Reshaping Institutions: Evidence on Aid Impacts Using a Pre-Analysis Plan

Katherine Casey
Stanford GSB

Rachel Glennerster
Abdul Latif Jameel Poverty Action Lab

Edward Miguel
University of California, Berkeley and NBER

FIRST DRAFT: APRIL 2011
THIS DRAFT: JUNE 2012

Abstract: Despite their importance, there is limited evidence on how institutions can be strengthened. Evaluating the effects of specific reforms is complicated by the lack of exogenous variation in institutions; the difficulty of measuring institutional performance; and the temptation to "cherry pick" estimates from among the large number of indicators required to capture this multi-faceted subject. We evaluate one attempt to make local institutions more democratic and egalitarian by imposing participation requirements for marginalized groups (including women) and test for learning-by-doing effects. We exploit the random assignment of a governance program in Sierra Leone; develop innovative real-world outcome measures; and use a pre-analysis plan (PAP) to bind our hands against data mining. The intervention studied is a "community driven development" program, which has become a popular strategy for foreign aid donors. We find positive short-run effects on local public goods and economic outcomes, but no evidence for sustained impacts on collective action, decision-making, or the involvement of marginalized groups, suggesting that the intervention did not durably reshape local institutions. We discuss the practical tradeoffs faced in implementing a PAP, and show how in its absence we could have generated two divergent, equally erroneous interpretations of program impacts on institutions.

JEL Codes: F35, H41, O4

*Acknowledgements: We wish to thank the GoBifo Project staff—Minkahil Bangura, Kury Cobham, John Lebbie, Dan Owen and Sullay Sesay—and the Institutional Reform and Capacity Building Project (IRCBP) staff—Liz Foster, Emmanuel Gaima, Alhassan Kanu, S.A.T. Rogers and Yongmei Zhou—without whose cooperation this research would not have been possible. We are grateful for excellent research assistance from John Bellows, Mame Fatou Diagne, Mark Fiorello, Maryam Janani, Philip Kargbo, Angela Kilby, Gianmarco León, Tom Polley, Tristan Reed, Arman Rezaee, Alex Rothenberg, and David Zimmer. David Card, Kevin Esterling, Jim Fearon, Macartan Humphreys, Brian Knight, David Laitin, Ed Leamer, Kaivan Munshi, Ben Olken, Biju Rao, Debraj Ray, Gerard Roland, Ann Swidler, Eric Werker and seminar audiences at Brown, CGD, CEGA, the Econometric Society (North American Summer Meetings), IGC, IMF, MIT, NEUDC, NYU, Princeton, Stanford, SITE, University of British Columbia, U.C. Berkeley, UCLA, UCSD, the Western Experimental Conference, and WGAPE have provided helpful comments. We thank the editor, Larry Katz, and five anonymous referees for excellent suggestions. We gratefully acknowledge financial support from the GoBifo Project, the IRCBP, the World Bank Development Impact Evaluation (DIME) initiative, the Horace W. Goldsmith Foundation, the International Growth Centre, the International Initiative for Impact Evaluation (3ie), and the National Bureau of Economic Research African Successes Project (funded by the Gates Foundation). All errors remain our own.

1. Introduction

Many scholars have argued that the accountability and inclusiveness of government institutions are key determinants of economic performance (Engerman and Sokoloff 1997, Acemoglu, Johnson and Robinson 2001, Banerjee and Iyer 2004). There is no consensus, however, on the reforms that will engender better functioning institutions, or on whether it is possible (or even desirable) for external actors like foreign aid donors to reshape power dynamics in less developed countries. This debate has played out vigorously in discussions of aid policy: while some argue that large infusions of foreign aid can themselves help build stronger institutions (Sachs 2005), others assert that historically rooted institutions and social norms are difficult to understand, let alone transform (Easterly 2006).

Progress towards resolving this question is complicated by the rarity of exogenous changes in institutional structure and the difficulty of measuring institutional performance. The context-specific nature of institutions means that there are few standard indicators to draw from, and reliance on subjective measures risks bias from "halo effects" (see Olken 2009 regarding corruption). Moreover, their multi-dimensionality makes a large number of outcomes potentially relevant, tempting the researcher to "cherry pick" a subset of results that may be statistically significant by random chance. We evaluate one attempt to transform local institutions in Sierra Leone and address these challenges by exploiting a randomly assigned governance intervention, developing objective measures of institutional performance, and using a pre-analysis plan (PAP) to bind our hands against data mining.

The intervention studied, a "community driven development" (CDD) project, provides both what we call "hardware" and "software" support to rural communities. Hardware includes block grants for local public goods, trade skills training and small business start-up capital. Software covers technical assistance that promotes democratic decision-making, the participation of socially marginalized groups, and transparent budgeting practices. The CDD approach attempts to bolster local *coordination*—for example, by setting up village development committees—and to enhance *participation*, by requiring women and "youths" (adults under age 35) to hold leadership positions, sign off on project finances, and attend meetings. The push for CDD reflects a broader intellectual movement in international development towards greater participation and empowerment of the poor (Chambers 1983, Narayan 2002, Sen 1985, 1999, World Bank 2001). CDD further resembles certain "War on Poverty" reforms in the 1960's

United States, particularly the Office of Economic Opportunity's Community Action Program, which bundled social service delivery with attempts to politically mobilize marginalized groups, especially African Americans (Rosener 1978, Germany 2007).[1] Donors currently channel large amounts of aid through these programs: Mansuri and Rao (2012) estimate that the World Bank alone has spent US$50 billion on CDD initiatives over the past ten years.

Advocates of participatory local governance promise a long and varied list of benefits ranging from more cost-effective construction of infrastructure, to a closer match between project choice and village needs, to the weakening of authoritarian village institutions.[2] Critics hold concomitant concerns that participation requirements serve as a regressive tax, widening political participation clogs up rather than expedites decision-making (Olson 1982), and external resources attract new leaders, crowd out the most disadvantaged (Gugerty and Kremer 2008) or are captured by elites if the program is unable to change the nature of *de facto* political power (Bardhan 2002). Any real world program risks manipulation during implementation, and skeptical observers fear that donors simply use the jargon of participatory development for political or public relations purposes while continuing to operate in a "top-down" manner. Few studies provide rigorous empirical evidence regarding these claims (Mansuri and Rao 2004).

Scholars have argued that the incompetence and elite domination of Sierra Leone's institutions – both in the central government and the traditional chieftaincy system – in the 1970's and 1980's were key contributors to the civil war that took place from 1991 to 2002 (Richards 1996, Keen 2003). Emerging from war with widespread poverty and a dearth of public services, the country fell to the very bottom of the United Nations Development Program Human Development Index that measures standards of living, health and education (United Nations 2003), with 2001 per capita income of just US$140 in exchange rate terms (World Bank 2003). To both facilitate recovery from and preclude a return to violence, one of the most high-profile reforms was the reconstitution of elected district-level governments. Housed within the

[1] There is remarkable similarity between CDD programs and the design and framing of these earlier U.S. efforts. Germany (2007: 15) writes that "the OEO pursued an aggressive, innovative and experimental agenda premised on empowering the poor and giving local people significant authority in fighting poverty. Envisioned by OEO administrators as an attack on the causes of poverty more than the symptoms, the War on Poverty was an ambitious effort to reform the psychology of the poor, the institutions of the ghetto, the systems necessary for upward mobility, and the patterns of black political participation." We thank David Card for drawing our attention to these parallels.

[2] For instance, Dongier et al. (2003) write that: "Experience demonstrates that by directly relying on poor people to drive development activities, CDD has the potential to make poverty reduction efforts more responsive to demands, more inclusive, more sustainable, and more cost-effective than traditional centrally led programs…achieving immediate and lasting results at the grassroots level."

government's Decentralization Secretariat and funded by the World Bank, the project we study, "GoBifo" (or "Move Forward" in Krio, Sierra Leone's lingua franca), further extended decentralization by providing financial assistance (of $4,667, or roughly $100 per household) and social mobilization to village-level committees. While the objective of making local government institutions more inclusive aimed to address some of the perceived root causes of the civil war, GoBifo's design is similar to many other CDD projects in non-post-conflict societies.

This paper assesses the extent to which GoBifo achieved its goals of reforming local institutions in rural Sierra Leone, and in so doing makes four contributions. The first general contribution is a discussion of how pre-analysis plans (PAP) can help avoid some common pitfalls in empirical research. The research and project teams agreed to a set of hypotheses regarding the likely areas of program impact in 2005 before the intervention began. As the project came to a close in 2009, we fleshed out this document with the exact outcome measures and econometric specifications we would use, and archived this pre-analysis plan before analyzing the follow-up data (see supplementary Appendix A). "Tying one's hands" in this way is potentially useful where researchers have wide discretion over what they report and may face professional incentives to affirm the priors of their academic discipline or the agenda of donors and policymakers. Explicit *ex ante* agreements between researchers and program sponsors can offer a layer of protection for "inconvenient" findings and thus reduce the scope for tendentious reporting. Adherence to a PAP reduces the risk of data mining or other selective presentation of empirical results ("cherry-picking") and generates correctly sized statistical tests, bolstering the credibility of the findings. More broadly, a system of registration for experimental trials would help round out the body of available research evidence, mitigating the publication bias that arises from underreporting null or counter-intuitive results. Registration of drug trials and pre-analysis plans is required by U.S. law but is uncommon in economics.[3] We hope our experience contributes to the emerging debate on the pros and cons of PAP's in social science.

The second contribution is the creation of novel measures of local institutions and collective action, or "structured community activities" (SCAs). These are concrete, real-world scenarios that allow us to unobtrusively assess how communities: (i) respond to a matching grant

[3] The FDA Modernization Act of 1997 led to the creation of the NIH-sponsored web registry clinicaltrials.gov in 2000, and a 2007 amendment requires results reporting and imposes financial penalties for non-compliance. In 2005, registration of clinical trials became a prerequisite for publication in any member journal of the International Committee of Medical Journal Editors. See Rosenthal (1979), Simes (1986), and Horton and Smith (1999).

opportunity; (ii) make a communal decision; and (iii) allocate a valuable asset among community members. We feel that these SCAs capture local collective action capacity, and uncover the decision-making processes that underlie it, more objectively than lab experiments, hypothetical vignettes or surveys alone.[4] The fact that the SCA's were carried out *after* the GoBifo program ended allows us to measure any persistent impacts on institutional performance.

This paper's evaluation of a CDD project will be of particular interest to development economists and practitioners. We use a randomized experimental design, which produces evidence on causal impacts in a large study sample of 236 villages and 2,832 households. The study's extended timeframe over four years (2005-2009) allows us to assess longer run impacts than is typically possible. While four years may be short relative to the lifetimes over which current institutions emerged, it is not short in comparison to most community development or other externally funded projects. To guide our empirical work, we develop a theoretical framework for understanding how CDD programs might impact local outcomes.

Fourth and finally, we contribute to the growing literature concerning the impacts of giving decision-making authority to marginalized groups. Research in India suggests that political quotas for women and members of scheduled castes shift the composition of public spending towards goods preferred by these groups (Chattopadhyay and Duflo 2004, Pande 2003) and reduces bias against female candidates (Beamen et al. 2009). By contrast, we find that requiring women and young adults to take on leadership positions, participate in project meetings, and sign off on project finances does not have any persistent effect on their participation in local decision-making or attitudes regarding their leadership ability. One explanation for this difference may be that while Indian quotas give members of historically excluded groups real power over sizeable resources within a formal state body (the *panchayat*), CDD takes a more indirect approach to *de jure* reforms—nudging communities towards more inclusion without explicitly challenging elites—and may not change the identity of *de facto* power holders (Acemoglu and Robinson 2008). Perhaps because sidelining the chiefs was not a program goal, chiefdom officials retained as much control over village development committees in GoBifo communities as they held over comparable organizing bodies in control villages.

Our analysis explores a wide range of measures, divided into two broad groups: project implementation, local public infrastructure, and economic outcomes (which we call family A),

[4] The measures of community negotiations Paluck and Green (2009) developed in Rwanda are a related approach.

and institutional and collective action outcomes (family B). We find that the GoBifo project was well implemented: it established village organizations and tools to manage development projects in nearly all cases, and provided the financing to implement them. The distribution of project benefits within communities was largely equitable and the leakage of project resources appears minimal. We further find immediate impacts on the stock and quality of local public infrastructure, such as schools and latrines. There is also more market activity in treatment communities, as well as increases in household asset ownership, suggesting economic benefits.

However, we find no detectable changes in the second, arguably more important, institutional domain (family B). We find no evidence that the program led to fundamental changes in the ability to raise funds for local public goods, decision-making processes, or social norms and attitudes. As an example, despite the experiences many women gained by participating in and managing GoBifo activities, after the project ended they were no more likely to attend or voice an opinion at community meetings. Similarly, there is no evidence that the establishment of a democratic organizing committee or the experience implementing projects led to more fundraising in response to a matching grant opportunity. In all, we find no evidence that the program reshaped village institutions, empowered minorities, or improved collective action beyond the activities stipulated by the project itself. The time horizon of the research over four years suggests that these findings cannot be dismissed simply as the result of a short term study.

The rest of the paper is structured as follows. Section 2 discusses the context, intervention, and theoretical framework. Section 3 covers the research design, pre-analysis plan, and econometric specifications. Section 4 discusses the empirical results and section 5 concludes.

2. Background

2.1. Institutions in Sierra Leone

Before describing the GoBifo program, we first consider why existing institutions in Sierra Leone might warrant reform. The country has a dual system of governance (common in many African countries, Mamdani 1996) in which the central government apparatus based in the capitol runs in parallel to the "traditional" local chieftaincy system, neither of which has historically been particularly democratic or inclusive. Authoritarian central government leaders in the 1970's and 1980's enriched themselves through illicit diamond deals while providing woefully inadequate public services (Reno 1995). President Siaka Stevens dismantled

democratic institutions, initially by abolishing elected district governments in 1972, and ultimately declaring the country a one-party state in 1978. One-party rule continued until the 1992 coup that roughly coincided with the start of the civil war (which ran from 1991 to 2002).

