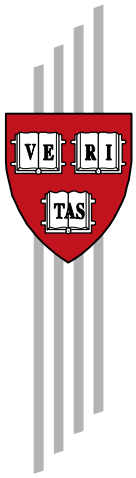
Can Transparency and Accountability Programs Improve Health? Experimental Evidence from Indonesia and Tanzania

Jean Arkedis, Jessica Creighton, Akshay Dixit, Archon Fung, Stephen Kosack, Dan Levy, and Courtney Tolmie

CID Faculty Working Paper No. 352

May 2019



Working Papers

Center for International Development at Harvard University

Can transparency and accountability programs improve health?

Experimental evidence from Indonesia and Tanzania[1]

Jean Arkedis

Jessica Creighton

Akshay Dixit

Archon Fung

Stephen Kosack

Dan Levy

Courtney Tolmie

[1] We are grateful to the broader T4D team, past and present, including Sarah Alphs, Matthew Bombyk, Eric Englin, Hannah Hilligoss, Jenna Juwono, Rohit Naimpally, James Rasaiah, Lindsey Roots, Rachmat Reksa Samudra, Astri Arini Waluyo, and J. Preston Whitt. For generous funding for this project, we thank the William and Flora Hewlett Foundation, the UK Department for International Development, and the Bill and Melinda Gates Foundation, and the Transparency and Accountability Initiative, which conceptualized and brokered the project. We are grateful as well to colleagues at our partner partner organizations: Hilda Kigola, Esther Mtumbuka and Happy Ndomba of the Clinton Health Access Initiative in Tanzania and Novita Anggraeni, Riska Amelia Hasan, Yulius Hendra Hasanuddin, Didik Purwandanu and Sad Dian Utomo of PATTIRO in Indonesia. This work would not have been possible without the tireless efforts of colleagues at our data collection partners, including Celine Guimas, Matt Wiseman and team at EDI Group; Samson Kiware, Isaac Lyatu, Diego Shirima and Godfrey Siwingwa of Ideas for Action; and Cristina Clerici, Rachel Jones, and Martin Zuakulu of Innovations for Poverty Action, all in Tanzania; and Kamti Ningsih, Setyo Pujiastuti, N. Wayan Suriastini and the rest of the team at SurveyMETER in Indonesia. We would also like to thank our partner J-PAL Southeast Asia, especially Lina Marliani, Eki Ramadhan, and Hector Salazar Salame. We also thank scholars Iqra Anugrah, Megan Cogburn, Mohmed Yunus Rafiq, and Kankan Xie for their ethnographic studies of several communities who participated in the program. We gratefully acknowledge the members of the T4D advisory committee: Yamini Aiyar, Jessica Cohen, Jonathan Fox, Anuradha Joshi, Dan Posner, and Bill Savedoff. Finally, our most important debt is to the thousands of community members in Indonesia and Tanzania who volunteered their time and energy to participate in the programs we evaluate here. The pre-analysis plan for this paper was registered at the AEA RCT registry (#655) prior to the endline data collection.

Abstract: We assess the impact of a transparency and accountability program designed to improve maternal and newborn health (MNH) outcomes in Indonesia and Tanzania. Co-designed with local partner organizations to be community-led and non-prescriptive, the program sought to encourage community participation to address local barriers in access to high quality care for pregnant women and infants. We evaluate the impact of this program through randomized controlled trials (RCTs), involving 100 treatment and 100 control communities in each country. We find that on average, this program did not have a statistically significant impact on the use or content of maternal and newborn health services, nor the sense of civic efficacy or civic participation among recent mothers in the communities who were offered it. These findings hold in both countries and in a set of pre-specified subgroups. To identify reasons for the lack of impacts, we use a mixed-method approach combining interviews, observations, surveys, focus groups, and ethnographic studies that together provide an in-depth assessment of the complex causal paths linking participation in the program to improvements in MNH outcomes. Although participation in program meetings was substantial and sustained in most communities, and most attempted at least some of what they had planned, only a minority achieved tangible improvements and fewer still saw more than one such success. Our assessment is that the main explanation for the lack of impact is that few communities were able to traverse the complex causal paths from planning actions to accomplishing tangible improvements in their access to quality health care.

I. Introduction

In 2017, over 4 million children died within the first 28 days of life (Dicker et al., 2018), most from diseases and complications that are readily preventable or treatable with proven, cost-effective healthcare services (You et al., 2015). In recent years, citizen and community scorecards, social audits, and other programs that try to increase transparency and accountability have been widely explored as tools to improve access to and quality of public services. One premise of these programs is that information empowers citizens to improve the responsiveness, accountability, and ultimately the effectiveness of their public services (Fox, 2007; Glennerster, 2005; McGee and Gaventa, 2011; Molina et al., 2017). This premise has been validated by several studies, notably Björkman and Svensson (2009), who find that a community scorecards program reduced infant mortality by a third in one year.[2] However, other studies show little or no effect of these programs.[3] The overall picture that emerges from the empirical literature is mixed (Fox, 2015; Joshi and Houtzeger, 2012; J-PAL, 2011; Joshi, 2010; McGee and Gaventa, 2011; Kosack and Fung, 2014).

In this paper, we present results from randomized controlled trials (RCTs) from the Transparency for Development (T4D) project of a non-prescriptive, community-led transparency and accountability program designed to encourage civic participation to improve access to high quality maternal and newborn healthcare. The T4D program was co-designed and piloted over a two-year period with local partner organizations[4] and then offered in one hundred communities

[2] Other experimental studies that have reported positive impacts include Banerjee et al. (2018), Andrabi, Das and Khwaja (2017), Fiala and Premand (2017), and Molina et al. (2017). See Section II.

[3] For example, Olken (2007), Banerjee et al. (2010) and Lieberman, Posner and Tsai (2014), Raffler, Posner and Parkerson (2018). See Section II.

[4] The partner in Tanzania was the Tanzania country office for the Clinton Health Access Initiative (CHAI), an organization focused on improving health service delivery in a number of domains. The partner in Indonesia, PATTIRO, is a research and policy advocacy organization, which focuses on regional and local governance issues in a number of sectors.

each in Indonesia and Tanzania, two contexts that differ substantially in terms of access to quality healthcare, choice among health service providers, as well as general levels of economic development. We leverage a mixed-methods approach combining the RCTs with focus group discussions, ethnographic studies, interviews, and systematic observations to evaluate the effects of the program on health outcomes and civic participation in these communities.

At a high level, the T4D program involved a trained facilitator from the partner organizations identifying a group of interested community members, informing them of MNH-related problems in their community, and helping them develop a plan of social actions to overcome problems with access to quality maternal and newborn health care in their communities. We focus on the impact of this program on the quality[5] and use of maternal and newborn health care services in communities in which the program was offered as well as infant health outcomes and civic participation and perceptions of civic efficacy among recently pregnant women in the broader community. For these pre-specified primary outcomes, we find that the T4D program had no significant average effects.

We use the interviews, surveys, focus groups, systematic observations, ethnographic studies, and other qualitative data to help explain these null average observed effects.[6] As others have noted, the causal chains linking transparency and accountability programs to improvements in public services are lengthy and complex.[7] We propose a framework of causal paths that may link participation in T4D to improvements in maternal and newborn healthcare services, health,

[5] Within the high-quality health-system framework proposed by The Lancet Global Health Commission (Kruk et al., 2018), our metrics refer to "Processes of care" and within that "Competent care and systems".
[6] Drawing on a wide variety of data sources provides important insights into the key processes that led to the observed null effects, addressing the black-box critique of impact evaluations (Rao & Woolcock, 2003; White, 2011; Deaton & Cartwright, 2016).
[7] Fox, 2015; J-PAL, 2011; Joshi and Houtzeger, 2012; McGee and Gaventa, 2011; Lieberman, Posner, Tsai, 2014; Kosack and Fung, 2014.

and civic empowerment, and assess the program's effects on each step of that framework. Overall, we find that community participation in both countries was substantial and sustained, and that the meetings created space for participants to leverage local knowledge and collectively plan diverse courses of action for improving their maternal and newborn health care. In most communities, those who participated in the meetings subsequently tried at least some of the actions they had planned. Many tried to educate their neighbors about the importance of delivering in a facility and seeking antenatal and postnatal care as well as to improve their health care services in other ways. In 87% of communities, participants in the final program meeting described having completed at least one of these activities. Yet interviews at the time with participants, and those with whom they had planned to engage, suggest that the proportion of communities in which participants succeeded at their goal was closer to 45%. When participants were asked to reflect on the program approximately a year and a half later, they could recall a specific, tangible improvement from their efforts in only 35% of communities (41% in Indonesia; 30% in Tanzania).[8] The proportion of communities in which participants were able to recall at least two tangible improvements that they had achieved—progress that might indicate broader rather than one-off improvements in access to quality care—was still lower: only 14% of communities in Indonesia and 4% in Tanzania.

In short, we do not find evidence that participants were able to achieve measurable improvements to maternal and newborn health care in the average community that was offered the program, and our main explanation for these findings is that participants in few communities were

[8] Note that we focus in this paper on impact among these broader communities. For an analysis of the experiences of those who participated directly in the program (attending meetings, planning and implementing social actions, etc.), see Kosack et al. (forthcoming).

able to traverse the complete path from planning actions to realizing improvements in their maternal and newborn health care.[9]

The rest of this paper is structured as follows. In Sections II and III, we present an overview of the existing literature and the contexts in Indonesia and Tanzania where the program was offered. We describe the program and research design in Sections IV and V. In Section VI, we present our findings from the RCTs. We unpack and triangulate the findings in Section VII, and conclude with Section VIII.

