

Supplemental Appendix

Time is not money: an experiment with community contribution requirements in cash and labour

Serena Cocciolo, Selene Ghisolfi, Md. Ahasan Habib, and Anna Tompsett*

January 30, 2026

*Cocciolo: World Bank. Ghisolfi: Catholic University of the Sacred Heart and LEAP, Bocconi University. Habib: NGO Forum for Public Health and KTH Royal Institute of Technology. Tompsett: Institute for International Economic Studies, Stockholm University and Beijer Institute for Ecological Economics, Royal Swedish Academy of Sciences.

Overview

A	Details on project implementation	2
A1	Timing	2
A2	Meeting attendance and participation	6
B	Recruitment, sampling, and data	7
B1	Community definition and eligibility	7
B2	Household sample	9
B2.1	Baseline	9
B2.2	Endline	9
B3	Matching households to water sources	10
B3.1	Water testing	12
B4	Sampling weights	14
B5	Monitoring data	15
C	Randomization checks	16
D	Simulations to support empirical specifications	21
D1	Cross-sectional data, comparison within treatment group only	21
D2	Panel data, balanced assignment to treatment across strata	22
D3	Panel data, unbalanced assignment to treatment across strata	25
D4	“Household-level” data	26
E	Using machine learning to understand selection	28
E1	Heterogeneous selection across treatment arms	28
E2	Characteristics that are correlated with selection on average	31
F	Extended data on contributions	32
F1	Commitment to contribute	32
F2	Differences in numbers of contributors across treatment arms	32
F3	Comparing project records to self-reported data	34
F4	Contribution amounts	35
F5	Within-household data on contributions	36
G	Follow-up experiment	40
H	Self-reported evaluations and motivations	42
I	Robustness of effects on program impact	48
I1	Declines in arsenic exposure in the control group	48
I2	Specification tests	48
J	Measuring location quality	50
K	Decomposition of well use	55
L	Selection	61
L1	Describing community-level selection into well installation	61
L2	Describing household-level selection into well use	69
L3	Describing household-level selection into contribution	76
M	Screening	84
N	Cost data	85

A Details on project implementation

A1 Timing

Appendix Figure A1 provides an overview of the overall project timeline. Appendix Figure A2 shows in more detail the timing of community meetings and well installations over the two-year project implementation period. Appendix Figure A3 plots the time between the first meeting held in a community and the first successful installation, for communities that successfully installed at least one well. Appendix Figure A4 shows the timing of meetings and installations relative to the local agricultural season. Community meetings were held throughout the year, while installations took place largely in two waves, both coincidentally primarily during the rain season. Appendix Figure A5 shows nonparametrically that the rate of success of installations is largely insensitive within treatment arms to the timing of meetings or installation attempts.

Figure A1: Project timeline

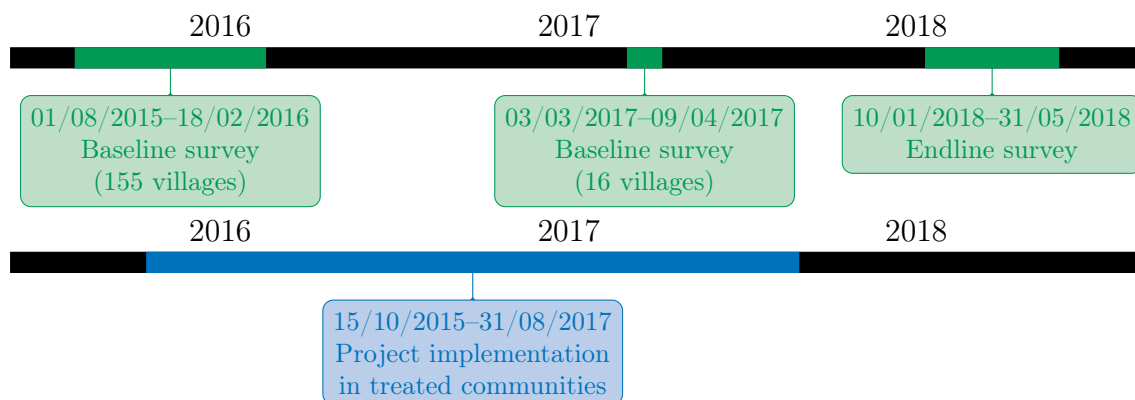
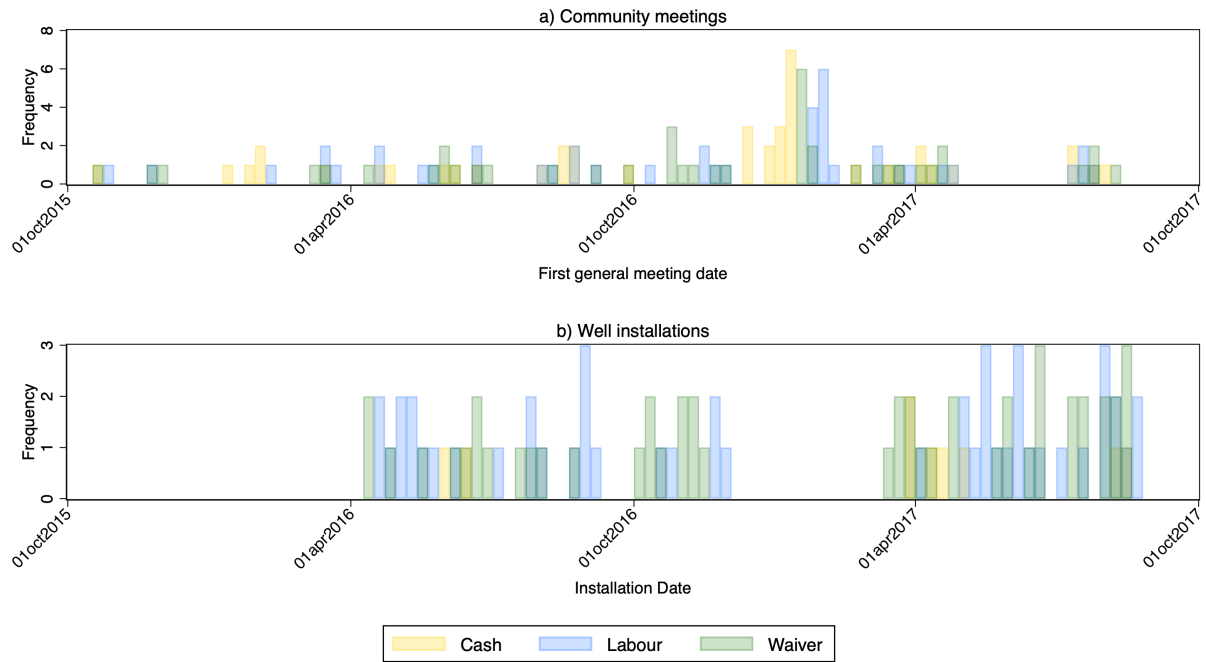
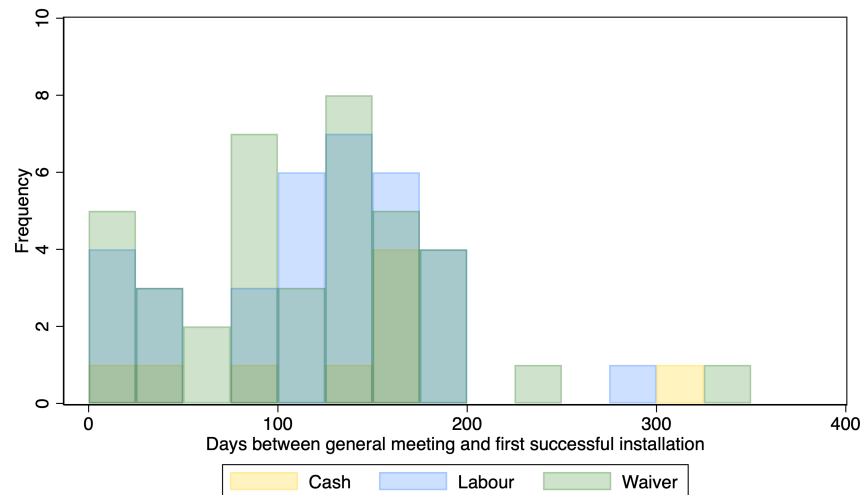