As background on the traditional system, the country's 149 paramount chiefs come from hereditary "ruling houses"; serve for life once appointed or elected (by a restricted electorate); exert considerable control over resource allocation, including land and labor; operate the local court system that presides outside the capital; and organize the provision of many local public goods (such as road maintenance). This system largely excludes both women (who are not even eligible to serve as chiefs in much of the country) and young men from decision-making. Political exclusion, growing frustration with government incompetence and corruption, and grievances against heavy-handed chiefs are seen as destabilizing factors that contributed to the war (Richards 1996, Keen 2003).

2.2 The GoBifo Project

After the war, the Government of Sierra Leone and its donor partners, including the World Bank, launched an ambitious institutional reform agenda, which included the re-establishment of district-level governments. The GoBifo pilot initiative was launched to support and deepen this reform by extending decentralization down to the ward and village levels.[5] The program had two main components: i) financial assistance in the form of block grants to fund local public goods provision and small enterprise development; and ii) intensive organizing to establish new structures to facilitate collective action (i.e. Village Development Committees) and institute participation requirements to elevate historically marginalized groups to positions of authority. As examples of the latter, GoBifo required that one of the three co-signatories on the community bank account be female; encouraged women and youths to manage their own projects (e.g., small business training for youths); made evidence of inclusion in project implementation a prerequisite for the release of funding tranches; and, as part of their internal review process, required field staff to record how many women and youth attended and spoke up in meetings. To formally link project activities to higher tiers of government, Village Development Committees (VDC) were required to submit their village development plans to the appropriate

[5] Wards are the lowest formal government administrative unit, each covering around 10,000 people on average, and the elected district councilor representing the ward chairs the Ward Development Committee. While the project we study also operated at the ward level, only the village-level intervention was randomized and is thus our focus.

Ward Development Committee (WDC) for review, endorsement and onward transmission to the new district councils for approval (GoBifo Project 2007).

The process of establishing new village institutions, training community members, and promoting social mobilization of marginalized groups was intense and accounted for a large part of GoBifo human and financial resources. Specifically, all project facilitators were required to reside in one of the six villages assigned to them and spend approximately one day per week in each of the villages. After the start of project work in January 2006 and through the completion of all village-level projects in July 2009, each village thus received roughly six months of direct "facilitation" over a three and a half year period (see the timeline in Appendix B). Furthermore, while just under half of the total GoBifo budget was dedicated to village- and ward-level block grants (US$896,000 or 47%), the balance covered "capacity development" in village- and ward-level planning (US$589,732 or 30%), project management and contingencies (US$255,320 or 14%), and monitoring and evaluation (US$177,300 or 9%). Thus for every dollar spent directly on grants, roughly one dollar was spent on capacity-building, facilitation and oversight.

Several different types of GoBifo village projects were common. The largest share of projects, at 43%, was in the construction of local public goods, with 14% in community centers or sports fields, 12% in education (i.e., primary school repairs), 10% in water and sanitation (i.e., latrines), 5% in health (including traditional midwife posts), and 2% in roads. Another 26% was in agriculture, including seed multiplication and communal farming; 14% in livestock (i.e., goat herding) or fishing; and 17% in skills training and small business development initiatives (i.e., blacksmithing, carpentry, soap making). Leakage of GoBifo funds also appears minimal: when we asked villagers to verify the detailed financial reports that were given to the research team by project management, community members were able to confirm receipt for 86.5% of the 273 transactions that were cross-checked.[6]

GoBifo is similar to CDD initiatives in other countries. The project implementation stages—establishing a local committee, providing facilitation that aims to shift social norms, and allocating block grants—are standard, as is the pervasive emphasis on inclusive, transparent and participatory processes. Compared to other projects (Olken 2007, Labonne and Chase 2008), the

[6] The discrepancies were of two types: i) the amounts in community records was markedly less than in project accounts; or ii) community members reported receiving building materials in kind and could not estimate their value. For each of the disputed transactions, the GoBifo accounting team produced hard copy payment vouchers signed by both a village representative (either the VDC Chair or Finance Officer) and a project field staff member.

most notable difference is that the village-level component of GoBifo did not involve any inter-community competition for funding. Regarding the scale of funding, GoBifo disbursed grants worth a bit under $5,000 to communities with 50 households, or 300 residents, on average (so roughly $100 per household, or $4.50 per capita annually over three and a half years).[7]

2.3 A Framework of Collective Action and External Aid

We next lay out a stylized local collective action framework that clarifies how an external intervention that provides financing and participation requirements might change local decision making, and derive implications that inform the empirical analysis; see Appendix C for the formal exposition of the model. In the model, a social planner determines the optimal investment in local public goods and sets a corresponding tax schedule, which is implemented with perfect compliance. Individual residents then decide whether or not to voluntarily participate in the planning and implementation of the public goods projects, taking their individual tax burden as given. We feel this framework is a reasonable approximation to the context of rural Sierra Leone (and similar societies with strong village headmen), where the traditional chief has the authority to levy fines and collect taxes to provide basic public goods, but there is variation in how involved residents are in actual decision making and implementation. In this setting, the external intervention lowers the *marginal* cost of local public goods provision through financial subsidies, and affects the *fixed* costs of collective action by imposing participation requirements and instilling democratic norms. We allow underrepresented groups (i.e., women) to have differential participation costs *ex ante*, which could be impacted by learning-by-doing or demonstration effects during project implementation.

We define three time periods that correspond to our data collection activities: the pre-program period when the baseline survey was fielded; the program implementation phase, where the first follow-up survey captured activities that had been completed during the intervention (and launched the structured community activities); and the post-program period, where the second follow-up survey explored what happened with the SCAs after the project had finished. Since the marginal cost reductions are tied directly to external financial assistance, while the fixed organizing cost reductions could be internalized and maintained, we can speculatively gain

[7] The Fearon et al. (2009) Liberia project provided roughly $20,000 to "communities" that comprised two to three thousand residents, so roughly $4 per capita annually over two years.

some leverage over which channels are at work by comparing impacts during the project versus post-program phases. Moreover, studying the post-program period allows us to evaluate the persistence and "sustainability" of impacts.

First consider the individual's decision of whether to contribute time and voluntary labor to the planning and provision of local public goods. While these decisions are taken in a decentralized fashion, they aggregate in a way that affects the cost of public goods provision facing the social planner. The fact that individuals ignore the aggregate effect of their voluntary labor captures the classic externality feature of collective action, and implies that even with perfect tax compliance, the planner will still fail to achieve the first-best level of public goods.

Individuals gain utility from consumption of the current stock of public goods, private consumption, and a psychic or social benefit of participating in collective action that captures the intrinsic value of civic involvement. Regarding the latter, Olken (2010) and Dal Bó et al. (2010) provide evidence that having a say in the decision-making process can have a large effect on satisfaction and cooperation even if the choice process has zero impact on the final policy outcome *per se*. Given historical legacies of exclusion, we assume that while some women and youth may derive positive utility from participation, they face additional social costs of speaking up and thus, on average, their net benefits of civic participation are lower than for elder male elites. All residents face the same opportunity cost of participating, which reflects the cost of time spent engaging in public goods provision instead of wage-earning activities, and must pay the tax set by the social planner. The first order conditions imply that the individual chooses to participate in collective action if and only if the net benefits are nonnegative.

The social planner chooses the level of local public goods investment with the objective of maximizing the sum of individual utilities. The cost of public goods provision has two components: a marginal cost capturing the price of construction materials, and a fixed coordination cost of collective action, which is a function of both the sum of individual participation decisions and the capacity of local institutions. Following the theory motivating participatory local development, we assume that the fixed costs of collective action are falling in both the capacity of local institutions and community participation; we assess the empirical validity of these assumptions below. The latter condition would be true if, for example, greater community involvement made public goods provision easier by creating greater support for the process. Importantly, even if participation has no effect on coordination costs, advocates argue

that local civic engagement carries intrinsic benefits, and therefore project participation belongs in the utility function and its enhancement becomes an appropriate objective for intervention.

Standard first order conditions imply that the planner chooses the optimal level of local public goods investment if affordable, or a smaller investment that exhausts the village budget (at a corner solution) if it is not. Given the poverty and extremely limited public services in rural Sierra Leone, it seems reasonable to assume the latter, where communities face a binding budget constraint that keeps public investment well below optimal levels. This means that there are plenty of public investments—in latrines, water wells, primary schools—whose village-wide marginal benefits exceed the marginal cost of construction, yet are simply unaffordable given the community's small tax base and inability to borrow (in light of pervasive financial market imperfections). Under these constraints, profitable investments become unaffordable because construction prices and/or coordination costs are prohibitively high.

Within this framework, participatory local governance interventions aim to have three distinct impacts. First, by subsidizing the cost of construction materials, the financial grants reduce the marginal cost of public goods provision. Second, the leadership quotas and participation requirements for women and youth aim to increase the benefits of participation for these historically marginalized groups. Such requirements should automatically translate into greater participation in collective activities during project implementation for these groups. Moreover, if women and young men learn-by-doing, or if their participation exerts positive demonstration effects on others that begin to shift social norms, this experience could trigger a persistent increase in their benefits of participation, sustainably raising participation levels into the post-program period. Third and finally, this increase in community participation, accompanied by the establishment of village development committees, plans and bank accounts, aims to reduce the fixed coordination costs of collective action. The idea is that once these are in place, the next village project should be less costly to identify and execute, both during project implementation and the post-program period. As such, the original GoBifo project proposal emphasizes the sustainability and broad mandate of these new structures, suggesting they will become "the focal point for development interventions" in the future (World Bank 2004).

This simple framework generates three empirical predictions to take to the data. First, the combination of financial subsidies and lower coordination costs should unambiguously increase public goods investment during the program implementation phase. To assess this, outcome

family A includes project implementation indicators to first evaluate whether the grants were in fact delivered to villages and new institutions established on the ground, and then a set of measures regarding the stock of local public goods to assess immediate impacts on investment levels. Second, as we move from project implementation to the post-program period, the marginal investment costs return to baseline levels while the fixed costs (potentially) remain reduced. To evaluate whether new village institutions lead to greater public investment in the post-program period, family B includes take-up of the building materials vouchers (in SCA #1), and other collective action measures beyond the direct program sphere. Third, if participation requirements for women and youth trigger a permanent enhancement in their benefits from participation, we should see more women and youths attending community meetings and taking part in decision-making post-program. This is captured by the outcomes in the gift choice component of SCA #2 and household survey responses concerning civic engagement in non-program areas. Moreover, enhancing participation by marginalized groups could initiate broader changes in social norms and attitudes (for instance, regarding the desirability of female leadership), as captured in several additional hypotheses under outcome family B. It remains an empirical question whether any of these predictions hold in reality, hence we turn to the data.

3. Research Design

3.1. Random Assignment

The 118 GoBifo treatment and 118 control villages were selected from a larger pool of eligible communities using a computerized random number generator. They were sampled from within the two study districts, which were chosen to strike a balance in terms of regional diversity, political affiliation, and ethnic identity, while simultaneously targeting poor rural areas with limited NGO presence (see Appendix D for a map). Bombali district is located in the Northern region dominated by the Temne and Limba ethnic groups and traditionally allied with the All People's Congress (APC) political party, one of Sierra Leone's two largest parties. Bonthe district is in the South, where the Mende and Sherbro ethnic groups dominate and where the other major party, the Sierra Leone People's Party (SLPP), is strong. Using the 2004 Population and Housing Census, the pool of eligible villages was restricted to those considered of appropriate size for a CDD project, namely between 20 and 200 households in Bombali and 10 to 100 households in Bonthe (where villages are smaller), and once the sample was chosen, the

villages were randomized into treatment and control groups, stratifying on ward.[8] There were 6 treatment and 6 control villages in each of 19 wards, plus one additional ward on Bonthe Island, where there were only 4 treatment and 4 control communities given the small size of the ward.

Statistics Sierra Leone staff randomly selected twelve households to be surveyed from the Census household lists in each village. Given interest in the dynamics of political exclusion and empowerment, the choice of respondent within each targeted household rotated among four different demographic groups in each subsequent household surveyed: non-youth male, youth male, non-youth female and youth female. All respondents are at least 18 years old, and note that the Government of Sierra Leone's definition of youth includes people up to 35 years of age (although the definition is a bit subjective in reality, especially since many Sierra Leoneans do not know their exact age). This data collection strategy means that for each community, and for the overall sample, responses are roughly balanced across the four demographic groups.[9]

The randomization procedure successfully generated two groups balanced along observable dimensions. Specifically, Table 1 lists the control group mean and the treatment minus control pre-program difference for a variety of community characteristics (including total households, distance to nearest road, average respondent years of education, and indices for civil war exposure and local history of domestic slavery) as well as an illustrative selection of pre-program values for outcome measures. There are no statistically significant mean differences across the treatment and control groups for any of these variables; Appendix F presents the same estimates for all 96 baseline measures and shows that the difference across treatment and control groups is significant at 90% confidence for only seven of these, roughly as expected by chance. One noteworthy pattern in the baseline data is the stark gender difference in local meeting involvement, with twice as many males (59%) than females (29%) speaking at village meetings.

3.2 Data Collection and Measurement

[8] We ran 500 computer randomizations and saved all resulting assignments that generated no statistically significant differences (at 95% confidence) between treatment and control groups in terms of the total number of households per village and the distance to the nearest road. Among these "balanced" assignments, one was then selected at random for the final treatment assignment. Following Bruhn and McKenzie (2009), we include the "balancing" observables in the regression analysis as covariates to generate correct standard errors. Treatment effect estimates are thus interpreted as impacts conditional on these observables, although results do not change with their exclusion (not shown). There were two minor data issues that led to a partial re-sampling of a small number of villages, however these did not affect the integrity of the randomization (see supplementary Appendix E).