II. The Experimental Literature

The promise of transparency and accountability programs for improving the quality of public services or governance has been subjected to a growing number of empirical tests in recent years. Conclusions have been mixed. A prominent study in Uganda found remarkable improvements in maternal and newborn health soon after a community scorecard program encouraged bottom-up community monitoring (Björkman and Svensson, 2009). Banerjee et al. (2018) find that citizens in villages in Indonesia who had been informed of their rice subsidy entitlements received 26% more subsidy on net, and Andrabi, Das and Khwaja (2017) find that providing information to citizens in Pakistan on test scores in education markets with multiple public and private providers improved subsequent education outcomes. Fiala and Premand (2017) find positive effects from grassroots monitoring of a community-driven development program on the quality of a range of community projects in Uganda.

[9] Both contexts were marked by noteworthy secular improvements in healthcare over the course of the evaluation period. We discuss this further in Section VII.

However, a large number of other studies have found little or no effects of these programs. Banerjee et al. (2010) and Lieberman, Posner and Tsai (2014) find no evidence of impact on education from offering parents information about the low quality of the education offered in their children's schools in India and Kenya, respectively. Olken (2007) finds that a community monitoring intervention in Indonesia had little impact on corruption. Raffler, Posner and Parkerson (2018) find that even an information and mobilization intervention modeled closely after Björkman and Svensson (2009) and evaluated fifteen years after the original study in the same country, Uganda, had no detectable effect on healthcare utilization or outcomes. Reviews and meta-analyses of the experimental literature have also failed to reach consensus: Holland and Schatz (2016) and Molina et al (2017) find mostly positive effects, 3ie (2018) find that very few have had much effect and that most struggled even to encourage much civic participation, and Kosack and Fung (2014) and Fox (2015) find mixed effects.[10] Communities are often able to improve their public services when they decide on their own to try (Mansuri and Rao, 2012). Yet overall, whether transparency and accountability programs can encourage communities to measurably improve public services remains debated.

III. Contexts

Indonesia is a lower-middle income country in Southeast Asia with 2017 GDP per capita of USD 3,846 (World Bank, 2017) that has experienced multiple peaceful transfers of power since

[10] Kosack and Fung review 16 experimental evaluations that varied on the service or sector they sought to improve as well as many other dimensions. They find that those associated with measured impacts tended to offer information on inputs as well as outputs, performance as well as rights, that was subjective as well as objective, that allowed comparisons of how public services were working relative to other places or benchmarks, and that not only offered information on public services but also encouraged participation to improve them. Fox (2015) reviews 25 evaluations and notes that they differ on two additional dimensions—breadth of the community they tried to organize, and length of time they offered information and capacity building and other support to those communities as they organized—and concludes that those that saw less impact tended to be low dosage.

the fall of an authoritarian regime in 1998 (Freedom House, 2018). The T4D transparency and accountability program was offered to communities in two provinces on two of Indonesia's eighteen thousand islands: Banten and South Sulawesi.

In a household survey conducted at baseline, most respondents in these provinces resided in sturdy dwellings with walls made of stone (48%) or wood (37%), and 99% households had electricity. Both also had relatively high levels of access to MNH care at baseline. The MNH service delivery system in Indonesia involves a diverse range of providers. The focal point of public services is usually a public health center that operates at the sub-district level, known as "puskesmas." Puskesmas provide comprehensive basic health services, generally including delivery services. They are the lowest level public health service center overseen directly by the Indonesian government, and every village is assigned to the catchment area of a puskesmas. A baseline survey of facilities in sample areas indicated that the average puskesmas served 8.8 villages and had close to 57 staff, with nearly 14 in the maternal/delivery unit alone. Nearly all surveyed puskesmas had electricity and 95.5% used water from an improved source. A puskesmas often oversees and is supported by a network of smaller health centers. Indonesia also has long had a nationwide community-based health program known as "posyandu." Staffed by a midwife and local volunteers, posyandus offer monthly antenatal care services to pregnant women, and growth monitoring and vaccination programs for children under the age of five. At baseline, 99% of villages in our sample had a puskesmas, a smaller health center overseen by a puskesmas, a village midwife operating under the supervision of the puskesmas, a birthing clinic, or a private provider located within the village. Almost all baseline respondents (82%) reported having more than one provider or facility from which they could seek care.

Yet respondents in sample areas also described a mixed experience with the quality of their care. For example, nearly all respondents received some form of antenatal care, and close to 90% of respondents said they had completed the recommended four ANC visits during pregnancy (Table 1). But only 55% of respondents gave birth at a facility, and only 22% of baseline respondents reported receiving three components of ANC: having their blood pressure checked, having a urine sample drawn and receiving a report of the results, and having a blood sample drawn and receiving a report of the results.[11]

Primary MNH Outcomes	Indonesia (% baseline respondents)	Tanzania (% baseline respondents)
Whether the respondent had a first antenatal care visit within the first trimester	69	19
Whether the respondent attended four or more antenatal care visits over the course of the pregnancy	87	43
Whether the respondent gave birth with a skilled provider	79	56
Whether the respondent gave birth at a health facility	55	56
Stunting – Whether the infant is below 2 standard deviations from the median WHO Child Growth Standards	16	27
Underweight – Whether the infant is below 2 standard deviations from the median WHO Child Growth Standards	16	9

Table 1. Baseline sample means for MNH outcomes in Indonesia and Tanzania[12]

[11] For detailed baseline statistics, please refer to the baseline report (Transparency for Development Project Team, 2016)

[12] There are two additional primary outcomes (utilization of postpartum and postnatal care) that are not included in this table because the associated questions were phrased differently at baseline and endline. At endline, respondents were asked about postpartum/postnatal care checks conducted *after* leaving the birth facility and within 7 days of giving birth. At baseline, however, respondents were asked about postpartum/postnatal care

Tanzania is a different context in several important respects. Its GDP per capita is only a quarter of that of Indonesia (World Bank, 2017), and although elections have been a regular feature of the political landscape since the 1990s, the ruling party has retained power for over half a century (Freedom House, 2018). The T4D program was implemented in two regions, Dodoma and Tanga, in which the average household was resource poor relative to those in Indonesia. In a household survey in communities at baseline, most respondents lived in dwellings made of mud (71.3%), and only 12.5% had access to electricity.

The ecosystem of public provision of MNH care in Tanzania also involves fewer service providers than in Indonesia. The primary public health facility for most communities in Tanzania is a dispensary. Similar to puskesmas in Indonesia, every village is assigned to the catchment area of a dispensary. However, Tanzanian dispensaries tend to be much smaller than Indonesian puskesmas: at baseline, the average dispensary we surveyed in Tanga and Dodoma served 2.7 villages and had 6.4 staff members. Electricity from the grid was the primary power source for only about a quarter of dispensaries at baseline (21.6%). Nearly half (44.4%) reported not having a regular supply of water. Facilities in Tanzania were also less common than in Indonesia: less than two-thirds (62.5%) of villages we surveyed had a dispensary located within the village. Nearly three quarters of respondents to our household survey (73.5%) reported not being able to choose care from more than one health facility.

Our baseline surveys also indicate that the quality of this care tended to be lower than in Indonesia. Antenatal care again offers a clear illustration: While the vast majority of women surveyed at baseline (98.4%) had received some form of antenatal care (ANC), less than half

checks conducted within 7 days of giving birth, *irrespective* of whether that was before or after leaving the birth facility.

(43.4%) had completed the recommended minimum of four antenatal visits during their most recent pregnancy (compared with 90% in Indonesia; see Table 1). Similar to Indonesia, slightly over half of respondents (56%) gave birth at a facility. The quality of ANC was mixed, with less than a third (30.9%) receiving the three basic components of ANC described above.[13]

In short, relative to Indonesia, Tanzania has fewer economic resources, a public health ecosystem with fewer providers in a lower resource setting offering less choice among providers for those seeking MNH care, and lower baseline outcomes in terms of utilization or quality of MNH care.

IV. The T4D Program

T4D was a transparency and accountability program developed over a two-year iterative co-design and piloting process with staff from partner organizations in each country. The initial design was based on the "community scorecard" approach that Bjorkman and Svennson (2009) found to have dramatically improved maternal and newborn healthcare in Uganda, and was subsequently defined, piloted, and iteratively refined over multiple rounds of design and feedback discussions.[14] The design process followed several principles. First, every aspect was co-designed with in-country partners with local knowledge of what was appropriate in their respective contexts. The focus was on improving health, rather than necessarily fixing health service delivery: in particular an area of health—the health of pregnant women and infants—of concern both in Indonesia and Tanzania as well as in the global health community. The program was designed to

[13] For detailed baseline statistics, please refer to the baseline report (Transparency for Development Project Team, 2016)

[14] The design process was modeled on "crawling the design space" in Pritchett et al. (2017). See the Intervention Design report for details (Transparency for Development project team, 2016).

be non-prescriptive and locally relevant: offering space for communities in widely different circumstances to leverage local knowledge and collectively decide on context-appropriate activities for improving maternal and newborn health. In addition, the program was designed to be community-driven (to emphasize the importance of participants using their own knowledge and capacity to understand and fix problems) and to be devoid of material, technical, or relational resources external to the community, so that participants would rely on their existing willingness and capacities to try to improve their maternal and newborn health. Finally, the program was designed to be relatively light-touch and scalable, so that it could be consistently implemented across a large number of diverse communities. To maximize comparability, the program's core components were similar across the two countries. The specifics of the approach were defined, iteratively refined, and validated over the two-year co-design process through a round of "pre-piloting" (quick tests of individual components) in Indonesia and complete pilots (the entire program from start to finish) in both Indonesia and Tanzania.

