included in the CRC and to ensure that the community felt like the CRC represented their views and experiences.

The household surveys were conducted in ten local languages with the help of 279 field staff trained by IPA Uganda. All data was collected using smart phones, with date and time stamps and GPS coordinates recorded and information transmitted to an encrypted server on a daily basis. In all, we completed 15,457 household surveys at baseline (the data from which was cleaned and turned into health center-specific CRCs for use in the units that involved the distribution of the report cards within two weeks of collection) and 14,598 household surveys at endline.

We also administered a comprehensive survey at each health center by interviewing the most senior health center staff member. If the in-charge was unavailable, we interviewed the next most highly ranked (or longest serving) health center staff member. There were three components to the health center surveys. The first was a brief questionnaire completed at the time of our enumerator's first contact with the health center. Since this visit was unannounced, it provided an opportunity for the collection of information about staff attendance, cleanliness, wait times and other clinic characteristics before the clinic staff was able to respond to the fact that they were being evaluated. The second component was the main health center staff survey, which collected information about the variety and quality of health services provided, utilization rates, staff structure and perceptions, funding mechanisms and drug stock-outs. This survey was conducted during a follow-up appointment scheduled at a time that minimized the impact on patient care. The third component involved the collection of administrative data on file at the health facility on monthly Health Man-

[13]From all eligible households (households with children under five or a pregnant woman), 40 were randomly selected to be interviewed. The random selection was made so that the number of households drawn from each village was proportional to the number of eligible households per village. Also, 7 replacement households were generated per village.

agement Information System (HMIS) forms and drug stock cards. Physical checks of drug stocks and storage conditions were also conducted to verify the accuracy of the health center records.

A notable difference between the data collection in ACT Health and P2P was the clear separation in our study between the organization that collected the baseline and endline data (IPA Uganda) and the organization(s) implementing the programming whose impact is being assessed (GOAL Uganda and its team of implementing partner organizations). In the P2P study, both the intervention and the evaluation were undertaken by the same organizations and staff.

This said, some of our analyses also take advantage of the comprehensive data that GOAL Uganda collected to monitor its own implementation activities. For example, we draw on information GOAL recorded on the issues that were included in the action plans developed by community members and health center staff (as well as the issues that made it into their joint social contracts) to assess whether citizens and providers have different preferences and perceptions of the main problems they face, and whether different treatments were associated with the highlighting of different kinds of issues. We also take advantage of GOAL's careful documentation of issues that arose during the course of implementation that might compromise our analysis to undertake robustness tests of our main findings.

5.2 Main Measures/Indices Used in the Analysis

We employ three different sets of measures in our analyses. The first are outcome measures constructed exactly as in P2P in order to maximize comparability. The second aim to capture the same set of outcomes as the first but deviate from the indicator construction in P2P in areas where we think the measures can be improved. Finally, we augment these outcome measures with additional indicators designed to test whether the intervention was carried out as intended, to help us better understand the mechanisms at work in generating any effects we detect, and to test for differential effects across different types of health centers and communities.

5.3 Dependent Variables Used in P2P

We constructed all the main dependent variables employed in the P2P study using the same survey questions and coding protocols. These measures cover five main topics:

1. The degree of monitoring that citizens undertake and the information they have about health provision, as measured by:
   (a) the presence of a suggestion box for complaints and recommendations
   (b) the presence of numbered waiting cards for patients

(c) the presence of posters informing patients about free services

(d) the presence of posters informing patients about their rights and obligations

(e) whether household members report that the local health center was discussed at the most recent LC1 meeting

(f) whether household members report that they had received information about the roles and responsibilities of the health unit management committee (HUMC).

2. Health worker behavior and treatment practices at the health center, as measured by:

(a) whether staff used any equipment during the examination

(b) the length of time the patient was made to wait before being examined

(c) the absence rate of health center staff

(d) the condition of the floors, walls and furniture, and the smell of the health center

(e) whether patients received information about the importance of visiting the health facility and the danger of self treatment

(f) whether patients received information about family planning

(g) the share of months in which stock cards indicated no availability of drugs.

3. Utilization rates, as measured by:

(a) the average number of patients visiting the health center per month for outpatient care

(b) the average number of deliveries at the health center per month

(c) the average number of antenatal visits at the health center per month

(d) the average number of family planning visits at the health center per month

(e) the share of visits to the project health center of all health visits, averaged over the catchment area

(f) the share of visits to traditional healers and self-treatment of all health visits, averaged over the catchment area.

4. Immunization rates, measured in terms of whether children under 5 years old had received at least one dose of measles, DPT, BCG, and polio vaccines and a vitamin A supplement, broken down by the following age categories:

(a) newborn

(b) under 1 year

(c) 1 year old

(d) 2 years old

(e) 3 years old

(f) 4 years old.

5. Health outcomes, as measured by:

   (a) the number of births per household during the year between baseline and endline

   (b) whether any women in the household are or were pregnant during the year between baseline and endline

   (c) under-5 mortality, per 1,000 live births

   (d) whether any children died in the household during the year between baseline and endline

   (e) weight-for-age for children under 18 months. For details of how these variables were constructed, we refer readers to the descriptions provided in Björkman and Svensson (2009).

5.4 Improved Dependent Variables

Our goal of replicating P2P requires that we employ a set of outcomes measures that are as close as possible to the ones used in the original study. However, we think that some of the measures used in P2P are not the best way to capture the impact of the intervention (or its component parts). Hence, we introduce a set of alternative outcome indicators that we think provide the strongest means of testing for treatment effects. These measures, listed below, fall into five broad categories; utilization, treatment quality, patient satisfaction, and health outcomes, including child mortality broken out as its own independent measure.. They are similar, but not identical to those employed in P2P.