[9] These four demographic groups each comprise roughly a quarter of the adult population in these two districts in the 2004 Census (ranging from 21 to 31%), indicating that our sample is quite representative.

This analysis draws on three main data sources: household surveys from late 2005 (baseline) and mid-2009 (follow-up); village-level focus group discussions held in 2005 and 2009; and three novel structured community activities (SCAs) conducted in late 2009 shortly after GoBifo activities had ended. The SCAs were introduced with the initial follow-up survey in May 2009 and then followed up in an unannounced visit five months later. The research team and enumerators were operationally separate from GoBifo staff at all stages of the project.

The 2005 household surveys collected data on baseline participation in local collective activities, as well as household demographic and socioeconomic information. To establish a panel, the field teams sought out the same respondents during the 2009 follow-up surveys, and the attrition rate was moderate: 96% of the same households were located and re-interviewed, as were 76% of the same individual respondents. Where the individual respondent from 2005 was unavailable, we picked another household member with the same gender and youth status (or same gender only, if no match on both criteria was available) to interview in 2009. This approach maintained the overall demographic composition of the respondent sample. In the 4% of cases where the entire household had moved permanently, we visited the dwelling located three doors down and interviewed someone with the same gender and youth status. Rates of attrition at both the individual and household level are balanced across treatment groups and do not vary significantly by treatment status interacted with several baseline characteristics including respondent gender, youth status, education, community meeting attendance, or household assets (see Appendix Table G). Note that our main analysis is conducted (and many of our outcome measures are collected) at the village level, which is the unit of treatment assignment and for which we have zero attrition.[10]

During the data collection visits in 2005 and 2009, the field team supervisor assembled key opinion leaders—including VDC members, the village chief, as well as women and youth leaders, among others—to describe the condition of local infrastructure and answer questions about local collective processes and activities. Research supervisors also made their own physical assessments of construction quality as a cross-check.

Given the difficulties in gauging institutional dynamics and collective action through survey responses alone, the third main type of data was gathered through the SCAs. These were

[10] For the outcome variables that rely on household-level responses, we construct village level averages using all individuals interviewed in the follow-up survey. However, none of our results are affected by limiting the sample to the original respondents who were resurveyed (see Table 3).

designed to measure how communities respond to three concrete, real-world situations: (i) raising funds in response to a matching grant opportunity; (ii) making a community decision between two comparable alternatives; and (iii) allocating and managing an asset that was provided for free. As opposed to hypothetical vignettes or laboratory experiments in the field, these exercises more directly, realistically and less obtrusively capture outcomes of interest. We discuss each SCA in detail here.

SCA #1 was designed to measure whether GoBifo produced persistent effects on villages' capacity for local collective action beyond the life of the project. Each community received six vouchers they could redeem at a nearby building materials store (in the nearest large town) if they raised matching funds. Specifically, each voucher was worth 50,000 Leones (roughly US$17) only if accompanied by another 100,000 Leones (US$33) from the community. Matching all six vouchers generated 900,000 Leones (US$300) for use in the supply store.

Since individuals had negligible savings and faced credit constraints, take-up of the vouchers is a measure of local capacity for cooperation. Voucher redemption was recorded by clerks at the building materials stores. Enumerators returned to all villages five months after the initial distribution of the vouchers to assess the distribution of project contributions and benefits (i.e., did they buy metal for a new roof for the primary school or for the chief's home?), the quality of final construction, and how inclusive and transparent the management of the resulting project had been. In the context of the model, higher take up in treatment communities implies that the program persistently reduced the *fixed* costs of collective action, as in this case the marginal component (i.e. the financial subsidies offered through the vouchers) was exactly the same for treatment and control villages.

Take-up of all the vouchers was always in the community's self-interest: given the subsidy (and even accounting for transport costs), the materials could be profitably resold immediately after purchase at the building material stores. To provide a sense of what types of projects this amount (US$300) could fund, the modal project was to purchase metal sheeting to upgrade the roofing on a community building like a school. In earlier GoBifo projects, villages were free to divide the funds between multiple projects, and roughly 20% of all projects were valued at or below US$300, indicating that this is a useful amount of funding.

SCA #2 was designed to measure the extent to which community decision-making is democratic and inclusive, and to assess the level of community participation. The day before

survey work, the enumerator teams met with the village head (the lowest level chiefly authority) and asked him to assemble the entire community for a meeting the next morning. At the subsequent meeting, the enumerators presented the community with a choice between two gifts each valued at roughly US$40—a carton of batteries (useful for radios and flashlights) versus many small bags of iodized salt—as a token of appreciation for participating in the research. We did extensive field piloting to identify two gifts between which community members would be largely indifferent and for which there was no normatively "correct" choice. The piloting suggested that there was more discussion when it was not obvious *ex ante* which option was preferred. (While it was not an outcome of interest in terms of program impacts, two thirds of the communities chose salt and one third the batteries in both the treatment and control groups.)

The enumerators – who were Statistics Sierra Leone employees and not GoBifo staff – emphasized that the community itself should decide how to share the gift and then withdrew from the meeting to observe the decision-making process from the sidelines. The enumerators remained "outside" the community meeting circle and recorded how the deliberation evolved without making any comments of their own. Among other things, the enumerators recorded who participated in any side-meetings; the degree to which the chief, village head and elders dominated the discussion; the extent of debate in terms of time and the number of comments; and a subjective assessment of the apparent influence of different sub-groups (e.g., women) on the final outcome. This exercise provided quantitative data on the relative frequency of female versus male speakers, and youth versus non-youth speakers in an actual community meeting.[11] Note that these are exactly the same metrics that the GoBifo facilitators were required to track as part of their internal impact assessments (GoBifo Project 2008).

SCA #3 was designed to gauge the extent of elite capture of resources, a common concern for decentralization reforms. During the first follow-up visit in 2009, the enumerators gave each village a large plastic tarpaulin sheet as a gift. Tarpaulins are frequently used in Sierra Leone as makeshift building materials (40% of households have potentially leaky thatched roofs), and in agriculture as a surface for drying grains (as fewer than a quarter of villages have a functional drying floor). During the second 2009 follow-up visit five months later, enumerators recorded which households had used the tarpaulin in the intervening period. This activity also

[11] Of the four enumerators, one focused his data collection on the participation of youths, one on women, one on all adults, and the fourth kept careful track of each person who spoke publicly.

captures an element of collective action, as enumerators assessed whether villages had been able to decide on a use for the tarp, and whether it had been put mainly towards a public (e.g., a communal grain drying floor) or private end (patching the roof of an individual's home).

We developed the SCA's precisely because we felt traditional survey measures of collective action and participation were potentially unreliable. Thus "validating" the SCA's using more standard measures is potentially problematic. Nevertheless, documenting a positive correlation between the SCA's and existing measures would provide some reassurance that there is an underlying "signal" of collective action capacity that is picked up by both. To provide suggestive evidence on the relevance of the SCA's, we selected variables from the baseline data that sought to assess the same concepts and tested whether they predicted SCA outcomes four years later. For SCA #1 concerning collective action capacity, Appendix Table H shows that the number of vouchers redeemed is positively and significantly predicted by the number of functional local public goods present in the community at baseline. For SCA #2 regarding the role of women and youth in decision making, Appendix Table H shows that the number of women (youths) attending the deliberation between salt and batteries is positive and significantly predicted by the baseline number of female (youth) respondents who reported that they had attended a community meeting in the past year. Similarly, the number of women (youths) who made a public statement is positively related to the baseline number of female (youth) respondents who claimed to have spoken up during a recent community meeting, although this correlation is not significant at traditional levels.

SCA #3 concerning elite capture was less successful in generating variation in performance across communities (as discussed in detail below), complicating the validation exercise. Specifically, we find that nearly all communities used the tarpaulin in a public way, as opposed to being privately "captured." Along similar lines, 57% of respondents reported that they had directly benefited from the tarp, and 90% reported that they received some of the salt or batteries. One possible explanation is that the highly public nature of the tarp and gift distribution – which occurred in the same open community meeting discussed above – may have curtailed the ability of local leaders to capture the asset. More speculatively, it remains possible that we may have seen more "capture" during the surprise follow-up visit if the field teams had instead surreptitiously handed the tarp to the local headman, which may more closely mimic the

way that transfers are sometimes made to rural communities; we leave this for future research.[12]

3.3 The Pre-Analysis Plan

The econometric analysis follows a pre-analysis plan (PAP) that was laid out in three steps: (i) an outline hypothesis document agreed to with the GoBifo project implementation team on October 10, 2005; (ii) a detailed pre-analysis plan listing all research hypotheses, the outcomes grouped under each hypothesis, and econometric specifications (including use of mean effects) that was archived with the Abdul Latif Jameel Poverty Action Lab Randomized Evaluation Archive on August 21, 2009 while data entry, cleaning and reconciliation was being carried out and prior to any analysis; and (iii) a supplement to the plan covering outcomes collected in the surprise 2009 follow-up visit (which was fielded five months after the first endline survey) that was archived on March 4, 2010. (The plan and supplement with time stamps are available online at http://www.povertyactionlab.org/Hypothesis-Registry and in supplementary Appendix A.)

The use of PAP's to "tie the hands" of researchers and limit the risks of data mining and specification search is common in medical trials. It is much less common, though not unknown, in economics. The finding of "author effects" amongst estimates of the impact of the minimum wage led to concerns about specification search and publication bias (Card and Krueger 1995). In response, in the first use of a PAP in economics (to our knowledge) Neumark (1999, 2001) pre-specified how data would be used to analyze the impact of changes in U.S. minimum wage laws in 1996 and 1997 before these data became available.

The interest in PAP's has grown with the spread of randomized evaluation methods in economics.[13] While the experimental framework naturally imposes some narrowing of econometric specifications, there is still considerable flexibility for researchers to define the outcome measures of interest, group outcome variables into different hypothesis "families" or domains, identify population subgroups to test for heterogeneous effects, and include or exclude covariates. PAP's are arguably particularly valuable, therefore, when there are a large number of plausible outcome measures of interest and when researchers plan to undertake subgroup

[12] For those interested, the detailed SCA supervisor field instructions are included in Supplementary Appendix I.
[13] At the time of writing, multiple efforts to establish registries for randomized control trials in economics are under discussion, including within the American Economic Association, the American Political Science Association, and the International Initiative for Impact Evaluation. There have also recently been calls for pre-analysis plans within psychology, see Simmons, Nelson and Simonsohn (2011). Some other recent papers in economics and political science that use or discuss pre-analysis plans include Alatas et al. (2012), Finkelstein et al (2012), Humphreys et al (2012), Olken et al (2010), Rasmussen et al. (2011), and Schaner (2011) .

analysis. The process of writing a PAP may have the side benefit of forcing the researchers to more carefully think through their hypotheses beforehand, which could in some cases improve the quality of the research design and data collection approach.

As with any attempt to "tie one's hands", PAP's are not without their risks. In particular, a leading concern is that important hypotheses will be omitted from the initial plan, perhaps due to simple oversight or to research progress in the discipline during the period between the writing of the PAP and data analysis, which could create a desire to carry out additional tests. Another risk is that the exact econometric specification laid out in advance does not describe the data as well as one that would have been chosen *ex post* if the authors had first "let the data speak", potentially leading to less precise estimates. Some of these risks can be mitigated, for example, by finalizing the set of outcomes after project implementation is completed (rather than before the start of the project), or by specifying a detailed algorithm through which the econometric specification will later be determined (for instance, based on patterns observed in the control group) rather than predetermining the exact specification up front. Such approaches provide the researcher with some degree of discretion, and underscore the fundamental trade-off in the practical implementation of PAP's between flexibility and commitment.

It is tempting to advocate a "purist" position that rules out any researcher discretion in order to provide the strongest possible safeguards against data mining, specification search, and other forms of tendentious reporting. This would entail specifying the complete plan in advance of program implementation and allowing no alterations. Any flexibility introduces the risk of manipulation: for example, if a researcher observed that a treatment village had experienced a large exogenous shock (orthogonal to the program being studied, e.g., a new factory opened there with many high paying jobs), she could add outcomes to the analysis plan (e.g., wage earnings) and falsely claim that any gains were due to the intervention itself. The countervailing concern is that rigidly conforming to the PAP will stifle learning. Moreover, if the rules governing the use of PAPs are too "tight", many researchers may resist their adoption and the benefits they offer will not be realized. Based on our experience, we advocate a compromise position that allows some researcher flexibility accompanied by the "price tag" of full transparency—including a paper trail of exactly what in the analysis was pre-specified and when, and public release of data so that other scholars can replicate the analysis—with the hope that this approach will foster the greatest research progress.

We found that there are a great many decisions involved in writing a PAP, and the economics profession has not yet agreed upon a set of best practices. In an effort to further this discussion, below we describe some of the tradeoffs we faced, the choices we made, and things we could have done better in writing and using a PAP. We focus on four issues: (i) the timing of writing and registering the plan; (ii) defining the research hypotheses and outcome measures; (iii) the level of econometric and analytical detail to include in the plan; and (iv) statistical adjustments for multiple testing.