Broadly, the resulting program was a series of six meetings between a facilitator employed by the partner organizations and a group of citizens from a village, organized over a period of approximately three months. The purpose of these meetings was to offer information and facilitated discussion that would encourage citizens to try to alleviate problems with their maternal and newborn health care that affected them and their neighbors. The program began with two day-long meetings of a small group of 15-16 people from the community whom they had identified over the previous weeks as interested in participating in the meetings as "community representatives" (CRs)[15]. At these meetings, the facilitator first shared information they had

[15] The facilitator recruited community representatives through consultation with village leadership and other community members. Community representatives were meant to have interest in working to improve their community's health care and/or experience with difficulties accessing quality care, and to represent a broad swath

gathered over the previous weeks on three key health indicators in the community: antenatal care visits of pregnant women (in Tanzania) or those with birth preparedness plans (in Indonesia), and how often women gave birth in health facilities and sought postnatal care services. As participants discussed why these indicators were not higher, the facilitator also shared information they had gathered from surveys in the community on a range of barriers that might be preventing more pregnant women from delivering at a facility or seeking antenatal or postnatal care, as well as stories illustrating approaches other communities in the area had taken to improve their services. Finally, the facilitator helped those attending the meetings formulate plans of activities that they could undertake to address the problems with maternal and newborn health in their community.

Immediately following the first two meetings was an open meeting, where those still interested in pursuing their planned activities presented them to the broader community. The last three meetings occurred at approximately 30-day intervals. At each, the facilitator met with the participants to learn about their progress and to help them reflect on challenges they may have faced. At the sixth and final meeting, the facilitator also encouraged them to plan for how they would continue to meet and sustain their progress after the facilitator was no longer organizing meetings.

No additional resources or help were provided to participants in these meetings.[16] Rather, throughout the program it was entirely up to the participants to decide what, if anything, to do to try to improve maternal and newborn health care in their particular context, using only their

of the village, including men, women, a range of ages, and those with a varying degree of previous leadership experience.

[16] The only exception to this is that participants in Tanzania were offered compensation for attending the first two meetings, though not for any of the subsequent meetings.

existing resources and capacities as members of their communities and citizens of Indonesia or Tanzania.

V. Research Design and Data

In this paper, we assess the hypothesis that participating in this program would empower people across diverse environments to act in ways that would improve the access and quality of the maternal and newborn health care services available to pregnant women and small children. The observable implication of this hypothesis is that across diverse contexts, the quality and use of their care would be significantly higher than in other communities.

We leverage an experimental design as well as focus group discussions, interviews, surveys, systematic observations, ethnographic studies, and other qualitative data to present a comprehensive assessment of this observable implication. In particular, we examine both *how* the program worked as well as *whether* it was effective at improving MNH outcomes in these two contexts.

The study consisted of RCTs in both Indonesia and Tanzania, with a "treatment" group of 100 villages and a "control" group of 100 villages in each country. In Indonesia, 85 of the 200 villages were in Banten and 115 in South Sulawesi, and in Tanzania 77 villages were in Dodoma and 123 in Tanga. This section describes the following aspects of the research design: (a) randomization into treatment and control groups, (b) baseline equivalence between treatment and control groups, (c) key outcomes measured, (d) data collection, and (e) estimation strategy.

a) Randomization

Our unit of randomization was the health facility – puskesmas in Indonesia, and dispensary in Tanzania. Our sample included 200 facilities in Indonesia and 153 facilities in Tanzania. Prior to randomization, we randomly selected one village in the catchment area served by each sample facility.[17] In Tanzania, we also randomly selected a second village in the catchment area of 47 sample dispensaries, for a total of 200 villages. We then randomly assigned each health facility and the one or two communities previously selected from its catchment area to a treatment or control group.

Random assignment to the treatment group was stratified on a few key variables. In Indonesia, we stratified on province (Banten or South Sulawesi) and the proportion of women in the village who had delivered in a health facility (above or below the sample median). In Tanzania, we stratified on three binary characteristics: region (Dodoma or Tanga), proportion of women in the village who had delivered in a health facility at baseline (above or below the sample median), and whether there were one or two sample villages in the catchment area of the health facility.

The T4D program was then offered in the randomly selected villages in the catchment area of facilities assigned to the treatment group; the selected villages in the catchment area of facilities assigned to the control group constitute the counterfactual.[18]

[17] This random selection was done after the application of certain criteria – for instance, we dropped urban communities (in Indonesia), and villages with a population of more than 10,000 or less than 1,000 (in Tanzania). For the detailed village selection protocol, please refer to the baseline report (Transparency for Development Project Team, 2016).

[18] Note that the program implementation occurred at the village level, in the sampled village(s) associated with treatment health facilities.

b) Baseline Equivalence

Before the program began, we verified balance between the treatment and control groups on variables used in stratifying as well as a host of other baseline characteristics. In Indonesia, the difference between the treatment and control group was statistically significant at the 5% significance level for only five of the 96 baseline variables tested (including key outcomes),[19] which falls within the expected bounds of random or naturally occurring sample variation. Similarly, in Tanzania, the difference between treatment and control groups was statistically significant at the 5% significance level for only six of 112 variables.[20] (For details, see Appendix A1.)

c) Outcomes

We focus on a set of primary outcomes pre-specified prior to endline data collection[21]: 1) utilization of maternal and newborn health services, 2) the content of these health services, 3) child health outcomes, and 4) civic participation and perceptions of civic empowerment among communities. We also analyze a set of secondary outcomes that are relevant for understanding the

[19] The variables that were statistically significantly different between treatment and control villages were: 1) ANC check - mother received urine sample results; 2) woman ever had an ANC visit because of a complication; 3) proportion of women paying for post-natal care; 4) in most recent effort government officials/political leaders listened to, and took seriously a community proposal; 5) in the past year, respondent or anyone in the household had participated in an information or election campaign.

[20] The variables that were statistically significantly different between treatment and control villages were: 1) whether or not anyone in the household owned a bicycle; 2) a dummy variable for whether or not the respondent felt they were properly informed of what was happening during recent visit to the health facility; 3) whether the respondent gave birth in a private hospital; 4) whether the respondent took a bicycle to the facility for delivery; 5) whether the respondent had taken public transportation to the facility for delivery; 6) and the proportion of children who were underweight (by weight-for age ratios).

[21] The Pre-Analysis Plan (AEA RCT #655) is available at https://www.socialscienceregistry.org/trials/655/

impact on the primary outcomes but are not used to assess the overall impact of the program.[22] (See Appendix A2 for outcome definitions.)

With multiple outcomes, there is increased risk of over-rejecting the null hypothesis of no effects. We address this risk in two ways. First, following Casey, Glennerster and Miguel (2012), we limit the number of outcomes by grouping related outcomes for content of care and civic participation into unweighted mean effects indices. In addition, we control the False Discovery Rate (FDR) using the approach in Benjamini, Krieger and Yekutieli (2006), which limits the expected proportion of rejections that are Type I errors.

Additionally, we analyze a range of intermediate outcomes to explore the mechanisms by which the activities participants planned in the meetings might have influenced their access to quality maternal and newborn health care service, and thereby influenced the primary outcomes. We explore a comprehensive set of characteristics spanning eleven dimensions of healthcare access and provision that might have been influenced by the types of activities participants in the meetings planned, and that correspond to the types of actions planned by the CRs (Table 2; see Appendix A3 for the definitions of the associated intermediate outcomes). We also explore whether women who recently gave birth were aware of more activities to improve access to quality maternal and newborn health care in villages in the sample group.

[22] Although the list of outcomes we explore in the two countries is generally similar, two differences are noteworthy. First, we analyze utilization and content of antenatal care as primary outcomes in Tanzania, but secondary outcomes in Indonesia. This is because baseline findings indicated that antenatal care outcomes in Indonesia were relatively high, and thus information on antenatal care was not included among the indicators that facilitators shared with community members during the first two meetings. Second, the components used to measure content of care (delivery, postpartum and postnatal) differed in the two countries based on their respective healthcare standards.

Finally, we triangulate our analysis of these intermediate outcomes with qualitative interviews and observations and ethnographic studies, described in Section VI.

d) Data

We rely on multiple data sources. The first is a survey of women who had given birth in the last year in sample villages at baseline and endline, which included questions related to our primary, secondary, and intermediate outcomes. At baseline, survey firms in Indonesia and Tanzania interviewed 5,398 recent mothers (2,398 in Indonesia and 3,000 in Tanzania). A second sample of 6,001 women in Indonesia and 6,008 women in Tanzania were interviewed for the endline survey, which was conducted a year and a half after the program ended and three years after the baseline. Respondents in each village were randomly sampled. Prior to data collection, survey teams conducted an extensive listing exercise with village and hamlet leaders, formal and informal healthcare providers, as well as other informants, to prepare a list of all women who had lived for at least six months in the village and had given birth in the prior 12 months.[23] The listing and sampling process is described in Appendix A4.