For each category, we construct an averaged z-score index consisting of the variables listed below as in Casey, Glennerster and Miguel (2012) and Kling, Liebman and Katz (2007). A z-score is constructed by subtracting the mean of the control group and dividing the variable by the standard deviation of the control group. The averaged z-score index is then constructed by averaging across the z-scores. Signs of individual z-scores are oriented such that a higher value implies a more beneficial outcome.

The advantage of this method is that it reduces the number of comparisons. The index can be interpreted as the average of the included measures, scaled to standard deviation units. Although we will report treatment effects on all index components to shed light on which outcome variables

are driving results, the averaged z-score indices and the under five mortality rate are the primary outcome variables.[14]

Utilization rates

- Vaccination rates of children under 36 months for polio, DPT, BCG, and measles; (based on household data), by age bracket
- Number of visits to the health center by sampled households in the past 12 months, based on household data
- Share of self-reported visits to sampled health center versus other providers (i.e., other government health center, private health center, traditional healers, or self-treatment), based on household data.

Treatment quality

- Whether household members declare that, during their most recent visit to the health center, equipment was used during examination
- Waiting time (total time spent at facility per visit, as reported by community members during their most recent visit to the health center) consisting in the total amount of time spent by the household members waiting for the initial consultation and the examination
- Attendance rate (percent of staff in attendance during the visit to the health center)
- Condition of clinic (cleanliness of floors and walls; whether the clinic smelled, as observed by the enumerator during unannounced visit to health center)
- Share of months in which stock cards indicated no availability of drugs (e.g., six key tracer drugs) in the past three months
- Whether household members declare that, during their most recent visit to the health center, they were examined by trained health care staff
- Whether household members declare that, during their most recent visit to the health center, they had privacy during their examination
- Whether household members declare that, during their most recent visit to the health center, lab tests were administered

[14] In cases where index components include health center level measures in addition to household level measures, we use the household as the unit of analysis.

- Whether household members declare that, during their most recent visit to the health center, their diagnosis was clearly explained to them.

Patient satisfaction

- Whether household members declare that the services currently offered at the health center are of "very high quality" or "somewhat high quality"
- Whether household members said that they were "very satisfied" or "satisfied" with the quality of care received during their most recent visits to the health center
- Whether household members declare that, during their most recent visit to the health center, the person conducting the examination behaved politely/showed respect
- Whether household members declare that, during their most recent visit to the health center, the person conducting the examination appeared to be interested in their health condition
- Whether household members declare that, during their most recent visit to the health center, the person conducting the examination listened to what they had to say
- Whether household members declare that, during their most recent visit to the health center, they felt free to express themselves to the person conducting the examination
- Whether household members declare that, compared to the year before, the availability of medical staff has improved at the health center
- Whether household members declared that they did not visit the health center for any of the following reasons: "staff is not available", "lack of drugs at facility", "poor quality of services", "poor staff attitude", "long waiting time", "lack of cleanliness", or "staff not well trained" (entering negatively).

Health outcomes

- Weight for age among children aged 0-18 months
- Weight for age among children aged 18-36 months
- Upper arm circumference among children aged 0-18 months
- Upper arm circumference among children aged 18-36 months.

Child mortality

We use a synthetic cohort life table approach to calculate neonatal and under-five mortality. Mortality probabilities for small age segments are combined into the common age segments for the two measures of mortality. This approach requires the date of birth, survival status and date or age of death for each child in the sample. Details for the mortality calculations are provided in the appendix.

- Neonatal mortality rate: Share of children in the catchment area younger than 28 days who have died in the last 12 months, as estimated from the household survey (entering negatively)

- Under-five mortality rate: Share of children in the catchment area under 5 years who have died in the last 12 months, as estimated from our household survey (entering negatively)

Note: we did not collect the dates of birth and death at baseline or midline for the children that were reported to have died during the recall period. These dates are needed to use the synthetic cohort life table approach. We will attempt to collect these missing dates during the endline household survey. At baseline and midline, we do ask the age at death, so for cases of missing dates of death we will impute the missing dates by assuming that they are evenly distributed throughout the year and randomly assigning a date of death for each reported child death. Subtracting the age at death from this date will give us an imputed date of birth.

5.4.1 Manipulation Checks

We measure a number of outcomes that serve as manipulation checks. We will report these to provide readers with the ability to make judgments about the fidelity with which the intervention was implemented. These findings are purely informational, however, as we estimate our treatment effects via an ITT analysis.

- Whether household members said that they have heard about the intervention components in response to the questions: "Some government health centers and their communities in this district were visited by [...] to discuss health service delivery. During the meeting, they disseminated a Citizen Report Card (CRC) showing the performance of the health center on a number of categories. According to your knowledge, did any such meetings take place in partnership with [...] in the past 12 months?" and "In some government health centers in this district, [...] organized meetings where health workers and the community came together to discuss how to improve health services. According to your knowledge, did any such meetings between health workers and the community take place here at [...] in the past year?"

- Whether health center in-charges said that they have heard about the intervention components in response to parallel questions to the above in the HC survey

- Whether household members said, in the household survey, that they have attended meetings in response to the questions: "Did you attend any such Citizen Report Card meeting?" and "Did you attend any such Interface Meeting?"

- Whether health center in-charges said that they have attended meetings in response to parallel questions to the above in the HC survey

- Whether household members correctly recall details on intervention (CRC, social action plan) in response to the following questions:

    - "Do you know what kind of activities took place during these meetings?"
    - "On what topics do you recall information being presented in the Citizen Report Card?"