Timing of Pre-Analysis Plans

First is the question of timing, and in particular, how early in the project and research implementation process the PAP should be written and archived. Many medical trials specify the entire analysis plan—including all outcomes and control variables—before the intervention or data collection have begun. Given that economic development programs are not typically administered with the same standardized protocols as drug regimens, we instead found it useful to follow a hybrid approach that pinned down the general domains of likely impacts before project implementation began, but fleshed out the exact outcome measures later on, before any analysis of the endline data. (We archived our PAP while data entry for the 2009 follow-up survey was taking place; to make the design airtight, future scholars should ideally archive their PAP before follow-up data collection has even begun.) This approach has several concrete advantages. Agreeing with the GoBifo project team to a precise, limited list of objectives in the October 2005 hypothesis document (which was finalized before baseline data collection had been launched or villages had been randomized into treatment groups) mitigated the risk that, upon finding no evidence of impacts on a set of outcomes *X*, the project's managers or funders would claim that, in fact, they had been seeking to impact a different set of outcomes *Y*. One concrete concern for us arose from the GoBifo project team's interest in effects on "social capital", which we felt was not a precisely defined concept, and thus risked later charges that we had simply not selected the right aspects of social capital to study.

At the same time, the flexibility to expand the set of outcomes over the course of project implementation allowed us to incorporate lessons learned during the baseline survey and the piloting of the SCAs, as well as respond to shifting emphases and implementation issues in the GoBifo project. As an example, when asking in the baseline survey how respondents would

have the community spend an amount of money comparable to the GoBifo grants, we omitted the response category of "business skills training", which were not part of our understanding of CDD practice at the time. Such training ended up accounting for a nontrivial share of project grants (as mentioned above), and thus the endline survey added a question to correct this omission. We believe it is appropriate to include in pre-analysis plans hypotheses that are generated while the project is ongoing, including those arising from field observation. Importantly, though, we refrained from dropping any of the original 2005 research hypotheses or outcome domains, with the view that documenting the absence of effects on areas that were originally viewed as within the remit of the project would be useful for the research community.

Another timing consideration is when to make the PAP public. Researchers may rightfully worry that PAPs include extensive research design details, and that making them public immediately would allow other scholars to copy novel insights and effectively "scoop" the authors of the PAP by beating them to publication. One option to address this legitimate concern is to simply include the PAP as a supplementary document when the paper is submitted to a journal so that referees can compare the paper to the original analysis plan, and then publicly release the PAP only upon final journal publication. In our view, however, the increased transparency that comes from prior publication of the PAP enhances the credibility of the process and we asked J-PAL to post this study's PAPs on their website before publication of the article.

There is a closely related set of issues around the benefits of creating public registries of planned trials and PAPs in the social sciences (similar to the website clinicaltrials.gov) in order to help limit publication bias. A main benefit of a public registry is that it would make it more transparent to other scholars which studies had been started on particular topics but for which papers were never published. To the extent that these projects had registered PAPs, there would also be a rich source of information on the details of how these studies were to be carried out, leading to a broader understanding of the field and enriched meta-analyses. In our view, it would thus be useful to set a "time limit" – potentially of up to several years – after which a registered PAP would be publicly released even if the results were not yet published.

Defining the Hypotheses and Outcomes

This ability to accrue new hypotheses over time was the main aspect of our own research where flexibility (accompanied by transparency) was most useful. Specifically, the 2009 PAP includes

one hypothesis not included in the original 2005 document—regarding impacts on social and political attitudes—that arose after observing and reflecting upon program activities. We further split one 2005 hypothesis into two hypotheses in the 2009 document, namely, separating impacts on public goods from other measures of collective action. While writing this paper, we further added a twelfth hypothesis (called hypothesis 1 below) by pulling together project implementation outcomes that had already been listed within the other hypotheses but were not explicitly specified as a distinct sub-grouping in the PAP. It is important to note that in making these adjustments no new outcome measures were added or excluded from the final PAP list in what we present below. Those who wish to consider only the results as exactly laid out *ex ante* can ignore hypothesis 1. However, we feel that the absence of a project implementation hypothesis was an oversight on our part and find the results of hypothesis 1 useful to consider. For transparency, our main table of results (Table 2) presents family-wide error rate adjusted p-values for both the original grouping of 11 hypotheses and for the *ex post* expansion to 12 hypotheses. Moreover, we also accommodate a "purist" approach by calculating the mean indices for only those hypotheses laid out in the 2005 document and including only those measures collected in the baseline data. As we discuss below, the main results of this paper are unchanged across these various approaches.

Perhaps more important is the fact that we grouped the various research hypotheses from the PAP into two distinct "families" while writing the paper, for ease of interpretation and to facilitate links to theory. While we did not specify these two families beforehand, we believe that the groupings—the development "hardware" of project implementation, public goods and economic activity (family A), and the "software" of local collective action (family B)—are compelling. Again, the reader is free to ignore the two family-level indices and focus exclusively on the treatment effects estimated for the hypothesis-level indices, as well as the particular outcome measures. We disclose complete results for all 334 unique outcome variables, including the exact survey question wording, in supplementary Appendix J.

An alternative and more sophisticated approach to accommodate flexibility that we did not use but consider promising is pre-specifying an algorithm that researchers will use to make subsequent judgments on the analysis that depends on information not available at the time the pre-analysis plan is written. Such approaches have already been extensively employed in statistics (see van der Laan and Rose 2011). As an illustration of the value of such an approach,

we decided not to include several outcome variables in the final 2009 PAP that had minimal variance in the baseline data. For example, in attempting to gauge respondents' knowledge of local government activities, we dropped measures that turned out to be far outside the realm of our respondents' experience: we dropped measures when we found that fewer than 1% of respondents at baseline knew exactly how much was collected in local taxes in their chiefdom section, or the official proportion of that tax that goes to chiefdom coffers versus elected local government officials. While we did not do so, it might have been preferable to include an explicit decision rule in the 2005 hypothesis outline document to define the variance threshold (in the baseline data) that we would use in deciding whether to exclude a variable from the final PAP.

We could similarly have specified a rule in the 2009 PAP to guide the "dropping" of new outcome measures (not collected in the baseline survey, such as our SCAs) that fell short of a particular variance threshold or had a high non-response rate in the endline data. In a related approach, for example, Finkelstein et al. (2012) examined the distribution of outcomes in their control group endline data to identify and exclude binary measures with minimal variance, i.e., with a mean very near zero or one. They further used the control group data to determine the most relevant margins along which to collapse categorical variables into binary measures, and allowed the observed degree of skewness in continuous variables to affect functional form choice. By contrast, in the analysis presented in this paper, we did not drop any outcome measures that had been specified in our 2009 PAP, and are thus left with several where the proportion of positive responses in the control group is sufficiently high to make the estimation of treatment effects largely uninformative—e.g., whether the community held a meeting to discuss use of the tarp in SCA #3 (in Table 5 below), which took place in 98% of communities.

Choosing the Optimal Level of Detail

Third is the issue of the level of detail to include in a pre-analysis plan. These *ex ante* commitments are critical for preventing *post hoc* specification searching and to generate appropriately sized statistical tests. To eliminate data mining, the final PAP defines both the sets of explanatory and dependent variables, as well as the precise specifications to test, including the set of interaction terms and population subgroups used to explore heterogeneous treatment effects (Leamer 1983, 1974). Our PAP specified that we would run the analysis both with and without covariates, on endline data only as well as incorporating baseline data where available,

and at each of the natural levels of aggregation in the data (individual, household and village). One shortcoming of the approach we took in the 2009 PAP is that we did not explicitly state which of these specifications was our primary test and which others would serve as "robustness checks". If writing our PAP again, we would instead select a single econometric specification to be our main approach – in this case, the most natural one would be the conservative approach of including minimal regression controls, using endline data only and carrying out village-level analysis, which we focus on in Table 2. Fortunately for us, the results in this paper are unchanged across different sets of controls, panel data, and levels of aggregation (as shown in Table 3), but in the future other scholars might prefer to eliminate such ambiguity from the PAP.

Accounting for Multiple Inference

Given the large number of outcome variables we consider, the other key risk is over-rejection of the null hypothesis due to the problem of multiple inference (Anderson 2008). Our plan thus first commits us to a mean effects approach that reduces the effective number of tests we conduct by identifying in advance which outcome variables to group together in testing a hypothesis (see O'Brien 1984; Kling, Liebman and Katz 2007). Note that the credibility of the mean effects approach depends critically on specifying in the PAP exactly which outcomes will be grouped under which hypotheses, lest the ability to reshuffle outcome measures across hypotheses opens up another avenue for tendentious reporting. Yet even with a mean effects approach, we are still testing multiple hypotheses, and so use the Westfall and Young (1993) free step-down resampling method for the family wise error rate (FWER), the probability that at least one of the true null hypotheses will be falsely rejected (as detailed in Anderson 2008). The PAP again lends credibility to this process by confirming that there were no hypotheses that were tested but excluded from the multiple testing adjustment. Grouping the hypotheses into two families, as we do, also helps to combat this issue by further reducing the number of statistical tests from twelve to two, but since we did not specify the two families in the PAP we place less emphasis on them.

Mean effects estimation and the accompanying FWER adjusted p-value is the primary metric by which we evaluate a hypothesis. We also provide results for the outcome measures individually to provide a sense of their magnitude and economic significance. Appendix J lists three distinct p-values for each particular outcome measure: i) the "naïve" or "per comparison" p-value, which is appropriate for a researcher with an *a priori* interest in a specific outcome (see

discussion in Kling, Liebman and Katz 2007); ii) the FWER adjusted p-values mentioned above that limit the probability of making a Type I error for *any* specific outcome within the hypothesis; and iii) the slightly less conservative false discovery rate (FDR) adjusted q-values that limit the expected *proportion* of rejections within a hypothesis that are Type I errors (Benjamini, Krieger and Yekutieli 2006, Anderson 2008).

3.4 Econometric Specifications

Under each hypothesis, we evaluate specific treatment effects using the following model:

$$Y_c = \beta_0 + \beta_1 T_c + X'_c \Gamma + W'_c \Pi + \varepsilon_c \quad (1)$$

where Y_c is an outcome (i.e., local school construction) in community c; T_c is the GoBifo treatment indicator; X_c is a vector of the community level covariates (controls); W_c is a fixed effect for geographic ward, the administrative level on which the randomization was stratified; and ε_c is the usual idiosyncratic error term. Elements of X_c always include the two village-level balancing variables from the randomization process—distance from a road and total number of households—and our results are robust to the inclusion of additional control variables specified in the PAP, including an index of civil war violence, ethnolinguistic fractionalization, and the historical extent of domestic slavery. The parameter of interest is β_1, the average treatment effect. As mentioned earlier, while some outcomes are measured at the household (e.g., radio ownership) or individual level (e.g., political attitudes), the natural unit of analysis is the village since some measures are only collected at that level (e.g., the existence of a village grain drying floor) and we thus measure all variables at this level, taking village averages as necessary. For the subset of outcome variables that were collected in both the 2005 baseline and 2009 follow-up surveys, our results are robust to leveraging the panel structure of the data.

As set out in the pre-analysis plan, we assess the degree of heterogeneous treatment effects by respondent gender, age, village remoteness, community size, war exposure, domestic slavery, and location in each of the two study districts. As we do not find evidence for heterogeneous effects along these dimensions, for reasons of space we have excluded this discussion from the text (see supplementary Appendix K for details).

The mean effects index for a hypothesis captures the average relationship between the GoBifo treatment and the K different outcome measures grouped in that hypothesis. Following

Kling, Liebman and Katz (2007), estimation of the mean treatment effect: (i) orients each outcome so that higher values always indicate "better" outcomes; (ii) standardizes outcomes into comparable units by translating each one into standard deviation units (i.e. by subtracting the mean and dividing by the standard deviation in the control group); (iii) imputes missing values at the treatment assignment group mean; (iv) compiles a summary index that gives equal weight to each individual outcome component; and (v) regresses the index on the treatment indicator as well as any control variables. This is the mean effects approach specified in our PAP, and is what we present as our main results for all full sample outcomes in Table 2.

As discussed below, the results are robust to using alternative index estimation techniques that were not specified in the PAP (Table 3). The seemingly unrelated regression (SUR) approach in Kling and Liebman (2004) accommodates item nonresponse (which is important when we extend the set of outcomes to include what we call "conditional outcomes", those that depend on the existence of a particular public good in the community and therefore are only measured for a subset of observations) and allows a flexible combination of panel and endline only analyses. The results are also robust to the approach described in Anderson (2008) that weights each component by the inverse of the appropriate element of the variance-covariance matrix (as measured in the control group) to maximize the information captured in the index. This approach "down weights" outcome measures that are highly correlated with each other, addressing the concern that, in effect, we may at times be repeatedly measuring the "same" outcome in a variety of slightly different ways under a given hypothesis.

4. Empirical Results

Column 1 of Table 2 presents a concise summary of the mean effect results for all twelve hypotheses, grouped into the two outcome families. Column 2 provides the corresponding "naïve" p-value that does not account for multiple inference; the remaining columns adjust this p-value to control the family-wise error rate (FWER) when considering the hypotheses as a group, where the group is defined as the full set of 12 hypotheses (Column 3), and the 11 hypotheses in the 2009 PAP (Column 4).

The three hypotheses under family A are that: "GoBifo creates functional development committees" (H1); "Participation in GoBifo improves the quality of local public services infrastructure" (H2); and "Participation in GoBifo improves general economic welfare" (H3).

The positive and significant (at 99% confidence) mean effect estimate of 0.298 standard deviation units for family A (hypotheses 1, 2 and 3) indicates that GoBifo achieved its most immediate objective of providing the organizational and financial means to encourage local public goods construction and small enterprise development. Specifically, the coefficient on hypothesis 1 indicates that the program was well executed, perhaps more so than many other real-world projects: GoBifo increased measures of local organization and linkages to facilitate collective action by 0.703 standard deviations on average. This strong implementation performance in turn led to immediate impacts on local infrastructure. The estimated mean effect of 0.204 for hypothesis 2 reflects positive effects on the stock and quality of local public goods; while the 0.376 coefficient for hypothesis 3 reflects gains in general economic welfare. The mean effects estimates for the first three hypotheses are significant at 99% confidence across all p-value adjustments. Reflecting back on the theoretical framework, these increases provide strong support for the prediction that the combination of lowering the marginal cost of public goods through grants, as well as reducing coordination costs through the establishment of new institutions, led to greater public investment. The next question is how much of this effect was driven by changes in institutions, norms and collective action.