We complement household data with a survey of healthcare providers in the sample health facilities, using these data mainly to explore intermediate outcomes related to the quality and use of maternal and newborn health care services. To explore how the activities of participants might have influenced the quality and use of these services, we rely as well on qualitative data. First, we analyze the plans of activities that participants developed to try to improve their access to quality

[23] Note that because the sampling design involves interviewing women who had given birth in the 12 months prior to the surveys (baseline and endline), the two waves of data collection constitute repeated cross-sections, not a panel.

maternal and newborn health care.[24] A year and a half after the last program meeting, we also conducted focus group discussions with former participants in all villages who were offered the program, in which they were asked to recall the activities they planned and tried, challenges they had encountered, and what if anything they thought had improved because of their efforts. An average of eight former participants in Indonesia and nine in Tanzania participated in these focus groups in each village.

In a sub-sample of treatment villages—41 in Indonesia and 24 in Tanzania[25]—we interviewed three participants individually to provide more details on how their activities went. Additionally, we conducted key informant interviews with the facilitator, the village head, the village midwife or someone from the health facility, several others with whom the three participants said they had engaged, and several other people randomly chosen in each community to verify what occurred and understand their perspective.

Finally, we rely on four ethnographic studies by four scholars who lived in or near eight communities—two each in Banten, South Sulawesi, Tanga and Dodoma—for 6-9 months spanning the period before, during, and after the intervention, and observed health care in a third nearby community in the control group.[26]

[24] The program was designed for participants to make four plans: one each in the second program (planning) meeting and in the 30-, 60-, and 90-day follow up meetings, and each facilitator was responsible for recording a copy of the social action plans developed during each of these meetings.
[25] These were randomly selected, stratified on the same characteristics as the ones described in Section V.
[26] The eight villages were selected purposively so that they were geographically close to each other to allow a single ethnographer to cover three villages.

e) Estimation strategy

Given that villages were randomly assigned to treatment and control groups, our basic method of estimating program impacts consists of comparing mean outcomes for the treatment and control groups of communities at endline, a year and a half after the program ended. We estimate the following regression equation:

(1) $Y_{ijk} = \beta_0 + \beta_1 TREAT_{jk} + \beta_2 STRATA_k + \varepsilon_{ijk}$,

where Y_{ijk} is the outcome of interest for mother/child i in village j in catchment area k; $STRATA_k$ is a vector of dummy variables that indicate the randomization strata; and $TREAT_{jk}$ denotes treatment assignment. We cluster the error term ε at the health facility level, which is the level of treatment assignment. The coefficient β_1 represents the impact estimate.

For robustness purposes, we also conduct a second set of regressions controlling for village-level averages in outcome variables at baseline (Y_j^0):[27]

(2) $Y_{ijk} = \beta_0 + \beta_1 TREAT_{jk} + \beta_2 STRATA_k + \beta_3 Y_j^0 + \varepsilon_{ijk}$

Finally, we estimate heterogeneity of impact on the primary outcomes for three subgroups specified in the pre-analysis plan – province or region within each country; time since the program ended; and the existing quality of the health care system. We use Equation (3) to estimate heterogeneity in impacts for the first two, region and time:

(3) $Y_{ijk} = \beta_0 + \beta_1 TREAT_{jk} + \beta_2 STRATA_k + \beta_3 GROUP_{ij} + \beta_4 TREAT_{jk} \times GROUP_{ij} + \varepsilon_{ijk,}$

[27] Some primary outcome variables were not measured in the baseline survey. In these cases, the baseline control is either omitted, or we substitute a similar proxy variable.

where $GROUP_{ij}$ is a dummy variable denoting the subgroup of respondent i in village j (either province or region, or whether the respondent gave birth between 0 and 6 months prior to being interviewed at endline). $TREAT_{jk} \times GROUP_j$ is an interaction between this dummy and the treatment dummy variable. To examine impact heterogeneity by variability in the quality of the existing health care system, we use a similar approach but include two group dummy variables for baseline health care quality in Indonesia.[28]

VI. Findings

In both Indonesia and Tanzania, we find that the core components of the T4D program were implemented in villages in the treatment group largely as designed, and that on average, the program had no statistically significant impacts on any pre-specified primary or secondary outcomes in either Indonesia or Tanzania. This section describes findings related to (a) implementation, (b) participation in the program, (c) impacts on primary and secondary outcomes, (d) impacts on intermediate outcomes, and (e) impacts on sub-groups.

(a) *Implementation*

The program in Indonesia was rolled out in two asynchronous waves over the course of seven months in 2015-2016. In Tanzania the program was administered in four overlapping waves

[28] We measure quality at baseline with a set of 12 indicators of minimum service delivery capacities and infrastructure that a health facility should have to be able to effectively provide maternal and newborn healthcare. Given the observed variation in the data, we classified baseline quality of care into three categories in Indonesia (high, medium and low) and two categories in Tanzania (high and low). Hence the inclusion of two group dummies (and their interactions with the treatment variable) for estimating sub-group impacts in Indonesia.

over a course of ten months in 2015-2016 (for details, see Transparency for Development Project Team, 2017). The core components of the program appear to have been implemented with a high degree of fidelity to random assignment and to the design of planned meetings and other activities.[29] In particular, program officers from partner organizations observed a selection of meetings directly and reviewed reports compiled by facilitators on each meeting, including information on who participated, which meetings they attended, and the activities they planned. In addition, external firms were hired to monitor the program in a third of the communities who were assigned to the treatment group and phone calls were made to their village heads to verify that the program had occurred there.

(b) *Participation in the Program*

Although participants in the program were unpaid and offered no additional resources with which to try to improve their access to quality maternal and newborn health care,[30] our data suggest that program meetings occurred largely as planned in all villages in the treatment group and that participants generally stayed engaged throughout the program. In Indonesia, facilitators reported that an average of 12 of the 15-16 people whom they had invited attended the first meeting; meeting observations in a sample of communities in the treatment group suggest similar participation. (The most poorly attended had three participants, but in several communities more people wanted to attend than the facilitator had invited; the largest meeting had 21). The final meeting had an average of nine participants, ranging from one to 17. In Tanzania, attendance at

[29] We explore how the program was perceived and understood in a forthcoming book based on the ethnographic studies of the experiences of eight of the 200 communities who were offered the program.
[30] As noted above, the only exception to this is that participants in Tanzania were offered compensation for attending the first two meetings, but not for any subsequent meeting.

the first meeting averaged 15, ranging from 12 to 16. At the final meeting the average number of participants was 12; the most poorly attended had three participants, and the most well-attended had 28 participants. Meeting observations also suggest that most participants engaged in the discussions rather than sitting quietly, and that in the majority of communities, participants frequently told local stories and offered examples of the more general topics under discussion, and generally conducted discussions and made decisions themselves rather than relying on the facilitator. In both countries, the majority of those who participated in the discussions were women.[31]

Participants in these meetings planned a large number of activities to try to improve their maternal and newborn health care. Participants in the average community in Indonesia planned between three and 17 activities, with a median of seven; in Tanzania, they planned between two and eight actions, with a median of four. In total, participants in the program planned 1,139 activities across the three months of meetings: 715 in Indonesia and 424 in Tanzania. In both countries, these activities were designed to alleviate problems with a wide range of issues, from cost of service to health facility accessibility to expectant mothers' awareness of recommendations and best practices in seeking care. For example, some planned to visit pregnant women to talk to them about the importance of giving birth at a health facility; others to meet with health facility staff to discuss the availability of medicine, supplies, and the high cost of delivery, create lists of community members whose cars could be used to transport patients to health facilities for deliveries and treatment of other illnesses or injuries, and work with their community to repair

[31] See Kosack et al. (forthcoming) for a more detailed consideration of participation in the program and an analysis of participants' expectations of being able to sustain improvements to their health care, both as they went through the meetings and in reflecting on their participation a year and a half after the program ended.

roads to allow easier access to the health facility. Table 2 groups planned activities into 12 broad categories.

Proportion of communities in which participants planned activities to:	**Both**	**Indonesia**	**Tanzania**
Increase awareness, knowledge & improved community attitudes	93.5%	92.0%	95.0%
Improve facility access	71.0%	79.0%	63.0%
Increase ability to pay (including demand-side cost solutions)	45.0%	44.0%	46.0%
Improve information transparency (cost, opening hours, etc.)	39.0%	42.0%	36.0%
Improve attitude, effort, or trust of provider	36.0%	41.0%	31.0%
Pass by-laws, develop partnerships with traditional providers, or other approaches aimed at health service uptake	35.0%	16.0%	54.0%
Increase availability of drugs, supplies, other inputs	28.0%	45.0%	11.0%
Improve facility infrastructure	28.0%	32.0%	24.0%
Increase staff (midwife, doctor, etc.)	17.5%	16.0%	19.0%
Improve facility cleanliness	6.0%	10.0%	2.0%
Improve provider knowledge	1.0%	2.0%	0.0%
Other community improvements – e.g. general hygiene or cleaning campaigns, planting medicinal gardens, or digging community wells	9.0%	18.0%	0.0%

Note: Proportions are based on the activities participants planned across the program meetings.

Table 2. Activities planned by participants

At the final program meeting, when facilitators asked participants which of their planned activities they had completed, they described 58% as complete (53% in Indonesia; 65% in Tanzania) and another 29% as ongoing (31% in Indonesia; 24% in Tanzania). Interviews suggest that the percentage of planned activities that were actually completed by the final program meeting may have been lower (36% in Indonesia; 52% in Tanzania). Nonetheless, both participants' plans

and interviews with participants and others in their community with whom they tried to engage suggest that participants attempted many of their planned activities.