5.5 Intermediate Outcomes

As discussed in section 2 above, the hypothesized theory of change underlying both P2P and ACT Health is that the intervention will generate a series of changes in community members' and health providers' knowledge, behavior, and attitudes that will lead to improvements in utilization, treatment practices and bottom-line health outcomes. Although Björkman and Svensson (2009) do investigate one key mechanism through which P2P has its impact – community monitoring – we analyze a significantly larger number of intermediate outcomes in order to gain a deeper understanding of the channels through which the P2P intervention is "doing its work." We do this with the help of seven indices, whose components are described below. As with the indices that capture our main outcomes of interest, we will test for project impact on each of these intermediate outcomes individually, and also on an index we will construct for each category using an averaged z-score approach.

Citizen knowledge

- Number of patients' rights that citizens are able to name correctly

- Number of patients' responsibilities that citizens are able to name correctly

- Number of services offered at the health center that citizens are able to name correctly

Health care provider knowledge

- Number of patients' rights that health center staff are able to name correctly
- Number of patients' responsibilities that health center staff are able to name correctly

Efficacy

- Whether household members think that they have "a lot of" or "some" influence in making this village a better place to live
- Whether household members agreed with the statement: "People like you have a say about how the GOVERNMENT provides health care to your community"
- Whether household members agreed with the statement: "People like you have a say about how HEALTH FACILITIES provide health care to your community"
- Whether household members think that they have "a lot" or "some" power to improve the quality of health care at their local health center in response to the following question: "Some people think that people in our community have no power to improve the quality of health care at [the HC]. Other people think that people in our community do have power to improve the quality of health care at [the HC]. What do you think?"
- Whether household members think they would be able to pressure a health worker to report to work on time in response to the following question: "If community members found out about a health worker regularly coming late, what do you think are the chances that they would be able to pressure that health worker to report to work on time?"
- Whether household members think they would be able to pressure a health worker to exert better effort in caring for patients in response to the following question: "If community members found out about a health worker not providing the effort that he/she should in caring for his/her patients, what do you think are the chances that they would be able to pressure that health worker to exert better effort in caring for patients?"

Community responsibility

- Whether household members think they are responsible for making sure that health workers come to work and provide high quality health services
- Whether household members replied "community members" in response to the question: "Who else do you think is responsible for making sure health workers come to work and provide high quality health services?"

Community monitoring

- Whether household members report having attended LC1 meetings in the last year
- Whether household members that attended the meeting declared that the local health center was discussed at the most recent LC1 meeting
- Whether household members have received information from the HUMC
- Frequency of HUMC monitoring at Health Center level
- Whether household members think that engaged community members would find out about it if a health worker at their health center did something wrong, like not providing the effort that he/she should in caring for his/her patients.
- Whether household members think that engaged community members would find out about it if a health worker at their health center did something wrong, like not reporting for work.

Relationship between health center and community

- Overall satisfaction of the health center staff on their relationship with the community, mentioning "satisfied" or "very satisfied" level
- Whether household members declared that they were "satisfied" or "very satisfied" with their relationship with the [HC] staff
- Whether household members declared that they trust health workers at [the HC]?
- Whether household members did not say that the facility staff would "refuse to see me" or "behave hostilely toward me" in response to the question: "If you had a complaint about the quality of services at [the HC] and you decided to talk to the facility staff, how would they respond?"

Health center transparency

- Whether a poster showing the health center's opening and closing hours is visible during an unannounced visit
- Whether a staff duty roster is displayed publicly, as ascertained during an unannounced visit
- Whether a suggestion box is present, as ascertained during an unannounced visit

- Whether information is posted listing services provided in the clinic, as ascertained during an unannounced visit

- Whether information is posted about patients' rights and responsibilities, as ascertained during an unannounced visit

5.6 Independent Variables

5.6.1 Treatment Indicators

Taking advantage of the factorial design, we test for the effect of three different definitions of treatment, which are captured in the following treatment variables:

- An indicator variable taking the value 1 if a health center catchment area was assigned to receive information and mobilization, 0 otherwise (the rows in Figure 1, $Info$)

- An indicator variable taking the value 1 if a health center catchment area was assigned to receive an interface meeting, 0 otherwise (the columns in Figure 1, $Interface$)

- An indicator variable taking the value 1 if a health center catchment area was assigned to receive the combined treatment (the bottom right cell in Figure 1, $Full$), and 0 if a health center catchment area was assigned to the pure control group (the top left cell in in Figure 1)

5.6.2 Controls

We include a vector of control variables in our regressions (all measured at baseline) to improve the precision of our estimates:

- An indicator variable taking the value 1 if a health center is a HCII

- An indicator variable for whether the facility provides official delivery services

- An indicator variable for whether a health center has staff houses

- Whether household members report having used this health facility within the 12 months prior to the baseline

- Average education level of the interviewed head of household in health center catchment area, measured in years of education

- Average household wealth level in the health center catchment area, calculated as the first component of a principal component analysis of the number of items of 17 assets (including cattle, radios, bicycles etc.) owned by the household and 3 measures of housing quality.

In addition to these variables, we will add the following to the vector of controls if balance tests indicate imbalance at the 10% significance level across our treatment arms at baseline: values of our utilization rate, treatment quality, patient satisfaction, and health outcome indices; infant mortality; the number of visits of VHTs to households in the last 12 months; the number of trained medical staff at the HC; whether the HC has piped water; whether the HC has electricity (grid or solar); distance to the nearest health provider (government or private/NGO) at the same or higher level; the population density of the 3 km radius around the health center (using LandScan GIS data); and the values of all five of our intermediate outcome indices.