The nine hypotheses in family B include: "Participation in GoBifo increases collective action and contributions to local public goods" (H4); "GoBifo increases inclusion and participation in community planning and implementation, especially for poor and vulnerable groups; GoBifo norms spill over into other types of community decisions, making them more inclusive, transparent and accountable" (H5); "GoBifo changes local systems of authority, including the roles and public perception of traditional leaders (chiefs) versus elected local government" (H6);[14] "Participation in GoBifo increases trust" (H7); "Participation in GoBifo builds and strengthens community groups and networks" (H8); "Participation in GoBifo increases access to information about local governance" (H9); "GoBifo increases public participation in local governance" (H10); "By increasing trust, GoBifo reduces crime and conflict in the community" (H11); and "GoBifo changes political and social attitudes, making individuals more liberal towards women, more accepting of other ethnic groups and 'strangers',

[14] The PAP states: "this is not an explicit objective of the GoBifo project leadership itself, but it is a plausible research hypothesis."

and less tolerant of corruption and violence" (H12)[15].

The small and not statistically significant mean effect estimate for family B (hypotheses 4 through 12), at 0.028 standard deviation units (standard error 0.020) provides no evidence that the experience of working together in GoBifo, and the introduction of new institutions and processes, durably changed the nature of local collective action. The program's democratic decision-making and "help yourself" approach did not appear to spill over into other realms of village life nor to persist into the post-program period. We find no evidence that GoBifo led to fundamental changes in local capacity to raise funds and act collectively outside of the project, the nature of decision-making, the influence of women or youths, or a range of social capital outcomes. In the context of the model, these null results suggest that GoBifo did not permanently increase the benefits of civic engagement for marginalized groups and that the organizing institutions established did not persistently reduce the fixed costs of collective action. Although this estimate is close to being significant at traditional confidence levels, with a p-value of 0.155, it is very close to zero and an order of magnitude smaller than the family A effect.

The results are robust to alternative specifications, including the Anderson (2008) re-weighting mean effects approach (Table 3, column 1), the Kling and Liebman (2004) SUR mean effects approach (column 2), including panel specifications where the data is available (column 3), including additional control variables (column 4), dropping endline survey replacement households to partially address attrition (column 5), including "conditional outcomes" that apply only to a subset of observations (column 6), and restricting the set of hypotheses and outcomes to those specified in the pre-program 2005 hypothesis document (column 7).

4.1 Family A: Development Infrastructure or "Hardware" Effects

The first hypothesis focuses on project implementation and measures the extent to which GoBifo successfully established Village Development Committees (VDCs); helped communities draw up development plans and open bank accounts; and created links between the villages and their local government representatives. The first panel of Table 4 present results for several outcomes under this hypothesis that demonstrated statistically significant treatment effects, where the first four "full sample" outcomes apply to all communities within the sample; while the remaining

[15] Regarding hypothesis 12, the pre-analysis plan notes: "this was not part of the original [2005] program hypotheses document but relates closely to GoBifo project objectives."

three "conditional" outcomes in Panel A depend on the existence of public infrastructure and thus only apply to those communities that have the particular good. Table 4 reports unadjusted p-values that are appropriate for those with an *a priori* interest in the individual outcome; the corresponding FWER and FDR adjusted values are presented in supplementary Appendix J.

Regarding interpretation, the treatment effect estimate in the first row of Table 4 indicates there was an increase of 40 percentage points in the existence of a VDC. VDCs already existed in many Sierra Leonean villages when GoBifo was launched, having been introduced by humanitarian assistance groups during the war-torn 1990's (Richards et al. 2004). By the post-program period, this treatment effect led the proportion of treatment communities with a VDC to be nearly double that of the controls. The corresponding coefficient in the second row indicates that GoBifo increased the likelihood that a community was visited by a member of its Ward Development Committee in the past year by 13 percentage points. Row 3 shows a positive treatment effect on the existence of village development plans by 30 percentage points, nearly a 50% increase on the base of 62% in the controls. Row 4 reveals an increase in having a village bank account of 71 percentage points, a nearly tenfold increase. The household survey also asked whether a member of the Ward Development Committee or district council was "directly involved in the planning, construction, maintenance or oversight" of local public goods. The positive and significant treatment effects on primary schools, grain drying floors, and latrines suggest that GoBifo successfully led local politicians to increase their involvement in village projects, consistent with its objective of supporting the broader decentralization process.

Hypothesis 2 explores treatment effects on the quantity and quality of local public goods. When considering individual outcomes, the measures under hypothesis 2 naturally form three sub-groups: those regarding the stock of local public goods, the quality of such goods, and community financial contributions to their construction and upkeep. Regarding the stock, the first three rows of Panel B in Table 4 present impacts for an illustrative sample of goods, where we find marked increases in the proportion of villages with a functional traditional midwife post by 17 percentage points, latrine by 21 points, and community center by 9 points.

The last three rows of Panel B show positive GoBifo impacts on the construction quality of three of the most common public goods—primary schools, grain drying floors and latrines—as determined through direct physical assessment by enumerators. These measures combine impacts from the GoBifo funded infrastructure projects, as well as any effects from better

maintenance of existing infrastructure. However, as there is no evidence that management practices did in fact change in treatment villages (as detailed below), the leading interpretation is that the positive impacts are being driven by the grants.

The final sub-group of outcomes concerns community financial contributions to existing infrastructure and these are omitted from Table 4 due to a lack of statistically significant effects. Combined with the negative and significant effect on whether the community approached another NGO or donor for financial support (in Panel B), these provide suggestive evidence that GoBifo funds may have served as a substitute for the community's own resources. At a minimum, they provide no evidence that GoBifo grants served as a catalyst for additional fund-raising nor that project experiences encouraged participants to seek out further development assistance. The SCA findings discussed below reinforce this view.

Hypothesis 3 relates to general economic activity and household welfare, since roughly one sixth of the grants were used to launch projects dedicated to job skills training or small business development—such as carpentry and soap-making—that, if well implemented, could translate into higher earnings. Along similar lines, another 40% of the grants went toward investments in agriculture and livestock, another common type of small business investment. Moreover, GoBifo injected cash grants into very poor communities, and as with any assistance, a portion of the funds are surely fungible.

The first two outcomes in Panel C of Table 4 refer to village-level outcomes, where we see a 30% increase in the number of petty traders (0.7 more traders on a base of 2.4 in the control group) and a 13% increase in goods locally available for sale. We also observe improvements in an asset ownership score (created using principal components analysis), where the underlying assets include common household durables (e.g., radios, mobile phones), amenities like drinking water source and sanitation, and the materials used in the dwelling's roof, walls and floor. The project tripled the proportion of respondents who had recently participated in skills training: a 12 percentage point increase on a base of 6% in control communities. We find no evidence that the program impacted total household income (not shown), however, income is quite difficult to measure among subsistence farmers and the treatment effect estimate is relatively imprecise.

4.2 Family B: Impacts on "Software": Local Institutions and Norms for Collective Action

The positive treatment effects for outcome family A suggest that investment in local public

goods did increase substantially during the project. To determine the role played by more effective local institutions (rather than the block grants alone), we next examine post-program outcomes after the block grants had been spent. The first hypothesis under family B (hypothesis 4) covers outcomes relating to collective action and contributions to local public goods. The mean effect for this hypothesis is not statistically distinguishable from zero under any p-value adjustment (0.012 standard deviations with a standard error of 0.037, Table 2); and of the 62 full sample and conditional outcomes evaluated, only seven treatment effects are significant at 95% confidence, with three positive in sign and four negative. The subset of outcomes relating to the matching grant opportunity (SCA #1) provides the most succinct and concrete illustration, as the ability to mobilize around a new opportunity, and raise funds for it, captures the essence of local collective action. In the top panel of Table 5 we cannot reject zero differential take-up of the subsidized building vouchers: 62 treatment (52%) and 64 control villages (54%) redeemed vouchers at local supply stores. Nor can we reject equality in the number of vouchers redeemed across treatment and control areas.

Other outcomes under this hypothesis consider household contributions to existing local public goods, where we expand the set of contributions to include labor, local materials, or food for project workers, yet continue to find no evidence of treatment effects. We also find no evidence for differences in contributions to several local self-help groups (i.e., rotating savings groups, labor gangs) nor in financial support for community teachers. Lastly, while treatment villages were much more likely to have a communal farm, by 23 percentage points (significant at 99% confidence), we cannot reject that the total number of respondents in treatment areas who had worked on a communal farm in the past year was the same as in controls. This presents a telling example of how our results document a proximate effect of project activities on a local organization established to capture that funding—i.e. the subsidized provision of seeds and tools led to the creation of a community farm—yet find no evidence that these translated into lasting impacts on participation in that organization or changes in behavior.

These findings raise questions about GoBifo's long term impacts. Clearly, community members gained experience in working together to implement projects over the nearly four years of the project. Yet we find no evidence that their GoBifo-specific experiences lead to greater capacity to take advantage of new opportunities that arose after the program ended. Most strikingly, while GoBifo often created new structures designed to facilitate local development by

reducing organizational costs—the VDC, a development plan, a bank account, and a communal farm—we do not find evidence that these structures left them better able to take advantage of the realistic matching grant opportunity in SCA #1.

Hypothesis 5 in family B includes outcomes relating to the civic involvement of socially marginalized groups. Since the inclusion of women and youth held great prominence in GoBifo's objectives and facilitator operating manuals, it also received special attention in the data collection. Covering an exhaustive battery of measures, the mean effect cannot be distinguished from zero and has a narrow confidence interval (see Table 2), providing no evidence of impacts on the role of women or youth in local decision-making, or on the transparency and accountability of decision-making more generally. Of 82 distinct outcomes, only seven were significant at 95% confidence, with four positive and three negative treatment effects.

Enumerator observations during SCA #2, when villages met to decide between salt and batteries, provide a clear illustration. In Panel B of Table 5 there is no evidence for treatment effects on the total number of adults, women and youths who attended the meeting or spoke publicly during the deliberation. On average, 25 women attended these meetings but just two of them made a public statement during the discussion about which item to choose. The estimated difference between the number of women who spoke in treatment versus control communities is only -0.20 (s.e. 0.22), and the proportion of males who spoke during the meeting remained twice as high as the proportion of females in the treatment villages, the same as at baseline. We similarly find no evidence of impacts on whether any smaller "elite" groups broke off from the general meeting to make the gift choice without broader consultation; the duration of the deliberation; or how democratic the decision-process appeared to the enumerators, e.g., by holding a vote. These patterns are consistent with the data from respondent reports recorded immediately after the meeting of how the tarpaulin allocation choice in SCA #3 was made, including which individuals had the final "say" and to what extent the decision was dominated by local elites (i.e., village headmen and male elders). Moreover, respondent opinions collected during the second 2009 follow-up survey reveal no evidence of treatment effects on reports about how decisions were made to distribute the salt or batteries (SCA #2); how to use the tarp (SCA #3); whether to raise funds for the building materials vouchers, and if so, how to mobilize funds, which items to purchase, and how to manage any construction (SCA #1).

Despite all of the effort in GoBifo to elevate the position of women and youth, we thus do

not observe any improvement in their role relative to older men in community decision making. Even for relatively low cost actions like speaking up in meetings, we find no evidence that the project translated into greater "voice" for marginalized groups. In the context of the theory, this suggests a lack of persistent gains in the individual benefits of participation for these groups, and provides additional evidence that the increase in public investment observed during project implementation was likely driven by the financial subsidy rather than fundamental changes in local institutions or *de facto* power.[16]

Hypothesis 6 (which we included in the pre-analysis plan out of research interest but was not an official aim of GoBifo project management) asks whether by espousing more democratic ways of managing local development, the project reduced the role of the traditional chiefly authorities. Taking all outcomes together, we cannot reject a coefficient of zero on the mean effect for hypothesis 6 (Table 2). Many outcomes under this hypothesis estimate the extent to which the village head and elders dominated the SCA decisions. While we find variation in how these decisions are made—at one extreme, in two villages the Chief decided between the salt and batteries in less than one minute without anyone else's input, while at the other an open discussion lasted nearly an hour and was followed by a formal vote—as mentioned above, we find no systematic differences across treatment and control villages.

A leading explanation for the apparent lack of institutional change, with some support in the data, is that elites exerted substantial control over the new organizations GoBifo created. As an example, we find that traditional elites retained their leadership of the VDC: in both treatment and control villages (for the roughly half of control communities with a VDC in 2009), approximately 88% of VDC chairs are men, 87% are older than 35, and 52% are traditional chiefdom authorities and elders. While participation requirements translated into some gains for women (a 6.6 percentage point increase in the proportion female members and a near doubling of the proportion of female Treasurers, 57% versus 31%), we cannot reject that the representation of youths remained at the same low level as in control areas (at 26%). These patterns highlight a tension inherent in the CDD approach: leveraging the capacity of existing institutions may be expedient for immediate project implementation while simultaneously limiting the likelihood of fundamental institutional transformation or changes in *de facto* power for marginalized groups.

[16] However, we cannot rule out that the subsidy was particularly effective (i.e., led to such notable increases in public goods) in part because of the project's facilitation and emphasis on participation and transparency.