(c) *Impact on primary and secondary outcomes*

In Figures 1 and 2, we present a summary of the T4D program's impact on pre-specified primary outcomes in Indonesia and Tanzania, respectively. Estimates of effect sizes fall short of 0.1 standard deviations across all outcomes, indicating that the differences between the treatment and control groups were small in both countries. Moreover, the 95% confidence intervals include zero in every case. Hence, we cannot reject the null hypothesis of no effect for any of the primary outcomes: utilization and content of MNH services, infant height-for-age and weight-for-age, or civic participation or perceptions of empowerment among recently pregnant women in communities who were offered the program. Similarly, we find no statistically significant differences between the average village in treatment and control groups on any of the secondary outcomes in Indonesia and Tanzania (see Table 3). These conclusions remain unchanged when we estimate Equation (2), controlling for the village average of the relevant outcome at baseline; see Appendix A5.

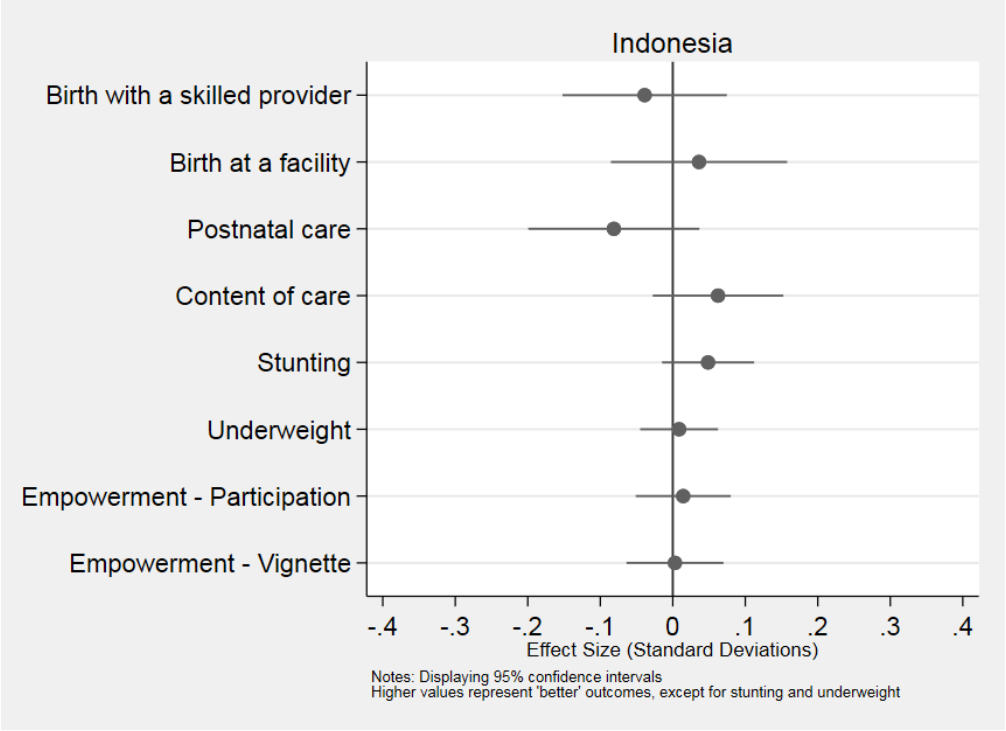


Figure 1. Impact of the T4D Program on Primary Outcomes in Indonesia

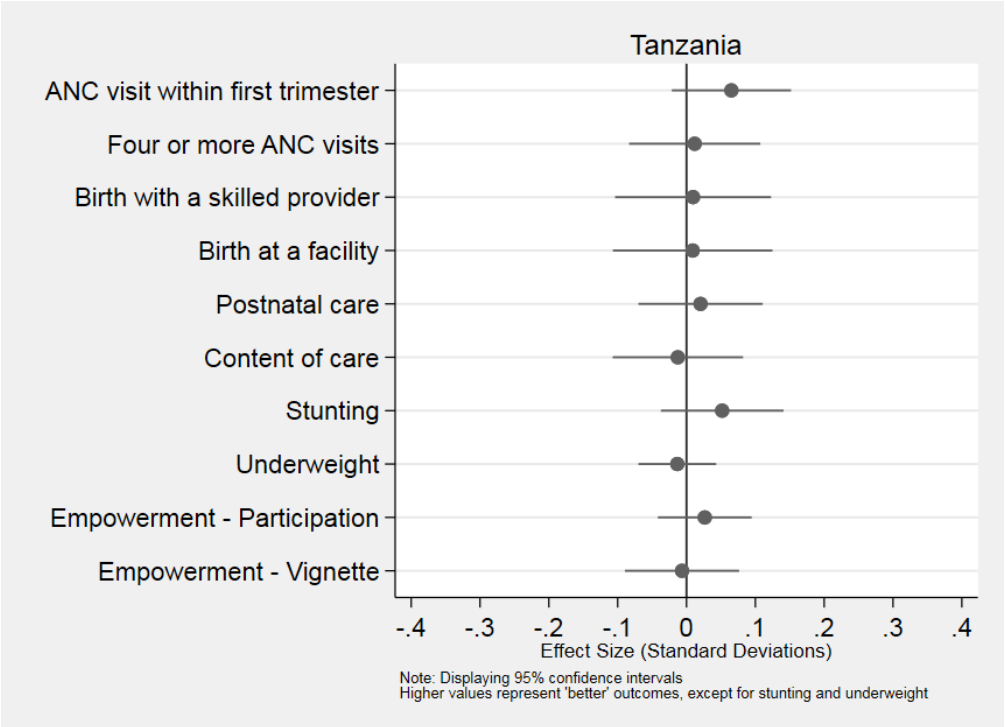


Figure 2. Impact of the T4D Program on Primary Outcomes in Tanzania

Indonesia						
	(1)	(2)	(3)	(4)	(5)	(6)
	Low birthweight	Maternal depression (K6 score)	Birth preparedness	Four or more ANC visits	First ANC visit within the first trimester	Content of Antenatal Care
Treatment	0.00449	-0.0882	0.0157	0.00587	-0.00931	-0.0161
	(0.00773)	(0.151)	(0.0727)	(0.0167)	(0.0175)	(0.0688)
Constant	0.0896***	18.28***	5.202***	0.875***	0.745***	5.979***
	(0.00820)	(0.151)	(0.0728)	(0.0177)	(0.0177)	(0.0657)
Observations	5,423	5,971	6,001	5,994	5,911	6,001
Control Mean	0.08	18.32	4.91	0.83	0.73	5.73
Tanzania						
Treatment	-0.00584	0.0752	-0.00235			
	(0.00611)	(0.182)	(0.0818)			
Constant	0.0521***	18.72***	5.661***			
	(0.00771)	(0.235)	(0.101)			
Observations	6,006	5,859	6,008			
Control Mean	0.05	18.11	5.04			

Notes: Robust standard errors clustered at the facility-level in parentheses. All regressions include strata-specific binary variables. Outcomes in columns (4)-(6) were included as part of the primary outcomes in Tanzania, and hence their impacts are not reported in this table. *** p<0.01, ** p<0.05, * p<0.1

Table 3. Impact of T4D on Secondary Outcomes in Indonesia and Tanzania

(d) *Impacts on Intermediate outcomes*

We also use Equation (1) to estimate the program's impact on the intermediate outcomes. Overall, we conclude that the program did not affect intermediate outcomes in the average village in the treatment group, in either Indonesia or Tanzania. Using the household and facility survey data, we analyze 106 intermediate outcomes in Indonesia, and find statistically significant differences at the 95% confidence level between treatment and control groups for nine (see Appendix A6 for detailed results). With 106 outcomes, five could be significant purely by chance. Once we control the false discovery rate, the adjusted p-values for these outcomes do not meet the conventional thresholds for statistical significance.

Our conclusions in Tanzania are similar. We analyze 126 intermediate outcomes from surveys of recently pregnant women and surveys and observations in health dispensaries, and find statistically significant differences at the 95% confidence level between treatment and control groups for eight (see Appendix A7 for detailed results). With 126 outcomes, six could be significant purely by chance, and as in Indonesia, the observed statistical significance of these eight outcomes in Tanzania does not survive correction for multiple hypotheses testing.

e) Sub-group analysis

As described in Section V, we estimate heterogeneity of impact on the primary outcomes for three subgroups specified in the pre-analysis plan – province or region within each country; time since the program ended; and the existing quality of the health care system. Overall, we see no systematic discernible variation of impacts within any of these three sub-groups (see Appendix A8 for details).

VII. Unpacking the Lack of Impacts

In this section, we use qualitative data on the program and participants' response to it to explore possible explanations for the lack of statistically significant impacts. We first explore methodological factors that could explain the lack of impacts and conclude that these are unlikely reasons. We then use a simple framework linking participation in the T4D program to MNH outcomes and qualitative evidence from interviews, focus groups, and observations to trace the causal paths from participation to impact. This helps us discard some explanations for the lack of average impacts, and offer evidence suggesting others that are more likely.

(a) *Methodological explanations*

One possibility is that the analysis was statistically underpowered. The 95% confidence intervals of the impact estimates presented above suggest that this is unlikely. In Indonesia, these intervals suggest an ability to rule out impacts of 0.1 standard deviations or more for five of the eight outcomes and impacts of 0.2 standard deviations or more for the remaining three. In Tanzania, the confidence intervals suggest an ability to rule out impacts of 0.1 standard deviations or more for four of the 10 outcomes, impacts of 0.15 standard deviations or more for another four outcomes, and impacts of 0.2 standard deviations or more for the remaining two. We conclude that the analysis was sufficiently powered to detect relatively small effects for most outcomes.