5.7 Attrition, Outliers, and Missing Values

5.7.1 Attrition

As noted, attrition between baseline and endline is low (< 6 percent). Our strategy for dealing with attrition is, following the protocols outlined in Lin, Green and Coppock (2015), to perform a two-tailed unequal-variances t-test of the hypothesis that treatment – defined as full P2P vs. pure control (i.e., the upper left and lower right cells in Figure 1), information and mobilization vs. not (i.e., the rows in Figure 1), and interface vs. not (i.e, the columns in Figure 1) – does not affect the attrition rate. We will consider a p-value below 0.05 as evidence for imbalanced attrition. In such an event, we will follow the approach outlined in Lin, Green and Coppock (2015) of consulting a disinterested jury of colleagues to decide whether the monotonicity assumption for trimming bounds is plausible. If so, we will report estimates of trimming bounds; if not, we will report estimates of extreme value (Manski-type) bounds, in addition to the unbounded analysis.

5.7.2 Outliers

We deal with outliers by capping unbounded variables at the 99th percentile of the observed values in our data.

5.7.3 Missing covariate values

To deal with missing values on our covariates, we will adopt the approach described in Lin, Green and Coppock (2015). Specifically, we will comply with the following rule:

1. If no more than 10% of the covariate's values are missing, recode the missing values to the overall mean.

2. If more than 10% of the covariate's values are missing, include a missingness dummy as an additional covariate and recode the missing values to an arbitrary constant, such as 0.

5.7.4 Missing dependent variables

To deal with missing values on our outcome measures, we will follow Kling, Liebman and Katz (2007) and impute missing values by setting them equal to the mean of each outcome variable for the relevant treatment group.

6 Analysis

6.1 Replication of P2P

A principal objective of ACT Health is to test whether the findings reported in P2P can be replicated. Hence, a first set of tests involve an exact replication of the main tables (tables 2-6) in Björkman and Svensson (2009). These tables report estimates of program impact on the outcomes described in section 5.3 above.

These analyses will use only the data collected in control units and in units that received the full P2P intervention (i.e., those in the upper left and bottom right cells in Figure 1. Since all of the outcome measures included in these analyses were collected in exactly the same way in ACT Health and in P2P, replicating the tables presented in Björkman and Svensson (2009) is straightforward. For details of the specific statistical models that we will run, we refer readers to the descriptions provided in Björkman and Svensson (2009).

6.1.1 Replication of P2P in the Subset of Units Whose Baseline Health Outcomes Match Those in P2P

As noted above, a challenge in comparing the findings of P2P and ACT Health is that the baseline conditions present at the time of the P2P study are different from those present at the time of ACT Health. To improve the comparability of the results, we will therefore replicate the full set of analyses described in section 6.1 in the subset of units in which baseline conditions at the time of ACT Health are roughly similar to those present at the time of P2P. Since under-5 mortality is the headline finding in P2P, we define "roughly similar baseline conditions" as baseline levels of under-5 mortality that are no more than one standard deviation lower than the under-5 mortality reported at baseline in P2P.

6.1.2 Replication of P2P in HCIIIs only

As also noted above, a significant difference between the P2P study and ACT Health is that the former only included HCIIIs, whereas the latter included both HCIIIs and HCIIs. Insofar as the two levels of health centers differ in ways that are potentially germane to treatment uptake, the

cleanest comparison between the findings in P2P and ACT Health will be in the subset of units in the latter study that are HCIIIs. We will therefore break out the full set analyses described in section 6.1 into the subset of units that are HCIIIs and HCIIs.

6.2 Replication of P2P Using Improved Outcome Measures

Next, we will test for the impact of ACT Health on the improved dependent variables described in section 5.4. Since these are our preferred outcome variables, these analyses will provided us with the clearest test of the replicability of P2P. We will again focus our analysis solely on the units in the upper left and lower right cells of Figure 1.

To estimate the overall effect of the treatment, we estimate the following ITT equation[15]:

$$Y_{ij} = \beta_0 + \beta_1 T_{ij}^{Full} + \beta_2 Y_{ij}^0 + \beta_3 X_{ij} + \phi_d + u_{ij} \quad (1)$$

where Y_{ij} is the outcome measure of household i in health center catchment area j. T_{ij}^{Full} is the treatment dummy for the combined treatment (the bottom right cell in Figure 1 in comparison to the pure control group, *Full*). Thus, we are leveraging only half of our sample in this specification. β_1 is the average treatment effect, Y_{ij}^0 is the baseline value of the outcome measure, X_{ij} is a vector of controls, ϕ_d are district fixed effects, and u_{ij} are robust standard errors clustered by the health center catchment area. Following Lin et al. (2013), we use Huber-White sandwiched standard errors.

We use this equation to test for the impact of treatment on each of the outcome indices listed in section 5.4, capturing utilization rates, treatment quality, patient satisfaction, health outcomes, and child mortality. While we consider the indices our main outcomes, we will also report results for each individual component.

Although we consider Equation 1 our main specification, we will also show simple difference-in-means (t-tests) for all primary outcomes and treatment arms. We will also run the following robustness tests: specifications without controls, without district fixed effects, with the outcome measures aggregated at the health center level, and with the outcome measure defined as the difference between post-treatment and pre-treatment values. As a tertiary analysis, we will run the main specification dropping units for which we have evidence that the quality of implementation was severely compromised. We will also test the hypothesis of non-effect on any of the four outcomes using the nonparametric combination approach, as proposed by Caughey, Dafoe and Seawright (Forthcoming).