We therefore tested the related hypothesis that CDD may enable local elites to capture a disproportionate share of economic benefits by distributing the tarp (SCA #3) during the first 2009 follow-up visit and observing how it was being used in the unannounced visit five months later. While the analysis finds no evidence of treatment effects on elite capture, it also reveals that the level of elite capture is, perhaps surprisingly, relatively low in the study communities. Panel C of Table 5 shows that for the 90% of communities that had used the tarp by the time of the second visit, 86% had put the tarp towards a public purpose, such as a communal rice drying floor or local ceremony. The most obvious example of elite capture would be use of the tarp to patch the roof of a single individual's house, which happened in fewer than 3% of all villages. That said, only 6% of villages were storing the tarp in a public place when not in use (with the vast majority storing it in the chief's residence) and several communities had not yet used the tarp, suggesting a failure to agree upon a use, or the risk of future elite capture, or both.

The next three hypotheses explore proxies for "social capital"—self-expressed trust of others (hypothesis 7), involvement in local groups and networks (hypothesis 8), and access to information (hypothesis 9)—emphasized alongside collective action and inclusion in the official GoBifo project objectives (World Bank 2004, GoBifo 2007). The analysis finds no evidence of treatment effects on social capital, with all three mean effects indices indistinguishable from zero. Beginning with trust, the only significant effect is an increase in reported trust of NGOs and donor projects: residents in treatment communities were 5.4 percentage points more likely to agree that NGOs or donors "can be believed" (the closest Krio translation for trust) as opposed to you "have to be careful" in dealing with them. There is no evidence for effects on the remaining eleven indicators, which include respondent self-reports on their trust for various groups and hypothetical vignettes, such as entrusting money to a neighbor to purchase goods on your behalf.

Enumerators asked respondents whether they were a member of a local self-help group (i.e., credit/savings group, school committee, women's group, youth group, among others) and if so, whether they had attended a meeting and contributed financially or in labor in the past month (hypothesis 8). We find no significant treatment effects on these indicators nor on other measures of local cooperation, such as whether the respondent had helped a neighbor re-thatch the roof of their house, a time-intensive activity that one cannot easily do alone.

There is also no evidence of treatment effects on households' access to information about local government or governance (hypothesis 9). Among 21 outcomes, only one—the proportion

of villages visited by a WDC member, discussed above—shows statistically significant effects. As examples, we fail to reject a zero treatment effect for measures of how much respondents know about what the community is doing with the building vouchers (SCA #1) and tarp (SCA #3); whether they can name their district council and chiefdom leaders; and their ability to answer objective questions about how local taxes are collected and used.

While the mean effect index for participation in local governance in Table 2 (hypothesis 10) is positive and statistically significant if one considers the unadjusted p-value, this result does not survive the FWER adjustments (columns 3 and 4). It is also largely driven by the outcomes already discussed under family A (certain outcome measures are included under multiple hypotheses). Specifically, we find large impacts on the existence of VDCs and village plans, and increases in the oversight of local public goods by chiefdom authorities that mirror earlier results on the involvement of local government representatives. There is no systematic evidence, however, of more active individual political engagement, such as self-reported voting or running for local office.

There is no evidence that the program affected the level of crime and conflict or the mechanisms through which they are resolved, leading to a zero mean effect for hypothesis 11 (Table 2). Of the ten indicators considered, only one—the 2 percentage point reduction in household reports of physical fighting over the past year—is significant at 95% confidence. While the nine null results suggest that project efforts to enhance conflict management capacity may not have created lingering benefits, on the positive side it provides some reassurance that the infusion of external grant money at least did not appear to spark increased conflict.

The twelfth and final hypothesis concerns the nature of individual political and social attitudes. The GoBifo program's emphasis on the empowerment of women and youth, and the transparency of local institutions, may have engendered a more equitable or "progressive" outlook toward politics and society more generally. Even if there are no changes in actual decision-making processes or local collective outcomes (as above), a marked change in expressed attitudes might still mean that the "seeds" for future social change had been planted. Enumerators gauged attitudes using pairs of opposing statements, such as "As citizens, we should be more active in questioning the actions of leaders" versus "In our country these days, we should have more respect for authority," and asking respondents which they agreed with more. These paired statements capture respondent views on a diverse range of topics including

the acceptability of violence in politics (a particularly salient issue in post-war Sierra Leone), domestic violence, youth and women in leadership roles, paying bribes, and coerced labor. Once again, there is no evidence of significant program effects, despite the concern that social desirability bias might lead some respondents to express views promoted by the program. The only significant impact is a positive 4 percentage point increase in agreement with the statement that young people can be good leaders. However, recall the lack of evidence that this change in opinions translated into more youths holding actual leadership positions on the VDC, or to more youth participation in the SCA meetings. Attitudinal change may be a necessary step toward changing future behavior, but almost four years of an intensive community driven development program did not lead to detectable changes in a range of expressed attitudes.

4.3 Robustness and Validity Checks

This section evaluates the robustness of the results. To start, we consider typical threats to randomized experiments. Fortunately, there were no problems with treatment non-compliance: all communities assigned to the treatment group received the program and none of those in the control group participated; and respondent attrition rates are no different in treatment and control areas. The baseline statistics presented in Table 1 and supplementary Appendix F also suggest that the randomization process successfully created two groups of villages that were similar along a wide range of observables. Note further that the results are unchanged in panel analyses that utilize baseline data when it is available (Table 3). Thus in order for spurious differences between the two groups to explain the positive impacts in family A, the treatment group would on average have had to be on a different trajectory than the controls, but there is no reason to believe this should systematically be the case given the randomized research design.

We next consider reasons why the treatment effect estimates might be underestimates. Given the moderate size of the grants and the fact that villages were geographically spread out, we feel that spillovers from village-level interventions in treatment areas to control communities are unlikely. Of greater concern would be the risk that the projects GoBifo simultaneously implemented at the ward level systematically benefited the control group at the expense of the treatment group. There was a separate pot of funding for each ward that was allocated by the Ward Development Committee (see section 2.2). Bias could result if WDC members took into account the placement of GoBifo village-level projects in deciding where to locate the ward

projects and targeted those areas that had not already benefited, perhaps as a way of compensating them for losing out on village-level assistance. However, there are no meaningful differences in the targeting of ward-level projects across treatment and control villages, and, if anything, treatment villages are slightly more likely to benefit (not shown).

A final concern is that the outcome measures were simply insufficiently refined to detect subtle decision-making, institutional, political or social differences between treatment and control communities. While some of our measures are certainly better than others, our main strength lies in the diversity of measures we use and the fact that they all produce similar results. We combine different data collection approaches, for example, employing both survey self-reports on the percentage of female and male respondents who spoke during the SCA meetings with direct enumerator observation. The research teams also gathered information from a variety of sources: they conducted interviews in respondent homes, held focus group discussions with key opinion leaders, observed a community decision as it unfolded, and recorded their own independent assessment of the construction quality of local infrastructure. Taking all these data together, the "zero" GoBifo program effects in family B are quite precisely estimated. To illustrate, the maximum true positive treatment effect on the proportion of women speaking (in the salt versus battery SCA #2 deliberation) that we may have incorrectly ruled out at 95% confidence is one additional female speaker per every 4.3 villages we visited, which is quite small. In the mean effects analysis, which combines many outcome measures, confidence intervals are tighter still. As an example, the 95% confidence interval for the mean effect across all outcomes in family B is (-0.012, 0.068) measured in standard deviation units, which is a narrow interval containing zero. A Type II error that incorrectly failed to reject the null for either value bracketing this interval would lead us to overlook an effect of negligibly small magnitude.

4.4 Alternative Interpretations and the Perils of Data Mining

Section 4.2 shows that evaluating the institutional change outcomes jointly under their pre-specified hypotheses generates no evidence of program impacts. Yet without the discipline of the pre-analysis plan and mean effects approach, we could have instead selected an assortment of individual treatment effects to tell a range of stories. While such data mining poses a risk for any analysis, it may be particularly problematic for assessing institutional outcomes. The multi-dimensionality of institutions—governing political, economic and social behaviors—implies a

large number of outcomes under family B, some of which will have statistically significant treatment effects by pure chance. Moreover, because institutions are amorphous and contextually determined, there is no commonly agreed set of standard measures defining the core of each domain, allowing the researcher to either deliberately or unintentionally "cherry-pick" a set of treatment effects whose selectively is difficult to detect from the outside.[17] To underscore just how misleading such data mining can be, Table 6 uses our data to construct two alternative interpretations—one negative, one positive—about GoBifo's impacts on institutions.

The selective collection of negative treatment effects in the top panel of Table 6 suggests that the heavy emphasis placed on participation during GoBifo implementation activities created "meeting fatigue" within treatment villages, which eventually translated into poor management of local development projects and political apathy. Specifically, respondents were less likely to report that they had attended a meeting to decide what to do with the tarp after the research teams had left the village. Tracing this initial backlash against participation through the course of the tarp SCA, we see that villagers were less likely to: report that "everyone had equal say" in deciding how to use the tarp; actually put the tarp to use; or be able to produce the tarp for inspection by the survey team. This deterioration in community participation appears to have further manifested in declining civic engagement more broadly, as evidence by decreased interest in holding local office (as a VDC member) and lower turnout in recent local elections.

The second panel of Table 6 presents the opposite story: these treatment effects suggest that the positive experiences communities gained implementing GoBifo projects catalyzed other collective activities and encouraged villagers to incorporate new democratic practices into other realms of decision making. These shifts in collective norms and behaviors in turn created space for new leaders in the community and incited greater interest in politics more generally. More specifically, the outcomes in Panel B reveal gains in non-project collective action, like increased training for community teachers and a greater prevalence of women's groups. Broad adoption of the democratic norms promoted by the project is evidenced by increased minute taking at community meetings, a greater likelihood of storing building materials in a public place, and local chiefs playing a less dominant role in managing the tarp. Finally, the CDD experience

[17] By contrast, any study regarding the returns to education would by necessity focus on individual wages. Of course, even in the measurement of labor outcomes, the analyst retains considerable discretion over outcomes, e.g., total earnings, hours worked, occupation, employment sector, etc., so issues of multiple testing and cherry-picking are likely to also be relevant in domains other than the study of institutions.

instilled a more accepting attitude towards youths taking on leadership roles, and increased citizen awareness of national politics, as seen by the greater ability to correctly name the date of the next general election.

These two plausible, opposite, and equally erroneous interpretations illustrate the risks of allowing researchers complete discretion to choose the subset of outcomes to highlight *ex post*, and the potential value of employing a pre-analysis plan.

5. Conclusion

This paper evaluates a well-implemented program that sought to provide public goods and change institutions in Sierra Leone. Our evidence suggests that the intervention was successful in setting up new village structures, improving local public goods and enhancing economic welfare. We do not, however, find evidence of lasting changes in village institutions, local collective action capacity, social norms and attitudes, or the nature of *de facto* political power.

The results run counter to the currently popular notion in foreign aid circles that community driven development (CDD) is an effective method to sustainably catalyze collective action and fundamentally alter local decision-making. There is no evidence that the establishment of local committees, development plans and bank accounts led to permanent reductions in the fixed organizing costs of collective action, likely because communities did not adopt and apply the new structures to communal endeavors beyond the immediate project. Exposure to democratic project processes similarly did not make traditional elites more willing to seek out the views of others in making community decisions, nor were villages any better able to raise funds in response to a matching grant opportunity. While "good" institutions may be critical for economic performance, our findings provide another piece of evidence that institutions and social norms are difficult to change. Consistent with this perspective, the related U.S. Community Action Program is thought to have had at best limited success in achieving its ambitious goals: Germany (2007: 8) writes that "by the early 1970s, the ebullient visions of the mid-1960's had been discarded."

At the same time, our results challenge the aid pessimist's view that external assistance cannot improve the lives of the poor in countries with "weak" institutions. We find that well-allocated external aid can have a positive impact on welfare. Indeed our results suggest that, in this context, the comparative advantage of the World Bank and other donors may lie more in

providing development "hardware," and less in instigating institutional and social change, at least with current tools such as CDD.

The results further suggest that participation requirements did not foster learning-by-doing or demonstration effects large enough to change attitudes, norms or behaviors towards marginalized groups. Despite requirements on the inclusion of women and youth in project decision making and intensive facilitation designed to enhance their influence, nearly four years later we see that women and youths are no more likely to voice opinions about how the community should manage new public assets. Returning to the comparison between informal interventions focused on reshaping norms, like the program studied here, and changes to the rules of formal institutions, like female leadership quotas, the existing evidence suggests that the latter may be a more effective way to alter *de facto* power dynamics and social perceptions in a modest timeframe (Chattopadhyay and Duflo 2004; Beaman et al. 2009). Importantly, however, we cannot rule out that part of GoBifo's success in using grants to deliver public goods was due to its emphasis on transparency and the inclusion of marginalized groups during the program.

Our findings also resonate with the mixed CDD impacts documented in related research. In the Philippines, Labonne and Chase (2008) find that CDD increased participation in village assemblies and interaction between residents and village leaders but did not initiate broader social change. Voss (2008) uncovers mixed impacts of the Kecamatan Development Program in Indonesian household welfare and access to services. Focusing on roads constructed in the same program, Olken (2007) finds that enhanced top down project monitoring through government audits was more effective in reducing corruption than increased grassroots participation in village-level accountability meetings. Fearon, Humphreys and Weinstein's (2009, 2011) randomized evaluation of a Liberian community driven post-war reconstruction project finds positive impacts on contributions to a public goods game in one of two treatment arms, but no evidence of program spillovers on contributions to existing public goods or speaking up in meetings. Beath et al (2011) show that an experimental CDD program in Afghanistan led to moderate positive impacts on community economic well-being and attitudes towards government, yet again with few impacts on collective action. Avdeenko and Gilligan (2012) find no CDD impacts on lab experiment public goods and trust games in Sudan.