A second possibility is that too much or too little time had elapsed between the conclusion of the program and the endline survey a year and a half later. If participants made noticeable improvements to their access to quality care during or immediately after the program but these improvements faded over time, perhaps the care for a cohort of infants conceived closer to the program's conclusion might have been more improved than the care for later cohorts. Our analysis suggests this is unlikely. If it were the case, we would have expected the sub-group analysis to indicate impacts of the program that were different for mothers who gave birth closer to the date of the survey (i.e. between 0-6 months) than those who gave birth later (i.e. between 6-12 months), which is not what the data suggest. An analogous concern is that *too little* time had elapsed between the program and the end-line surveys. Although we cannot rule out this possibility, end-line surveys in both Indonesia and Tanzania commenced approximately 21 months after the completion of the first two program meetings, a period long enough for the birth of a new cohort of infants and to see most short- and medium-term effects of participants' planned activities.

Overall, given the above evidence, the identification strategy and the evidence from other qualitative data, we conclude that it is highly improbable that we failed to detect a causal impact of the program for methodological reasons.

(b) *Exploring the causal paths from participation to impact*

As noted above, the causal pathways by which a transparency and accountability program might lead to improvements in a specific public service are long, complex, and varied. To further explore the lack of average effects from the T4D program, we utilize the framework in Figure 3, which summarizes several possible causal chains that may link participation in the program to improvements in maternal and newborn health as, first, "inputs", leading to "outputs," and then to "intermediate outcomes," "service outcomes," and finally "health outcomes." Column A represents the T4D program: the "inputs" offered to each community in the treatment group. Column B represents any activities that participants in the program tried to improve access to high quality maternal and newborn health care services in their communities. Column C represents intermediate improvements in certain aspects of the facility, patients' experiences when seeking care, and increase in demand among the community for care that might lead to overall increases in the use of higher quality health care services (Column D), which might in turn lead to healthier mothers and infants (Column E).

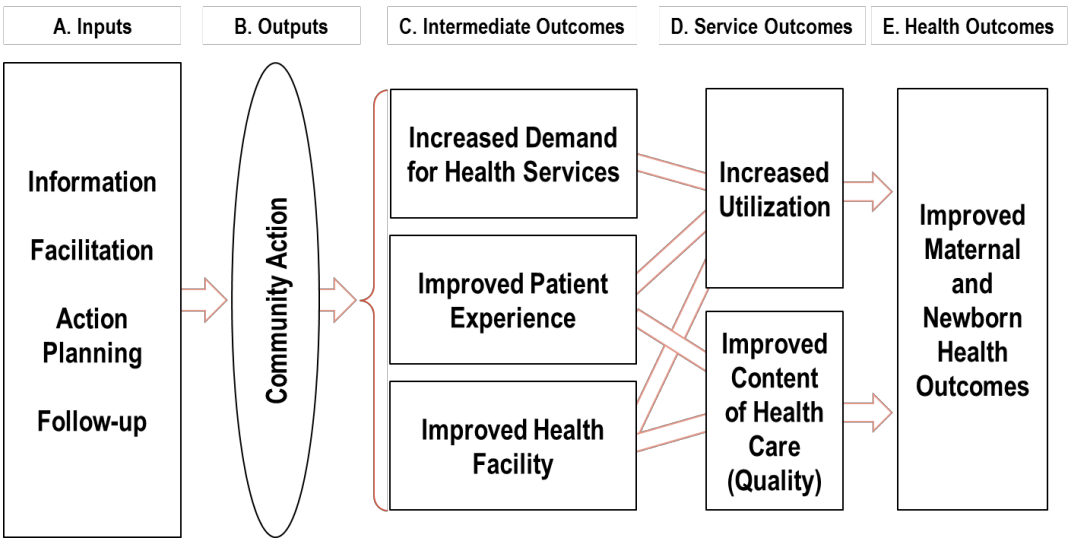


Figure 3. Framework linking participation in the T4D program to MNH outcomes

Sections VI (a)-(b) above describe observational and interview evidence that strongly suggests that in the average community, there were some people who were willing to participate in the meetings and to try the plans of activities that they had designed. In particular, the evidence suggests that the programs were implemented largely as designed in both Indonesia and Tanzania, and that participants in the program meetings planned a large number of activities (Column A of Figure 3). Focus groups, interviews, observations of meetings, and ethnographic studies all indicate that many did in fact try out some of what they had planned (Column B). The remaining evidence described in Section VI also suggests that these efforts did not improve maternal and newborn health care and outcomes in the average community in the treatment group (Columns D and E).

One possible reason is that there were significant secular improvements in maternal and newborn healthcare and outcomes in the years between the program and the endline surveys, which obscured the improvements that participants in the program were able to achieve. Overall, access to quality health care did improve in both treatment and control communities for several

indicators.[32] However, an environment of steady improvements is not a complete or satisfactory explanation of our findings. First, the improvements were not observed across every indicator (e.g., the proportion of underweight sample infants in Indonesia was quite similar at baseline and endline: 14% and 13%, respectively) and overall communities in Tanzania saw much less of an improvement than communities in Indonesia. Second, in nearly all communities where the program was offered, there were still people who thought their care could be even better and were willing to try civic activities to improve it.

Instead, the evidence suggests that a more likely explanation for the lack of difference in intermediate outcomes between communities in the treatment and control groups (Column C) is that participants' activities only rarely added meaningfully to the activities in maternal and newborn health care and outcomes that most communities were experiencing anyway.

First, in nearly all (93.5%) communities, participants' planned activities included efforts to educate, inform, or "socialize" others in their community, suggesting that they believed that lack of knowledge was preventing more pregnant women from seeking care. In both Indonesia and Tanzania, these efforts typically focused on women of childbearing age and young families, and aimed to improve awareness of the importance of seeking maternal and newborn health services in health facilities before, during, and after birth: for example, the importance of seeking antenatal care and birth preparedness planning, the importance of giving birth at a facility or of seeking postnatal care services, or of men about accompanying their spouses to the health facility. Some went door-to-door to convey messages; others organized educational events involving providers

[32] For example, birth at a facility increased from baseline to endline substantially in the overall samples for both countries. It increased from 55% to 74% in Indonesia, and from 56% to 67% in Tanzania. Some of the other key primary outcomes also increased from baseline to endline, but the increases were much less substantial.

from the health facility to speak with women of childbearing age; others asked midwives to integrate education into their monthly outreach services.

Despite their prevalence among planned activities, we find little evidence to suggest that these efforts added significantly to the large number of existing efforts in most communities to improve knowledge of or attitudes toward maternal and newborn health services and standard practices.[33] In Indonesia, about two-thirds of expecting mothers were aware of an education or socialization campaign in their village in both the treatment and the control groups (row (2) of Table 4). In Tanzania, 44% of respondents in the treatment group were aware of these kinds of campaigns, only about 5.6 percentage points (0.1 standard deviations) more than respondents in the control group (row (2) of Table 5).

[33] The ethnographic studies suggest that in some places, participants' efforts may have had some effect on *how* education or socialization activities were run. In particular, their activities seemed to be more participatory, and less prescriptive in terms of telling community members what to do than the typical efforts led by others.

	Potential health activities	Treatment Mean	Control Mean	Difference	p-value	Effect Size	Sample Size
(1)	Total number of potential health activities	5.716	5.570	0.146	0.398	0.046	5999
Proportion of respondents aware of –							
(2)	Socialization campaign aimed at encouraging women to visit health facility	0.652	0.642	0.010	0.597	0.021	5811
(3)	Request for a new ambulance	0.272	0.239	0.033	0.230	0.078	5454
(4)	Attempts to improve the stock of drugs/equipment at the health facility	0.487	0.460	0.027	0.242	0.054	5064
(5)	Attempts to improve the attitude or performance of health facility staff	0.530	0.492	0.039*	0.056	0.077	5312
(6)	Public posting of the cost of service at the health facility	0.188	0.185	0.003	0.852	0.008	5755
(7)	Community members building or requesting a new health facility	0.238	0.251	-0.012	0.471	-0.029	5466
(8)	Attempts to improve health facility infrastructure	0.483	0.474	0.009	0.680	0.018	5518
(9)	Improvement to the road leading to the health facility	0.627	0.629	-0.003	0.910	-0.005	5811
(10)	Attempts to reduce the cost of mother and child health services	0.307	0.309	-0.002	0.919	-0.004	5600
(11)	Creation of a community savings group	0.069	0.044	.0246*	0.072	0.120	5666
(12)	Improvements to the posyandu	0.678	0.652	0.026	0.225	0.054	5664
(13)	Community organized transportation to a health facility	0.085	0.087	-0.002	0.911	-0.005	5754
(14)	Hygiene or cleaning campaign	0.482	0.445	.0374*	0.086	0.075	5769
(15)	Partnership between midwives and baby dukun	0.622	0.658	-0.035	0.153	-0.075	5577
(16)	Additional staff allocated to this village or the health facility	0.471	0.466	0.005	0.837	0.011	4959
	Number of Respondents	3016	2985				
	Number of villages	100	100				

Notes: Treatment means are regression adjusted. *** p<0.01, ** p<0.05, * p<0.1

Table 4. Indonesia: Community awareness regarding potential health activities in their village