[15]Equation 1 differs from the models employed in Bjorkman and Svensson (2009). But, having adopted their modeling choices in the pure replication in section 6.1, we feel this is the right specification for testing the impact of the intervention.

6.3 Heterogeneous Treatment Effects

Although Björkman and Svensson do not explore this issue in their 2009 paper, there are a number of hypotheses in the literature that might lead us to expect to find varying treatment effects across different types of HCs and/or in communities with different characteristics. We therefore undertake a number of tests of heterogeneous treatment effects. We consider these tests for heterogeneous treatment effects secondary analyses, so we do not include them in the set of primary families to correct for multiple comparisons, discussed below. For these analyses, we estimate treatment effects on just the five main outcome indices (utilization rates, treatment quality, patient satisfaction, health outcomes, and mortality rates) rather than for each individual outcome measure. To test for heterogeneous treatment effects, we estimate the equation:

$$Y_{ij} = \beta_0 + \beta_1 T_{ij}^k + \beta_2 T_{ij}^k * Het_{ij} + \beta_3 Het_{ij} + \beta_4 Y_{ij}^0 + \beta_5 X_{ij} + \phi_d + u_{ij} \quad (2)$$

where T_{ij}^k is the treatment dummy for treatment indicator k. k indicates which treatment component we are assessing: the provision of information and mobilization (the rows in Figure 1, *Info*), the provision of the interface meeting (the columns in Figure 1, *Interface*), or the combined treatment (the bottom right cell in Figure 1 in comparison to the pure control group, *Full*). Het_{ij} is an indicator variable of the subgroup for which we are testing heterogeneous treatment effects, which for this purpose is not included in the vector of covariates X_{ij}. X_{ij} includes a dummy indicating assignment to the interface if k = *Info* and a dummy indicating assignment to the provision of information and mobilization if k = *Interface*. All other terms are the same as in equation 1. β_2 is the marginal increase in treatment effect in health center catchment areas in this subgroup.

As with our main analyses in section 6.2 above, we will also run these specifications without controls, without district fixed effects, and with the outcome measures aggregated at the health center level.

6.3.1 Facility-level characteristics

We test for differential treatment effects with respect to three characteristics that vary across health facilities.The first is whether, the health center is a HCII or HCIII-level facility. These analyses allow us to compare the effects of the intervention across the two different levels of health centers that we study.

The second is whether, at baseline, the health center is performing above or below the median level in its district in our treatment quality index, as measured at baseline.[16] The rationale for this test is that the impact of the intervention is likely to be different for well-performing and poorly-

[16] As a rule of thumb, we define 'above the median' as at or above the median.

performing health centers, in part because the nature of the information contained in the CRC about the quality of the health services provided at the health center will be different.

A third health facility-level characteristic we will investigate is the availability of alternative healthcare options. The greater the availability of alternative sources of health care, the greater the likelihood that community members will respond to the receipt of information about poor service provision at their own health center by exiting rather than by exercising voice. We will measure this via an averaged z-score index comprised of two components:

1. the distance to the nearest other health centers, in kilometers, including both government health centers and private/NGO-run clinics, and including health centers at the same and at higher levels, as measured from GIS data

2. the share of self-reported visits to sampled health center versus other providers (i.e., other government health center, private health center, traditional healers, or self-treatment)

We will test for heterogeneous treatment effects between health centers that are above and below the median value for this index.

6.3.2 Catchment area characteristics

In addition to these facility-level attributes, we also test for heterogeneous treatment effects across units located in catchment areas with different characteristics. The first of these is the catchment area's collective action potential at baseline. This is likely to be important insofar as the P2P intervention depends on the ability of community members to work together to monitor HC staff and sanction them if they are found to be underperforming. We will measure the community's collective action potential constructing a z-score index of the following two components:

1. the share of people who respond "yes" to the question: "If you wanted to take action to improve the quality of care provided at [the HC], do you think other community members would join you in your efforts?" in a health center catchment area

2. the ethnic heterogeneity of the catchment area, calculated from our household survey data and constructed using a Herfindahl index (entering negatively).

We then construct an indicator taking the value 1 if this z-score index is above the median. The rationale for the second measure of collective action potential stems from the large literature indicating that ethnically diverse communities have a more difficult time achieving collective ends (e.g., Miguel and Gugerty (2005), Khwaja (2009), Algan et al. (2016)). Consistent with these findings, Björkman and Svensson (2010) report that the impact of the P2P intervention was stronger in health facilities that were located in more ethnically homogeneous districts.

Insofar as P2P is about community monitoring, we might also expect communities that are already actively engaged in monitoring will respond differently to treatment than communities whose baseline levels of monitoring are lower. We test this expectation by comparing outcomes in communities whose community monitoring index values (as defined in section 5.5) are above and below the median.

A final catchment area-level characteristic that may plausibly affect treatment uptake is the community's baseline level of efficacy. As Lieberman, Posner and Tsai (2014) argue, even individuals who are motivated to act by the receipt of information about poor service delivery in their community may be dissuaded from acting if they do not believe that they have the power to effect change. To the extent that this is so, we might expect communities in which baseline levels of efficacy, as measured by the efficacy index (as described in section 5.5), are above the median to respond more strongly to the intervention.