Turning to empirical methods, this paper underscores the importance of pre-analysis plans (PAPs) to limit data mining and generate appropriately sized statistical tests, and discusses

some of the practical tradeoffs we faced in implementation. We confront the fundamental tension between researcher discretion versus commitment, and argue that flexibility to explore questions that arise as the research and project unfold is sometimes desirable yet should only be exercised in tandem with complete transparency over deviations from the *ex ante* specifications. In the context of a PAP, limited flexibility with full transparency allows the scholarly community to make its own assessments about the credibility of different results. We show how misleading an undisciplined interpretation of treatment effects can be in the absence of a PAP by constructing two opposing and equally erroneous narratives based on our data.

As the results of this paper concern one program in one country, any general policy implications are clearly speculative. However, we can conclude with certainty that more research is needed to identify the interventions that can successfully promote inclusive collective action. Employing a pre-analysis plan may enhance the credibility of such pursuits.

REFERENCES

Acemoglu, Daron, Simon Johnson and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review*, 91(5): 1369-1401.

Acemoglu, Daron and James A. Robinson. 2008. "Persistence of Power, Elites, and Institutions." *American Economic Review,* 98(1): 267-293.

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken and Julia Tobias. 2012. "Targeting the Poor: Evidence from a Field Experiment in Indonesia." *American Economic Review*, 104(2): 1206-1240.

Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedaian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association,* 103(484):1481-1495.

Avdeenko, Alexandra, and Michael J. Gilligan. 2012. "Community-Driven Development and Social Capital: Lab-in-the-Field Evidence from Sudan", unpublished working paper, NYU.

Banerjee, Abhijit and Lakshmi Iyer. 2005. "History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India." *American Economic Review,* 95(4): 1190-1213.

Bardhan, Pranab. 2002. "Decentralization of Governance and Development." *Journal of Economic Perspectives*, 16(4): 185-205.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. 2009. "Powerful Women: Does Exposure Reduce Bias?" *Quarterly Journal of Economics*, 124(4): 1497-1540.

Beath, Andrew, Fotini Christia, and Ruben Enikolopov. 2011. "Winning Hearts and Minds through Development Aid: Evidence from a Field Experiment in Afghanistan", unpublished working paper.

Bruhn, Miriam, and David McKenzie. 2009. "In pursuit of balance: Randomization in practice in development field experiments." *American Economic Journal: Applied Economics*, 1(4):

200-232.
Card, David and Alan B. Krueger. 1995. "Time-Series Minimum-Wage Studies: A Meta-analysis." *American Economic Review*, 85(2):238-243.
Casey, Katherine, Rachel Glennerster and Edward Miguel. 2011. "The GoBifo Project Evaluation Report: Assessing the Impacts of Community Driven Development in Sierra Leone." Final report submitted to The World Bank.
Chambers, Robert. 1983. *Rural Development: Putting the First Last*. London: Longman.
Chattopadhyay, Raghabendra and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica,* 72(5): 1409-1443.
Dal Bó, Pedro, Andrew Foster and Louis Putterman. 2010. "Institutions and Behavior: Experimental Evidence on the Effects of Democracy." *American Economic Review*, 100(5): 2205-2229.
Dongier, Philippe, Julie Van Domelen, Elinor Ostrom, Andrea Rizvi, Wendy Wakeman, Anthony Bebbington, Sabina Alkire, Talib Esmail and Margaret Polski. 2003. "Chapter 9: Community-Driven Development." In *The Poverty Reduction Strategy Sourcebook Volume 1*, 301-331.Washington DC: The World Bank.
Easterly, William. 2006. *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Little.* New York: Penguin.
Engerman, Stanley L. and Kenneth L. Sokolof. 1997. "Factor Endowments, Institutions, and Differential Paths of Growth Among New World Economies: A View from Economic Historians of the United States." In *How Latin America Fell Behind*. Stanford: Stanford University Press.
Fearon, James, Macartan Humphreys and Jeremy M. Weinstein. 2009. "Development Assistance, Institution Building, and Social Cohesion after Civil War: Evidence from a Field Experiment in Liberia." Center for Global Development Working Paper 194.
Fearon, James, Macartan Humphreys and Jeremy M. Weinstein. 2011. "Democratic Institutions and Collective Action Capacity: Results from a Field Experiment in Post-Conflict Liberia", unpublished working paper, Stanford University.
Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph Newhouse, Heidi Allen, Katherine Baicker, and the Oregon Health Study Group. 2012. "The Oregon Health Insurance Experiment: Evidence from the First Year", forthcoming *Quarterly Journal of Economics*.
Germany, Kent B. 2007. *New Orleans after the Promises: Poverty, Citizenship, and the Search for the Great Society*. University of Georgia Press: Athens, GA.
GoBifo Project. 2006. "GoBifo Overall Budget." Project mimeograph.
GoBifo Project. 2007. "Operations Manual: Version 2 June 2007." Project mimeograph.
GoBifo Project. 2008. "Results towards Goals and Objectives." Project mimeograph.
Gugerty, Mary Kay and Michael Kremer. 2008. "Outside Funding and the Dynamics of Participation in Community Associations." *American Journal of Political Science,* 52(3):585-602.
Horton, R. **and R. Smith.** 1999. "Time to register randomized trials." *British Medical Journal*, 319:865.
Humphreys, Macartan, Raul Sanchez de la Sierra and Peter van der Windt. 2012. "Fishing, Commitment and Communication: A Proposal for Comprehensive Nonbinding Research Registration" forthcoming, *Political Analysis*.
Keen, David. 2003. "Greedy Elites, Dwindling Resources, Alienated Youths: The Anatomy of

Protracted Violence in Sierra Leone." *International Politics and Society,*2: 67-94.
Kling, Jeffrey R. and Jeffrey B. Liebman. 2004. "Experimental Analysis of Neighborhood Effects on Youth." Princeton University Manuscript.
Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects", *Econometrica*, 75(1): 83-119.
Labonne, Julien and Robert Chase. 2008. "Do Community-Driven Development Projects Enhance Social Capital? Evidence from the Philippines." World Bank Policy Research Working Paper 4678.
Leamer, Edward E. 1974."False Models and Post-Data Model Construction." *Journal of the American Statistical Association*, 69(345):122-131.
Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review,* 73(1): 31-43.
Loder E, Groves T, MacCauley D. 2010. "Registration of observational studies: the next step toward research transparency." *British Medical Journal,* 340:375–376.
Mamdani, Mahmood. 1996. *Citizen and Subject: Contemporary Africa and the Legacy of Late Colonialism*. Princeton, NJ: Princeton University Press.
Mansuri, Ghazala and Vijaynedra Rao. 2004. "Community-Based and -Driven Development: A Critical Review." *World Bank Research Observer,* 19(1): 1-39.
Mansuri, Ghazala and Vijayendra Rao. 2012. *Localizing Development: Does Participation Work?* Washington, DC: The World Bank.
Narayan, Deepa. 2002. *Empowerment and Poverty Reduction: A Sourcebook*. Washington, Dc: The World Bank.
Neumark, David. 1999. "The Employment Effects of Recent Minimum Wage Increases: Evidence from a Pre-specified Research Design", NBER Working Paper #7171.
Neumark, David. 2001. "Evidence on Employment Effects of Recent Minimum Wage Increases from a Pre-Specified Research Design", *Industrial Relations*, 40(1): 121-144.
Oates, Wallace E. 1999. "An Essay on Fiscal Federalism." *Journal of Economic Literature*, 37(3):1120-1149.
O'Brien, Peter C. 1984. "Procedures for Comparing Samples with Multiple Endpoints." *Biometrics,* 40(4):1079-1087.
Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy,* 115(2): 200-249.
Olken, Benjamin A. 2009. "Corruption Perceptions vs. Corruption Reality." *Journal of Public Economics,* 93(7-8): 950-964.
Olken, Benjamin A. 2010. "Direct Democracy and Local Public Goods: Evidence from a Field Experiment in Indonesia." *American Political Science Review*, 104(2):243-267.
Olken, Benjamin A., Junko Onishi, and Susan Wong. 2010. "Indonesia's PNPM Generasi Program: Interim Impact Evaluation Report." Jakarta: The World Bank.
Olson, Mancur. 1982.*The Rise and Decline of Nations.* New Haven: Yale University Press.
Paluck, E.L., and D.P. Green. 2009. "Deference, dissent, and dispute resolution: An experimental intervention using mass media to change norms and behavior in Rwanda", *American Political Science Review*, 103, 622-644.
Pande, Rohini. 2003. "Can Mandated Political Representation Provide Disadvantaged Minorities Policy Influence? Theory and Evidence from India." *American Economic Review*, 93(4):1132-1151.
Rasmussen, Ole Dahl, Nikolaj Malchow-Møller and Thomas Barnebeck Andersen. 2011.

"Walking the talk: the need for a trial registry for development interventions." Manuscript.
Reno, William. 1995. *Corruption and State Politics in Sierra Leone.* Cambridge and New York: Cambridge University Press.
Richards, Paul. 1996. *Fighting for the Rainforest: War, Youth and Resources in Sierra Leone*. London: James Currey & Portsmouth, NH: Heinemann for the International African Institute.
Richards, Paul, Khadija Bah and James Vincent. 2004. "The Social Assessment Study: Community-driven Development and Social Capital in Post-war Sierra Leone." Manuscript.
Rosener, Judy B. 1978. "Citizen Participation: Can We Measure its Effectiveness?", *Public Administration Review*, 38(5): 457-463.
Rosenthal, Robert. 1979. "The file drawer problem and tolerance for null results", *Psychological Bulletin*, 86(3): 638-641.
Sachs, Jeffrey D. 2005.*The End of Poverty: Economic Possibilities for Our Time.* New York: Penguin Press.
Schaner, Simone. 2010. "Intrahousehold Preference Heterogeneity, Commitment and Strategic Savings: Theory and Evidence from Kenya." MIT Working Paper, http://econ-www.mit.edu/files/6221.
Sen, Amartya. 1985. *Commodities and Capabilities*. Amsterdam: Elsevier.
Sen, Amartya. 1999. *Development as Freedom*. New York: Knopf.
Simes, R.J. 1986. "Publication bias: The case for an international registry of clinical trials", *Journal of Clinical Oncology*, 4:1529-1541.
Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant", *Psychological Science*, 22(November): 1359-1366.
United Nations. 2003. *Human Development Report 2004*. New York: United Nations Development Program, Oxford University Press.
van der Laan, Mark J., and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Series in Statistics: New York.
Voss, John. 2008. "Impact Evaluation of the Second Phase of the Kecamatan Development Program in Indonesia." The World Bank, Jakarta Research Paper.**World Bank.** 2001. *World Development Report 2000/2001: Attacking Poverty.* Washington, DC: The World Bank.
World Bank. 2003. *World Development Report 2003: Sustainable Development in a Dynamic World.* Washington, DC: The World Bank.
World Bank. 2004. "Japan Social Development Fund Grant Proposal: Capacity Development to Strengthen Social Capital in Sierra Leone." Project mimeograph.

Table 1: Baseline (2005) Comparison between Treatment and Control Communities

	Baseline mean for controls	T-C difference at baseline	N
	(1)	(2)	(3)
Panel A: Community Characteristics			
Total households per community	46.76	0.30	236
		(3.67)	
Distance to nearest motorable road in miles	2.99	-0.32	236
		(0.36)	
Index of war exposure (range 0 to 1)	0.68	-0.01	236
		(0.02)	
Historical extent of domestic slavery (range 0 to 1)	0.36	0.03	236
		(0.06)	
Average respondent years of education	1.65	0.11	235
		(0.13)	
Panel B: Selected Variables from "Hardware" Family A			
Proportion of communities with a Village development committee (VDC)	0.55	0.06	232
		(0.06)	
Proportion visited by Ward Development Committee (WDC) member in past year	0.15	-0.01	228
		(0.05)	
Proportion of communities with a functional grain drying floor	0.23	0.05	231
		(0.05)	
Proportion of communities with a functional primary school	0.41	0.08	230
		(0.06)	
Average household asset score	-0.06	0.11	235
		(0.08)	
Proportion of communities with any petty traders	0.54	-0.01	226
		(0.06)	
Panel C: Selected Variables from "Software" Family B			
Respondent agrees that chiefdom officials can be trusted	0.66	-0.01	235
		(0.02)	
Respondent agrees that Local Councillors can be trusted	0.61	0.00	235
		(0.02)	
Respondent is a member of credit / savings group	0.25	-0.03	235
		(0.02)	
Among males who attended a community meeting, respondent spoke publicly	0.59	-0.02	235
		(0.04)	
Among females who attended a community meeting, respondent spoke publicly	0.29	0.03	229
		(0.04)	
Respondent claimed to have voted in last local elections	0.85	-0.01	235
		(0.02)	

Notes: i) significance levels indicated by + p <0.10, * p <0.05, ** p <0.01; ii) robust standard errors; iii) the T-C difference is the pre-program "treatment effect" run on the baseline data aggregated to the village-level mean, using a minimal specification that includes only fixed effects for the district council wards (the unit of stratification) and the two balancing variables from the randomization (total households and distance to road); iv) regressions for the two balancing variables in rows 1 and 2 exclude the outcome from the set of controls; and v) see Appendix F for the T-C difference for all 94 outcomes collected in the baseline survey.