	Potential health activities	Treatment Mean	Control Mean	Difference	p-value	Effect Size	Sample Size
(1)	Total number of potential health activities that respondent was aware of (ranging from 0-15)	4.679	4.247	0.432***	0.010	0.141	6003
Proportion of respondents aware of –							
(2)	Socialization campaign aimed at encouraging women to visit health facility	0.438	0.382	0.056**	0.010	0.114	5838
(3)	Creation of a new bylaw relating to mother and baby health	0.358	0.323	0.036	0.179	0.076	5820
(4)	Attempts to improve the stock of drugs/equipment at the health facility	0.289	0.286	0.003	0.877	0.007	5430
(5)	Attempts to improve the attitude or performance of health facility staff	0.275	0.265	0.010	0.612	0.022	5359
(6)	New complaint or suggestion box at the health facility	0.416	0.348	0.068**	0.025	0.143	5408
(7)	Community members building or requesting a new health facility	0.425	0.362	0.064*	0.054	0.132	5764
(8)	Attempts to improve health facility infrastructure	0.455	0.432	0.023	0.338	0.047	5778
(9)	Improvement to the road leading to the health facility	0.398	0.365	0.033	0.195	0.068	5880
(10)	New mobile clinic or other outreach services from the health facility	0.285	0.262	0.023	0.425	0.052	5910
(11)	Creation of a community savings group	0.185	0.159	0.025	0.159	0.069	5883
(12)	Construction of a placenta pit	0.474	0.440	0.033	0.241	0.067	5391
(13)	Registry of men who do not support their wives in accessing health services	0.062	0.054	0.009	0.304	0.038	5696
(14)	Creation of a maternity home for women to wait near the health facility	0.209	0.183	0.025	0.120	0.066	5877
(15)	Campaigns aimed at educating TBAs	0.382	0.356	0.026	0.239	0.055	5761
(16)	Additional staff allocated to the dispensary or health center	0.269	0.274	-0.005	0.807	-0.012	5772
	Number of Respondents	2971	3037				
	Number of villages	100	100				

Notes: Treatment means are regression adjusted. *** p<0.01, ** p<0.05, * p<0.1

Table 5. Tanzania: Community awareness regarding potential health activities in their village

Two-thirds of the activities that participants planned were not focused on education; they instead sought other improvements in their access to quality maternal and newborn health care services. But according to the household survey data, expectant mothers in the treatment communities were by and large no more aware than expectant mothers in the control communities of health activities in their communities (Tables 4 and 5).

Moreover, the evidence suggests that most of these activities did not succeed. Ethnographic studies and interviews in a sample of villages in the treatment group with participants, those with whom they tried to engage as they pursued their activities, and other members of their communities suggest that these activities were successful in only about two-fifths of the villages. And in fewer still were participants later able to recall any tangible improvements in their access to quality maternal and newborn health care that they had achieved as a result of their efforts. During the endline focus group discussions in all villages in the treatment group (a year and a half after the program ended), participants were asked if they thought their activities had improved the health care available to mothers and infants in their communities. Participants in 41% of communities in Indonesia and 30% in Tanzania described achieving at least one tangible outcome from their efforts (Table 6): for example, a new ambulance, a generator, or other equipment for the health facility, a new health facility or easier access to an existing one, or healthcare staff who had come to reside in the village.[34] The percentage of communities in which participants recalled achieving tangible improvements from at least two of their efforts—suggesting that their efforts may have led to broader improvements in their care rather than one-off successes—is still lower: 14% in Indonesia, and 4% in Tanzania. Participants in a similar

[34] Considering that the broader environment was one of steadily improving maternal and newborn health care, perhaps some of these improvements would have happened anyway, without the efforts of the participants.

proportion seemed to think that their efforts had led to improvements in their access to quality care. When asked to reflect on whether maternal and newborn health care had improved in their community because of their activities overall, participants in only 12% of the communities in Indonesia and 4% in Tanzania described tangible improvements.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	% villages that designed at least one non-education action	% villages that attempted at least one non-education action*	% of villages that completed at least one non-education action	% villages that completed at least one non-education action*	% villages where at least one non-education action was successful*	% villages where CRs recalled at least one tangible outcome 1.5 years later	% villages where CRs recalled at least two tangible outcomes 1.5 years later
Indonesia	*100%*	*93%*	*84%*	*59%*	*44%*	*41%*	*14%*
Tanzania	*100%*	*100%*	*91%*	*83%*	*46%*	*30%*	*4%*

Notes: * Estimates based on a sample of villages in the treatment group. Sources as follows: (1) Based on participants' plans from all villages (columns 1 and 3); key informant interviews and ethnographic studies in 41 villages in Indonesia and 24 in Tanzania (columns 2, 4, and 5); and responses in endline focus group discussions with participants in all villages in the treatment group to the question "In the end, what was the outcome from this activity?" (columns 6 and 7). Minor discrepancies in interviews about planned activities led to dropping 11 activities from these proportions (5 in Tanzania, 6 in Indonesia).

Table 6. From plans to outcomes

The diversity of villages and actions included in our sample makes it unlikely that there is a single explanation for why the actions planned and executed by participants did not generate more of an impact on health outcomes. Here we consider several factors that may jointly help to explain the disconnect.

First, it is possible that participants or their communities did not think their maternal and newborn health care needed further improvement. As noted, the general context turned out to be one in which maternal and newborn health care would improve substantially already, even without the efforts of participants in this program. Baseline surveys with recently pregnant women also suggest that the context was one in which maternal and newborn health care services were

perceived by many to be performing well even before the program began. In 79% of the communities in the treatment group in Indonesia, more than three quarters of recently pregnant women told interviewers at baseline that the respect their provider had shown them during their most recent pregnancy was good or excellent; 68% said that the availability of drugs and equipment was good or excellent; and 79% that the overall quality of the care they had received was good or excellent. (Most also seemed to value it highly.[35]) Experiences with quality care seemed less common in Tanzania, though among communities in the treatment group more than three quarters of recently pregnant women told interviewers that the respect the provider had showed them during their most recent pregnancy was good or excellent in 67 of the 100 communities; that the availability of drugs was good or excellent in 23, and that the overall quality of care they had received was good or excellent in 39.

Yet there is little evidence of consensus in most communities at baseline that their access to quality care did not need improvement. Instead, baseline surveys suggest that the particular ways in which recently pregnant women thought their care could be improved differed widely, one reason that participants in the meetings may have planned such a wide variety of activities to try to improve it (Table 2 above). Although baseline surveys of recently pregnant women suggest that the three aspects of care described in the previous paragraph were perceived to be working well in most places, at least three quarters of recently pregnant women told interviewers that *all* three of these aspects of their recent care were good or excellent in only half of communities in Indonesia

[35] For example, in Indonesia, prior to the program, a majority of recently pregnant women in only 11% of communities (22 of 200) told interviewers that it was not important for a woman who was experiencing no complications with a pregnancy and who had already had a baby without complications to seek antenatal care. In 16% of communities, a majority told interviewers that it is fine for a woman in labor to wait to go to the health facility until she is having complications, and in 26% a majority told interviewers that it is just as safe to deliver at home with a traditional birth attendant as to deliver in a facility. A majority offered interviewers all three of these views in only 4% communities. More than three quarters of recently pregnant women offered these views to interviewers in only 2%, 2%, and 4% of communities respectively.

and 13 communities in Tanzania. When asked about seven aspects of the quality of care during their recent birth, there were only 40 communities in Indonesia and 8 in Tanzania in which even a simple *majority* of recently pregnant women said that they thought that all seven were good or excellent. The surveys also indicate some doubts about the importance of receiving care in a facility unless the mother was experiencing complications, one reason that many participants may have tried to inform and educate about the importance of receiving care from the facility during pregnancy and birth (Table 2 above).[36] And as noted, in nearly all communities where the program was offered, there was still an average of 12 people who thought their care could be even better and were willing to participate in meetings to try to improve it. Indeed, observations of meetings in 81 communities in the treatment groups suggest that at least some participants seemed skeptical of the importance of improving their maternal and newborn health in only 10 (5 in Indonesia and 5 in Tanzania), and that in only 2, participants discussed no stories of local examples of problems with maternal and newborn health or the need to improve it.

Rather than a lack of perceived need for further improvement, we propose three further possibilities why participants in most communities were not able to achieve measurable improvements in their access to high quality care. First, it is possible that the activities that participants tried would not have improved health outcomes even if they had led to more measurable or concrete improvements in the community's access to quality maternal and newborn health care, because the link between these improvements and health outcomes was too weak or

[36] In Indonesia, in 51% of communities, more than a quarter told interviewers that it was not important for a woman who was experiencing no complications with a pregnancy and who had already had a baby without complications to seek antenatal care. In 52% of communities, more than a quarter told interviewers that it is fine for a woman in labor to wait to go to the health facility until she is having complications, and in 46% more than a quarter told interviewers that it is just as safe to deliver at home with a traditional birth attendant as to deliver in a facility. At least one recently pregnant woman offered these views to interviewers in 92%, 93%, and 86% of communities respectively.

indirect. For example, some planned activities were cosmetic changes at the health facility such as planting a garden or cleaning the premises, rather than more systemic changes with the potential to lead to transformative improvements. Indeed, in endline surveys of recently pregnant women, there is little correlation between many of the aspects of care that participants in the T4D program tried to improve—such as whether staff were present or reachable, and the delivery room was clean, well-ventilated, with soap and water and no mold or dust—and whether or not mothers and children were healthier or used the facility more around birth. Further, many participants may not have approached people in positions of authority, who might have helped them make more measurable or concrete improvements, because of a lack of connections or relational resources.