For additional context, we also test for heterogeneous treatment effects based on the interval in each unit between treatment and the collection of our endline data. Although we endeavored to keep this interval equal across health centers, the complexity of managing dozens of field teams managed by two different organizations in hundreds of units led to some variation in the number of days that had elapsed between the delivery of our treatment (by GOAL Uganda and its partners) and the collection of our endline data (by IPA-Uganda). Insofar as the impact of the treatment might decay over time (or, conversely, take time to generate changes in our outcomes of interest), this variation might plausibly be associated with different measured treatment effects. To test whether they are, we compare units with above- and below-median values for the interval between treatment and endline data collection.

6.4 Mechanisms

A major advance of ACT Health over P2P is our ability to investigate the mechanisms through which P2P operates. We do this via two different strategies. First, we test for the effect of the treatment on a series of intermediate outcomes that capture different channels through which the treatment may operate. Then, employing our full data set, we compare our main treatment outcomes across the rows and columns of Figure 1. That is, we compare treatment outcomes in units that did and did not receive the information and mobilization treatments and in units that did and did not receive the interface meeting treatment. By comparing the size of these differences we can get some purchase on the relative importance of these two different components of the broader intervention. We also take advantage of questions included in the household survey at endline to learn about the mechanisms underlying any differential effects we may find across these two treatment components.

6.4.1 Uncovering Mechanisms by Investigating Intermediate Outcomes

We re-estimate equation 1, redefining Y_{ij} in terms of the intermediate outcomes described in section 5.5. The logic underlying this approach is that if the treatment affects health care delivery through its impact on intermediate outcome Q, then we should see an effect of the treatment on Q. Of course, we should also see an effect of Q on health care delivery, but estimating this second stage is much more complicated (Imai, Keele and Tingley, 2010), so we limit our analysis to the first stage, which provides at least suggestive evidence for whether or not the causal pathway in question is relevant.

We test for the relevance of each of the seven intermediate outcomes described in section 5.5. For ease of exposition, we test for the impact of treatment on each index value rather than on each individual component of each measure. The rationale for each mechanism can be found in the theoretical discussion in section 2. Specifically, we test whether the P2P treatment has its impact by

- increasing citizen knowledge about their rights and responsibilities, and about the services offered at the HC
- increasing health providers' knowledge about patients' rights and responsibilities
- increasing the sense of efficacy felt by community members
- increasing the extent to which community members believe they are responsible for monitoring health workers to make sure they are doing their job
- increasing the extent to which community members are, in fact, engaged in monitoring their HC
- improving the relationship between HC staff and the community
- increasing the extent to which HC staff are transparent

6.4.2 Leveraging the Factorial Design

As explained in the theoretical discussion in section 2, whereas the information component of P2P address the problem of "noise" in the ability of community members to determine how hard health care workers are working on their behalf, the interface component addresses the problem of unobservability. To test which of these components is responsible for any treatment effects we observe, we break the P2P intervention into two components: the rows in Figure 1, which record whether or not the unit received the information and mobilization part of the overall treatment, and the columns, which record whether or not the unit received the interface meeting. To test for the

differential effects of these two components, we estimate the following equation for all the main outcome measures described in section 6.1:

$$Y_{ij} = \beta_0 + \beta_1 T_{ij}^{Info} + \beta_2 T_{ij}^{Info} T_{ij}^{Interface} + \beta_3 T_{ij}^{Interface} + \beta_4 Y_{ij}^0 + \beta_5 X_{ij} + \phi_d + u_{ij} \quad (3)$$

where are all terms are the same as specified in equations 1 and 2. As in our main analyses, we will also run these specifications without controls, without district fixed effects, and with the outcome measures aggregated at the health center level. To test for the effect of being assigned to the interface treatment, we conduct an F-test on the hypothesis that $\beta_1 + \beta_2 = 0$. To test for the effect of being assigned to the information/mobilization treatment, we conduct an F-test on the hypothesis that $\beta_2 + \beta_3 = 0$.

In contrast to the analyses described thus far, which employed data only from the upper left and bottom right cells in Figure 1, these analyses use the entirety of our data. We will therefore have double the power to estimate the effects of these two components of the full P2P intervention.

Exploring mechanisms responsible for different effects across treatment arms

To explain any differences we may find across the information/mobilization and interface treatment arms, we undertake two different types of analyses. First, we re-estimate equation 3, redefining Y_{ij} in terms of each of the intermediate outcomes described in section 5.5. Then, we explore the effect of each treatment arm on a series of survey responses, collected at endline, regarding community members' perceptions of the effects of each treatment.

First, we compare likelihood that community members exposed to the *Info* and *Interface* treatments say:

- that they would "ever share your complaint about the HC with somebody?"
- that "this intervention changed the relationship between community members and health workers?"
- that "this intervention changed the health seeking behavior of community members?"
- that "this intervention changed the behavior of health workers in providing health care services?"
- that "this intervention changed the relationship between community members and health workers?"
- that, "if [they] had the opportunity to receive an updated CRC annually/attend another interface meeting between health workers and the community every year, they would"

Second, we can exploit the fact that community members in some catchment areas received the full treatment (i.e., both $Info$ and $Interface$). This put them in a unique position to weigh in on their impression of the relative usefulness/power of each component. Among this subset of community members, we therefore can learn from answers to the questions:

- "Overall, if you had the opportunity to either receive an updated Citizen Report Card every year or to attend another Interface Meeting between health workers and the community every year, which would you choose?"

- "Overall, in your opinion, what intervention activity did you find most useful?"

We are also interested in testing some specific hypotheses about why each component works. For example, we hypothesized that the importance of the interface meeting lies in the opportunity it affords community members to better observe HC staff behavior. To get at this issue, we will compare across the interface and information/mobilization treatment arms:

- the share of "don't know" responses to the question: "Sometimes when health centers are not performing as well as they should, it is because the staff is not doing its job properly/is not hardworking. But sometimes the staff is working as hard as it can but the problem lies in the challenges they face. Which do you think is the bigger issue at [the HC], health workers not working hard enough or other challenges they don't have control over?"