Table 2: GoBifo Treatment Effects by Research Hypothesis

Hypotheses by Family	GoBifo Mean Treatment Effect Index	Naïve p-value	FWER adjusted p-value for all 12 hypos	FWER adjusted p-value for 11 hypos in 2009 PAP
	(1)	(2)	(3)	(4)
Family A: Development Infrastructure or "Hardware" Effects				
Mean Effect for Family A (Hypotheses 1 - 3; 39 unique outcomes)	**0.298**** **(0.031)**	**0.000**		
H1: GoBifo project implementation (7 outcomes)	0.703** (0.055)	0.000	0.000	
H2: Participation in GoBifo improves the quality of local public services infrastructure (18 outcomes)	0.204** (0.039)	0.000	0.000	0.000
H3: Participation in GoBifo improves general economic welfare (15 outcomes)	0.376** (0.047)	0.000	0.000	0.000
Family B: Institutional and Social Change or "Software" Effects				
Mean Effect for Family B (Hypotheses 4 - 12; 155 unique outcomes)	**0.028 (0.020)**	**0.155**		
H4: Participation in GoBifo increases collective action and contributions to local public goods (15 outcomes)	0.012 (0.037)	0.738	0.980	0.981
H5: GoBifo increases inclusion and participation in community planning and implementation, especially for poor and vulnerable groups; GoBifo norms spill over into other types of community decisions, making them more inclusive, transparent and accountable (47 outcomes)	0.002 (0.032)	0.944	0.980	0.981
H6: GoBifo changes local systems of authority, including the roles and public perception of traditional leaders (chiefs) versus elected local government (25 outcomes)	0.056 (0.037)	0.134	0.664	0.667
H7: Participation in GoBifo increases trust (12 outcomes)	0.042 (0.046)	0.360	0.913	0.914
H8: Participation in GoBifo builds and strengthens community groups and networks (15 outcomes)	0.028 (0.037)	0.450	0.913	0.914
H9: Participation in GoBifo increases access to information about local governance (17 outcomes)	0.038 (0.037)	0.301	0.913	0.913
H10: GoBifo increases public participation in local governance (18 outcomes)	0.090* (0.045)	0.045	0.315	0.322
H11: By increasing trust, GoBifo reduces crime and conflict in the community (8 outcomes)	0.010 (0.043)	0.816	0.980	0.981
H12: GoBifo changes political and social attitudes, making individuals more liberal towards women, more accepting of other ethnic groups and "strangers", and less tolerant of corruption and violence (9 outcomes)	0.041 (0.043)	0.348	0.913	0.914

Notes: i) significance levels (naive p-value) indicated by + p <0.10, * p <0.05, ** p <0.01; ii) robust standard errors; iii) includes fixed effects for the district council wards (the unit of stratification) and the two balancing variables used in the randomization process--total households per community and distance to nearest motorable road; iv) these mean effect estimates are limited to endline data only and the full sample set of outcomes that excludes all conditional outcomes (i.e. those that depend on the state of another variable--for example, quality of infrastructure depends on the existence of the infrastructure); v) construction of the mean effects index in Column 1 gives equal weight to each component (following Kling, Liebman and Katz 2007) as specified in the PAP; vi) familywise error rate (FWER) adjusted p-values limit the probability of making any Type I errors when considering the hypotheses as a group, where the group is defined as the final set of 12 hypotheses or the original 11 hypotheses in the pre-analysis plan (Westfall and Young 1993 free step-down resampling method as detailed in Anderson 2008); and vii) for the complete list of all variables under each hypothesis--including the exact wording of survey questions and treatment effect estimates--see Appendix J.

Table 3: GoBifo Treatment Effects by Hypothesis, Alternative Specifications

Hypotheses by Family	Covariance weighting (Anderson 2008)	SUR approach (Kling and Liebman 2004)	Include panel data	Include full set of controls	Exclude replacement households (attrition)	Include conditional outcomes	Restrict to 2005 hypotheses
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Family A: Development Infrastructure or "Hardware" Effects							
H1: Project Implementation	0.922**	0.700**	0.688**	0.695**	0.706**	0.471**	
	(0.056)	(0.052)	(0.063)	(0.055)	(0.056)	(0.058)	
H2: Local public services	0.233**	0.203**	0.179**	0.206**	0.205**	0.099*	0.149**
	(0.040)	(0.040)	(0.040)	(0.039)	(0.039)	(0.040)	(0.048)
H3: Economic welfare	0.565**	0.371**	0.362**	0.362**	0.375**	0.271**	0.222**
	(0.050)	(0.046)	(0.047)	(0.045)	(0.048)	(0.037)	(0.057)
Family B: Institutional and Social Change or "Software" Effects							
H4: Collective action	-0.043	0.016	0.038	0.011	0.014	-0.040	0.134*
	(0.036)	(0.036)	(0.042)	(0.036)	(0.037)	(0.031)	(0.059)
H5: Inclusion of vulnerable groups	0.000	0.001	0.002	0.000	0.004	0.015	0.067
	(0.029)	(0.030)	(0.030)	(0.031)	(0.032)	(0.027)	(0.116)
H6: Local authority	0.050	0.056	0.051	0.052	0.039	0.053	-0.006
	(0.035)	(0.036)	(0.036)	(0.037)	(0.037)	(0.033)	(0.070)
H7: Trust	0.039	0.042	0.047	0.036	0.048	0.028	0.021
	(0.046)	(0.044)	(0.061)	(0.046)	(0.046)	(0.043)	(0.050)
H8: Groups	0.031	0.027	0.03	0.027	0.045	0.007	-0.048
	(0.037)	(0.035)	(0.039)	(0.037)	(0.037)	(0.034)	(0.054)
H9: Information about governance	0.017	0.037	0.028	0.031	0.045	0.033	0.097*
	(0.038)	(0.035)	(0.040)	(0.036)	(0.037)	(0.035)	(0.043)
H10: Participation in governance	0.160**	0.092**	0.084+	0.082+	0.088+	0.131**	0.088+
	(0.044)	(0.043)	(0.045)	(0.044)	0.046	(0.045)	(0.050)
H11: Crime and conflict	0.041	0.010	0.027	0.014	-0.013	0.011	0.010
	(0.048)	(0.041)	(0.054)	(0.043)	(0.042)	(0.039)	(0.068)
H12: Political and social attitudes	-0.011	0.040	0.040	0.035	-0.011	0.005	
	(0.044)	(0.041)	(0.041)	(0.044)	(0.046)	(0.037)	

Notes: i) significance levels (naive p-value) indicated by + p <0.10, * p <0.05, ** p <0.01; ii) robust standard errors; iii) includes fixed effects for the district council wards (the unit of stratification) and the two balancing variables from the original randomization--total households per community and distance to nearest motorable road; iv) outcomes included per hypothesis vary by column: Columns 1 - 5 include full sample outcomes only (184 unique outcomes in total), Column 6 includes both full sample and conditional outcomes (i.e. those that depend on the state of another variable--for example, quality of infrastructure depends on the existence of the infrastructure, 334 unique outcomes in total), and Column 7 include 63 unique outcomes (see xi below); v) Column 1 weights each index component by the inverse of the appropriate element of the variance-covariance matrix (as in Anderson 2008) where the matrix is estimated in the control group (zero replaces any negative estimated weights); vi) Column 2 uses stacked OLS outcome-by-outcome as in Kling and Liebman 2004; vii) Column 3 uses the Kling and Liebman 2004 approach incorporating panel data where available; viii) Column 4 uses Kling et al. 2007 approach with the full set of control variables specified in the PAP; ix) Column 5 uses Kling et al. 2007 and excludes all endline survey replacement individuals and households; x) Column 6 uses Kling and Liebman 2004 and includes outcome measures that apply only to a subset of observations (note five variables from the PAP were omitted due to insufficient observations: community financial contributions to peripheral health unit, palava hut, market and grainstore (H2 and H4) and existence of football equipment (H2)); and xi) Column 7 uses Kling et al. 2007 restricted to the hypotheses written down in the 2005 pre-program document and to full sample outcomes included in the baseline 2005 survey.

Table 4: Illustrative Selection of Statistically Significant Treatment Effects, Family A

Outcome Variable	Mean in Controls	Treatment Effect	Standard Error	N
	(1)	(2)	(3)	(4)
Panel A: Hypothesis 1 - Project Implementation				
Village development committee	0.46	0.40**	(0.05)	235
Visit by WDC member	0.21	0.13*	(0.06)	234
Village development plan	0.62	0.30**	(0.05)	221
Community bank account	0.08	0.71**	(0.05)	226
A local politician was involved in managing the infrastructure:				
Primary School	0.42	0.18**	(0.06)	138
Grain drying floor	0.24	0.13*	(0.06)	115
Latrine	0.22	0.16**	(0.04)	169
Panel B: Hypothesis 2 - Local Public Services				
Functional traditional midwife post in the community	0.08	0.17**	(0.04)	235
Functional latrine in the community	0.46	0.21**	(0.06)	234
Functional community center in the community	0.03	0.09**	(0.03)	236
Community took a proposal to an NGO or donor for funding	0.29	-0.15**	(0.05)	229
Supervisor's physical assessment of construction quality (index from 0 to 1):				
Primary School	0.58	0.11+	(0.06)	123
Grain drying floor	0.38	0.16*	(0.08)	101
Latrine	0.27	0.18**	(0.05)	154
Panel C: Hypothesis 3 - Economic Welfare				
Total petty traders in village	2.43	0.70*	(0.34)	225
Total goods on sale of 10	4.45	0.57*	(0.24)	236
Household asset score	-0.16	0.30**	(0.09)	236
Attended trade skills training	0.06	0.12**	(0.02)	235

Notes: i) significance levels (per comparison p-value) indicated by + p <0.10, * p <0.05, ** p <0.01; ii) treatment effects are estimated on endline data only; iii) robust standard errors in parentheses; iv) includes fixed effects for the district council wards (the unit of stratification) and the two balancing variables from the original randomization: total households per community and distance to nearest motorable road; and v) where indicated, outcomes are conditional on the existence of functional infrastructure in the community.

Table 5: Illustrative Treatment Effects, Structured Community Activities (SCAs)

Structured Community Activity (SCA) Outcome:	Mean for Controls	Treatment Effect	Standard Error
	(1)	(2)	(3)
Panel A. Collective Action and the Building Materials Vouchers			
GoBifo Mean Effect for SCA #1 (17 outcomes in total)	**0.00**	**0.00**	**(0.05)**
Proportion of communities that redeemed vouchers at building materials store	0.54	-0.02	(0.06)
Average number of vouchers redeemed at the store (out of six)	2.95	0.06	(0.35)
Proportion of communities that held a meeting to discuss the vouchers	0.98	-0.05*	(0.02)
Panel B. Participation in the Gift Choice Deliberation			
GoBifo Mean Effect for SCA #2 (33 outcomes in total)	**0.00**	**0.00**	**(0.04)**
Duration of gift choice deliberation (in minutes)	9.36	1.54	(1.12)
Total adults in attendance at gift choice meeting	54.51	3.57	(2.88)
Total women in attendance at gift choice meeting	24.99	1.98	(1.59)
Total youths (approximately 18-35 years) in attendance at gift choice meeting	23.57	2.06	(1.32)
Total number of public speakers during the deliberation	6.04	0.22	(0.40)
Total number of women who spoke publicly during the deliberation	1.88	-0.20	(0.22)
Total number of youths (approximately 18-35 years) who spoke publicly	2.14	0.23	(0.24)
Proportion of communities that held a vote during the deliberation	0.10	0.07	(0.04)
Panel C. Community Use of the Tarpaulin			
GoBifo Mean Effect for SCA #3 (18 outcomes in total)	**0.00**	**-0.03**	**(0.05)**
Proportion of communities that held a meeting to discuss use of the tarp	0.98	-0.03	(0.02)
Proportion of communities that stored the tarp in a public place	0.06	0.05	(0.04)
Proportion of communities that had used the tarp (5 months after receipt)	0.90	-0.08+	(0.04)
Given tarp used, proportion of communities using the tarp in a public way	0.86	0.02	(0.05)
Proportion of households that directly benefited from the tarp	0.57	-0.01	(0.04)

Notes: i) significance levels (per comparison p-value) denoted by + $p < 0.10$, * $p < 0.05$ and ** $p < 0.01$; ii) robust standard errors; iii) treatment effects estimated on follow-up data; (iv) includes fixed effects for the district council wards (the unit of stratification) and the two balancing variables from the original randomization: total households per community and distance to nearest motorable road; v) sample size varies between 225-236 for all outcomes in the table save the last, which is conditional on having used the tarp and has N = 161; and vi) the SCA-wide mean effect estimates follow Kling and Liebman (2004) to accommodate the mixture of full sample and conditional outcomes.

Table 6: Erroneous Interpretations under "Cherry Picking"

Outcome Variable	Mean for controls	Treatment effect	Standard error	N	Hypo
	(1)	(2)	(3)	(4)	(5)
Panel A: GoBifo "Weakened" Institutions					
Attended meeting to decide what to do with the tarp	0.81	-0.04+	(0.02)	236	H5
Everybody had equal say in deciding how to use the tarp	0.51	-0.11+	(0.06)	232	H5
Community used the tarp (verified by physical assessment)	0.90	-0.08+	(0.04)	233	H4
Community can show research team the tarp	0.84	-0.12*	(0.05)	232	H5
Respondent would like to be a member of the VDC	0.36	-0.04*	(0.02)	236	H10
Respondent voted in the local government election (2008)	0.85	-0.04*	(0.02)	236	H10
Panel B: GoBifo "Strengthened" Institutions					
Community teachers have been trained	0.47	0.12+	(0.07)	173	H4
Respondent is a member of a women's group	0.24	0.06**	(0.02)	236	H8
Someone took minutes at the most recent community meeting	0.30	0.14*	(0.06)	227	H5
Building materials stored in a public place when not in use	0.13	0.25*	(0.10)	84	H5
Chiefdom official did not have the most influence over tarpaulin use	0.54	0.06*	(0.03)	236	H6
Respondent agrees with "Responsible young people can be good leaders" and not "Only older people are mature enough to be leaders"	0.76	0.04*	(0.02)	236	H6, H12
Correctly able to name the year of the next general elections	0.19	0.04*	(0.02)	236	H9

Notes: i) significance levels (per comparison p-value) indicated by + p <0.10, * p <0.05, ** p <0.01; ii) robust standard errors; iii) treatment effects estimated on follow-up data; and iv) includes fixed effects for the disctrict council wards (the unit of stratification) and the two balancing variables from the randomization (total households and distance to road) as controls.