A second possibility, strongly supported by the ethnographic studies, is that history of other development projects in these places caused people there to have certain expectations of how these types of projects work – for instance, expectations of payment in return for participating – that led at least some to go along with the meetings the facilitator held but "wait and see" before following through beyond participation in the meetings and initial attempts at the activities that they had designed.

A final set of factors relate to the design of the program. A few key design features are worth reiterating. The program was designed to be: (i) non-prescriptive (it provided information to communities about maternal and newborn health care without suggesting any particular course of action to improve it), (ii) community-driven (it emphasized the importance of the community using their existing knowledge and capacities to understand and fix problems) and, (iii) devoid of external resources (it offered no new materials, technical support, or relational resources). Finally, the program was designed to be relatively light-touch and scalable so that it could be offered consistently in diverse communities across large regions of two countries with varied economies,

politics, and health systems. As our assessment of Columns A and B above reveals, a program designed with these goals did create the space for many communities to leverage local knowledge and collectively plan widely varied courses of action for improving their access to quality maternal and newborn health care services. However, these activities appear in most cases to have had indirect or weak links to health outcomes or to have been insufficient to overcome some of the contextual factors cited above, at least over and above the improvements in health care that were already happening.

VIII. Conclusion

In recent years, transparency and accountability programs have been used increasingly in attempts to improve welfare in developing countries. In this paper, we evaluate the impact of a non-prescriptive, community-driven, health-focused transparency and accountability program in Indonesia and Tanzania. We find no evidence of average impact on healthcare, health, or perceptions of empowerment among recently pregnant women in communities who participated.

Notwithstanding the many differences between the two countries where this program was offered – notably in terms of resource levels and healthcare provider choice – we also find substantial similarities in the reasons why the program did not have an average impact. Using qualitative evidence to trace a simple framework depicting the causal links between the program and health outcomes, we observe that the missing links in these causal chains were similar in both countries. We find that in both countries, participation in the program was substantial and sustained in most communities, and that participants planned and tried a wide variety of civic activities to try to improve their maternal and newborn health care. However, we also find that the efforts at

informing and educating neighbors, which were common in both countries, were often small, localized or one-off community events that did not add much knowledge or dramatically increase use of healthcare among pregnant women. Among the wide range of other types of activities that participants planned, only a minority in both Indonesia and Tanzania a year and a half later recalled that their efforts had translated into a tangible improvement, such as an ambulance or a new or improved facility or pharmacy, new staff, or more privacy. In future work, we plan to explore whether some participants in the program were actually able to succeed in achieving tangible outcomes from their efforts and to understand why some were more successful than others in that regard. In parallel work (Kosack et al. forthcoming) we also focus on how the program influenced those who directly participated in the meetings and activities, as opposed to the focus in this paper on the program's effects on average welfare in their broader communities.

Overall, our findings and assessment in this paper highlight the complexity of the paths linking transparency and accountability programs to health outcomes, and leads us to be skeptical that one like the T4D program—a non-prescriptive, community-driven program that provided no additional resources—is sufficient, on average, to empower communities to measurably improve health outcomes across diverse contexts. It is possible that a transparency and accountability program that selected communities or participants differently or offered them more or more varied material or relational resources, facilitation, or support of other kinds to plan and execute actions would have made more of a difference in their capacities to navigate the paths from planning to achieving improvements in health outcomes. Future research into improving well-being or public services through community-led accountability would benefit from exploring these possibilities further.

References

3ie. (2018). “Does Community-Driven Development Build Social Cohesion or Infrastructure?” Governance Working paper brief. London: International Initiative for Impact Evaluation.

Andrabi, T., Das, J., & Khwaja, A.I. (2017). Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets. American Economic Review, 107(6), 1535-1563.

Banerjee, A.V., Banerji, B., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. American Economic Journal: Economic Policy, 2:1, 1-30.

Banerjee, A., Hanna, R., Kyle, J., Olken, B. A., & Sumarto, S. (2018). Tangible information and citizen empowerment: Identification cards and food subsidy programs in Indonesia. Journal of Political Economy, 126(2), 451-491.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. Biometrika, 93(3), 491-507.

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

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

Deaton, A., & Cartwright N. (2016). Understanding and misunderstanding randomized control trials, NBER Working Paper No. 22595.

Dicker, D., Nguyen, G., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., ... & Abdelalim, A. (2018). Global, regional, and national age-sex-specific mortality and life expectancy, 1950–2017: a systematic analysis for the Global Burden of Disease Study 2017. The lancet, 392(10159), 1684-1735.

Fiala, N., & Premand, P. (2017). Social accountability and service delivery: Evidence from a large-scale experiment in Uganda. Mimeo.

Fox, J. A. (2015). Social accountability: what does the evidence really say?. World Development, 72, 346-361.

Fox J. A. (2007). Accountability politics: power and voice in rural Mexico. New York: Oxford University Press.

Freedom House (2018). Freedom in the World 2018 – Indonesia Country Report. Retrieved from: https://freedomhouse.org/report/freedom-world/2018/indonesia

Freedom House (2018). Freedom in the World 2018 – Tanzania Country Report. Retrieved from: https://freedomhouse.org/report/freedom-world/2018/tanzania

Glennerster, R. (2005). Can information catalyze reform? Sierra Leone and rural India. Abdul Latif Jameel Poverty Action Lab, MIT.

Holland, J., & Schatz, F. (2016). Macro evaluation of DFID's policy frame for empowerment and accountability. Annual technical report 2016: What works for social accountability. Available at: http://itad. com/wp-content/uploads/2017/06/EA-Macro-Evaluation-Technical-report-Dec16-FINAL. pdf [Accessed February 6, 2018].

Joshi A, Houtzager PP. (2012). Widgets or Watchdogs? Public Management Review. Mar 2012;14(2):145-162.

Joshi, Anuradha. (2010). Review of Impact and Effectiveness of Transparency and Accountability Initiatives: Annex 1 Service Delivery. Institute of Development Studies.

J-PAL. (2011). Governance Review Paper: JPAL Governance Initiative. Abdul Latif Jameel Poverty Action Lab, MIT.

King, E., Samii, C., & Snilstveit, B. (2010). Interventions to promote social cohesion in sub-Saharan Africa. Journal of development effectiveness, 2(3), 336-370.

Kosack, S., & Fung, A. (2014). Does transparency improve governance?. Annual review of political science, 17, 65-87.

Kosack, S., Bridgman, G., Creighton, J., Tolmie, C., & Fung, A. (2018). Encouraging Participation. Working Paper.

Kruk, M. E., Gage, A. D., Arsenault, C., Jordan, K., Leslie, H. H., Roder-DeWan, S., ... & English, M. (2018). High-quality health systems in the Sustainable Development Goals era: time for a revolution. The Lancet Global Health, 6(11), e1196-e1252.

Lieberman, E. S., Posner, D. N., & Tsai, L. L. (2014). Does information lead to more active citizenship? Evidence from an education intervention in rural Kenya. World Development, 60, 69-83.

Mansuri, G., & Rao, V. (2012). Localizing development: Does participation work?. The World Bank.

McGee R, Gaventa J. (2011). Shifting Power? Assessing the Impact of Transparency and Accountability Initiatives. IDS Working Papers. 2011(383):1-39.

Molina, E., Carella, L., Pacheco, A., Cruces, G., & Gasparini, L. (2017). Community monitoring interventions to curb corruption and increase access and quality in service delivery: a systematic review. Journal of Development Effectiveness, 9(4), 462-499.

Narayan, D., Patel, R., Schafft, K., Rademacher, A., & Koch-Schulte, S. (2000). Can anyone hear us? Voices of the poor. The World Bank.

Olken, B. A. (2007). Monitoring corruption: evidence from a field experiment in Indonesia. Journal of political Economy, 115(2), 200-249.

Pritchett, L., Samji, S., & Hammer, J. S. (2013). It's all about MeE: Using Structured Experiential Learning ('e') to crawl the design space. Center for Global Development Working Paper, (322).

Raffler, P., Posner, D. N., & Parkerson, D. (2018). The weakness of bottom-up accountability: Experimental evidence from the Ugandan health sector. Unpublished manuscript.

Rao, V., & Woolcock, M. (2003). Integrating qualitative and quantitative approaches in program evaluation. The impact of economic policies on poverty and income distribution: Evaluation techniques and tools, 165-190.

Transparency for Development Project Team (2016). Baseline Report. Retrieved from: https://ash.harvard.edu/files/ash/files/baseline_report.pdf

Transparency for Development Project Team (2016). Transparency for Development Intervention Design. Retrieved from: https://ash.harvard.edu/files/ash/files/intervention_design_0.pdf

World Bank (2017). GDP per capita (current US$) from World Bank national accounts data, and OECD National Accounts data files. Retrieved from: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

World Bank (2004). World Development Report 2004: Making services work for poor people. Washington, DC: World Bank.

White, H. (2011). Achieving high-quality impact evaluation design through mixed methods: The case of infrastructure. Journal of development effectiveness, 3(1), 131-144.

You, D., Hug, L., Ejdemyr, S., Idele, P., Hogan, D., Mathers, C., ... & Alkema, L. (2015). Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Inter-agency Group for Child Mortality Estimation. The Lancet, 386(10010), 2275-2286.