- the share of respondents who say they are "very" or "somewhat" informed in response to the question: "How informed are you about the attendance of health workers at [the HC]?"

A second expectation regarding the effects of the interface meetings is that they will improve the nature of the relationship between community members and health center staff and, in particular, increase the extent to which community members trust the HC staff. To test these expectations we will compare our "relationship between health center and community index" across respondents exposed to the Info and Interface treatment arms. We will also analyze the impact on each individual component of the index, as described in section 5.5 above.

Similarly, we hypothesized that, to the extent that the receipt of the CRC changes community members' behavior, it is because the information the CRC contains is felt by community members to be relevant. To test this, we will exploit the answer to the question that was asked of community members who received the CRC:

- "On a scale from 1 to 10, with 10 being the highest score, how relevant was the information you received?"

6.5 Multiple Testing Correction

Given the number of outcome variables in our study, multiple testing is a concern. We will report both uncorrected p-values and results from the Benjamini and Hochberg (1995) False Discovery Rate correction. This simple step-up procedure is slightly less punitive than a Bonferroni correction since it focuses exclusively on correcting for the false discovery rate (type I errors).

For the primary analyses of indices, the family is defined by the treatment definition. In other words, the five main indices form three families, (1) in the analysis comparing the provision of information and mobilization against no provision of information and mobilization (*Info*), (2) in the analysis comparing the interface (*Interface*) against no interface, and (3) in the analyses comparing the full P2P intervention (*Full*) against the pure control. For secondary analyses for heterogeneous treatment effects, the family is similarly defined as the five outcome indices in each of the seven types heterogeneous analyses times the three treatment definitions, yielding 21 families. For supplementary analyses with index components as dependent variables, the family is defined as the components of an index, by treatment definition.

The families for the primary analyses are outlined in table 5 in Appendix C.

7 Ethical Considerations

IRB protocols have been approved at IPA (Protocol ID: 0497) and at the Uganda National Council for Science and Technology (UNCST) (Protocol ID: ARC157). Approval for the project was also received from UNCST itself (Protocol ID: SS3559) and Office of the President, Uganda. Participation in the study is voluntary and all respondents need to have given their informed consent in order to participate. Respondents do not receive any compensation for their time. In order not to distract health workers from performing their duties, enumerators were instructed to interrupt the survey when a health worker was busy and to resume when she was again available. All data collection is electronic, using SurveyCTO, an ODK based platform. PDAs are password protected and data is uploaded to an encrypted server on a daily basis, networks permitting. Data is stored on password protected computers using encryption and removing all PII from the datasets.

References

Algan, Yann, Camille Hémet, David Laitin et al. 2016. "The social effects of ethnic diversity at the local level: A natural experiment with exogenous residential allocation." *Journal of Political Economy* 124(3).

Andrabi, Tahir, Jishnu Das and Asim Ijaz Khwaja. 2014. "Report cards: The impact of providing school and child test scores on educational markets." HKS Working Paper No. RWP14-052.

Arrow, Kenneth Joseph. 1974. "Essays in the theory of risk-bearing." *Journal of Business* 47(1).

Banerjee, Abhijit V, Rukmini Banerji, Esther Duflo, Rachel Glennerster and Stuti Khemani. 2010. "Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India." *American Economic Journal: Economic Policy* pp. 1–30.

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 289–300.

Besley, Timothy. 2007. *Principled agents? The political economy of good government*. Oxford University Press.

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

Björkman, Martina and Jakob Svensson. 2010. "When is community-based monitoring effective? Evidence from a randomized experiment in primary health in Uganda." *Journal of the European Economic Association* 8(2-3):571–581.

Björkman Nyqvist, Martina, Damien De Walque and Jakob Svensson. 2014. "Information is power: experimental evidence on the long-run impact of community based monitoring." *World Bank Policy Research Working Paper No 7015* .

Casey, Katherine, Rachel Glennerster and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Pre-Analysis Plan." *Quarterly Journal of Economics* 127(4):1755–1813.

Caughey, Devin, Allan Dafoe and Jason Seawright. Forthcoming. "Testing Elaborate Theories: A Nonparametric Framework." *Journal of Politics* .

Chong, Alberto, L Ana, Dean Karlan and Leonard Wantchekon. 2015. "Does corruption information inspire the fight or quash the hope? A field experiment in Mexico on voter turnout, choice, and party identification." *The Journal of Politics* 77(1):55–71.

Clemens, Michael A. 2015. "The meaning of failed replications: A review and proposal." *Journal of Economic Surveys* .

Hölmstrom, Bengt. 1979. "Moral hazard and observability." *The Bell journal of economics* pp. 74–91.

Humphreys, Macartan and Jeremy Weinstein. 2012. "Policing politicians: Citizen empowerment and political accountability in Uganda." Working Paper.

Imai, Kosuke, Luke Keele and Dustin Tingley. 2010. "A general approach to causal mediation analysis." *Psychological Methods* 15(4):309.

Khwaja, Asim Ijaz. 2009. "Can good projects succeed in bad communities?" *Journal of Public Economics* 93(7):899–916.

Kling, Jeffrey R., Jeffrey B. Liebman and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75(1):83–119.

Lieberman, Evan S, Daniel N Posner and Lily L Tsai. 2014. "Does information lead to more active citizenship? Evidence from an education intervention in rural Kenya." *World Development* 60:69–83.

Lin, Winston, Donald Green and Alexander Coppock. 2015. "Standard Operating Procedures: A Safety Net for Pre-Analysis Plans." Forthcoming in *PS: Political Science & Politics*, 2016.

Lin, Winston et al. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *The Annals of Applied Statistics* 7(1):295–318.

Miguel, Edward and Mary Kay Gugerty. 2005. "Ethnic diversity, social sanctions, and public goods in Kenya." *Journal of Public Economics* 89(11):2325–2368.

Ross, Stephen A. 1973. "The economic theory of agency: The principal's problem." *The American Economic Review* 63(2):134–139.

Appendix

A Synthetic Cohort Life Table Approach to Measuring Mortality

A **synthetic cohort life table** approach is used in which mortality probabilities for small age segments based on real cohort mortality experience are combined into the more common age segments. This approach allows full use of the most recent data and is also specific for time periods.

The synthetic cohort life table approach requires each child's date of birth, survival status and date or age of death. To collect the required information, we extend the household child roster of living children to also ask about deaths.

We will then have a complete list of all children alive at the time of the endline survey and their months of birth along with a list of all children who died in the period between the baseline and endline surveys, their months of birth, and the month that they died. We can then take the following steps to calculate the under-five mortality rate.

- Step 1: Create an age variable for each month of the recall period between baseline and follow up. Let's say that our first baseline survey is conducted between September and December, 2014 and the endline is conducted between September and December of 2016. We would create a variable for the age in months beginning in September 2014: $age_{sep13}; age_{oct13}; \ldots; age_{dec13}; \ldots; age_{dec14}$.

- Step 2: Create a binary variable for the month that a child died: $died_{sep13}; \ldots; died_{dec14}$.

- Step 3: Create an age at death variable for each month: $agedied_{sep13}; \ldots; agedied_{dec14}$.

- Step 4: For each month, find the total number alive at each age 0-59 months.

- Step 5: Sum the total number for each age across all months. (Use the mean number for each age of the current and previous months.)

- Step 6: For each month, find the total number that died at each age.

- Step 7: Sum the total number that died at each age across all months.

- Step 8: Calculate the mortality rate for each age in months (number of deaths at age A/total number of months at age A).

- Step 9: Calculate the survival rate for each age in months (1-mortality rate).

- Step 10: Calculate the overall survival rate by multiplying the individual age-specific survival rates across the relevant ages (i.e. 0-59 months for under-five survival rate).

- Step 11: Find the overall mortality rate (1-survival rate).

The other mortality measures (neonatal, infant, child) can be calculated by using the corresponding periods.

B P2P and ACT Health Interventions Compared

Please see Table 4 on the next page.

C Families for Multiple Comparison Corrections

Table 5: Families for Multiple Comparison Corrections

Primary analyses			
Family	Dependent variables	Treatment	Specification
(1)	Utilization rates	T^{Full}	Equ. 1
	Treatment quality	T^{Full}	Equ. 1
	Patient satisfaction	T^{Full}	Equ. 1
	Health outcomes	T^{Full}	Equ. 1
	Child mortality	T^{Full}	Equ. 1
(2)	Utilization rates	T^{Info}	Equ. 3
	Treatment quality	T^{Info}	Equ. 3
	Patient satisfaction	T^{Info}	Equ. 3
	Health outcomes	T^{Info}	Equ. 3
	Child mortality	T^{Info}	Equ. 3
(3)	Utilization rates	$T^{Interface}$	Equ. 3
	Treatment quality	$T^{Interface}$	Equ. 3
	Patient satisfaction	$T^{Interface}$	Equ. 3
	Health outcomes	$T^{Interface}$	Equ. 3
	Child mortality	$T^{Interface}$	Equ. 3

Table 4: P2P and ACT Health Interventions Compared

Area	Difference	P2P Intervention	ACT Health Intervention
Program Intervention	Treatment arms (N)	Full P2P program (25); Control (25)	Full P2P program (92); control (95); interface only (97); information and mobilization only (94)
	Involvement of community-based organisations (CBOs)	Worked through 18 CBOs, many of which had been active in promoting health improvements in area prior to intervention	Worked in consortium with 4 implementing partner organizations, none of whom had been involved in health-related programming prior to intervention
	Number of participants at community dialogue meetings	150 community members	75 community members
	Number of participants at interface meetings	Unclear, but certainly <75 because participants were chosen at community meeting	27-75, depending on treatment arm
	Participation of sub-county officials	Unclear	Sub-county chief, Community Development Officer (CDO) and Health Inspector invited to observe, but rarely attended[DP1]
	Role-playing in interface meetings?	Yes	No
	Six-month follow-up meeting?	Yes	Yes
	Who ran the intervention?	Stockholm University and World Bank	GOAL Uganda
Research Design	Districts included in study	9 districts	16 districts
	Number of HCs studied	50 government run HCs	379 government run HCs
	Level of HCs studied	HCIIIs only	HCIIs (226) and HCIIIs (152)
	HC catchment area definition	All villages within 5km radius around the HC	Three villages closest to the HC
	HH sample Size	Panel of 5,000 households surveyed at baseline and endline; 100 per HC catchment area	Panel of 14,598 households surveyed at baseline and endline; 55 per HC catchment area (40 long surveys; 15 short surveys)
	Attrition across baseline and endline	12 percent	5.6 percent
	Study dates	Baseline: end of 2004 Intervention: 2005 Endline: start of 2006	Baseline: end of 2014 Intervention: 2015 Endline: end of 2015
	Who ran the evaluation?	Stockholm University and World Bank	IPA Uganda