Figure A2: Timing of community meetings and installation



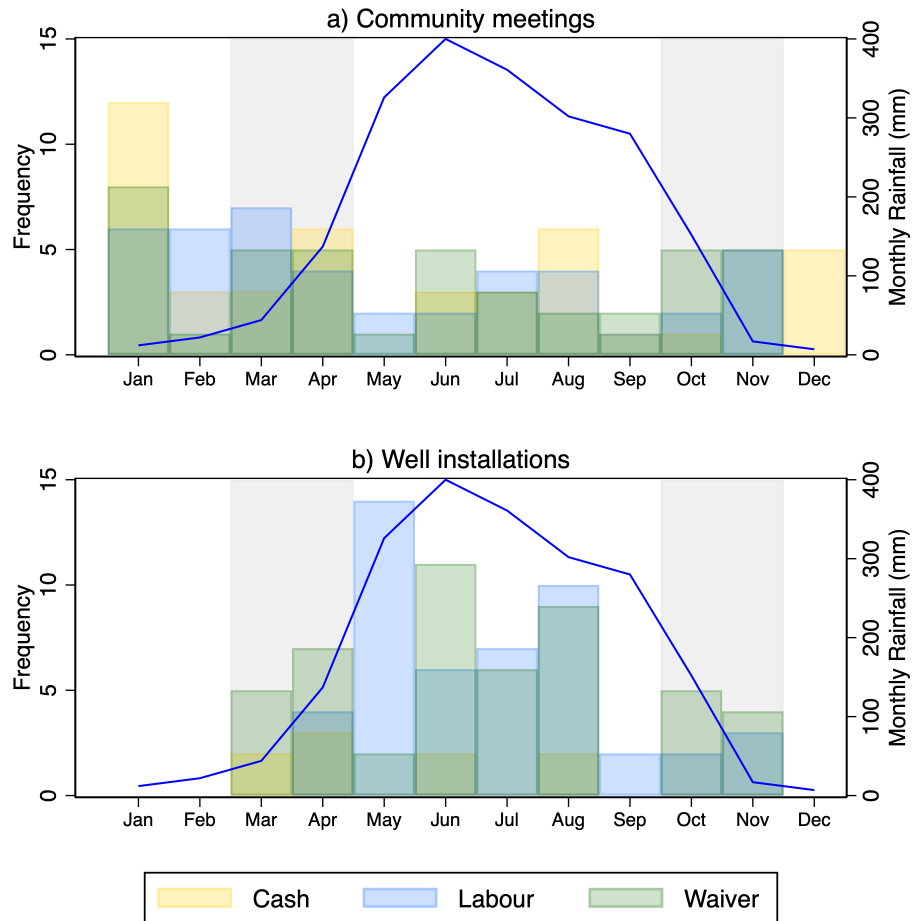
Notes Histograms show frequency of community meetings in panel a) and well installations in panel b).

Figure A3: Time between community meeting and installation



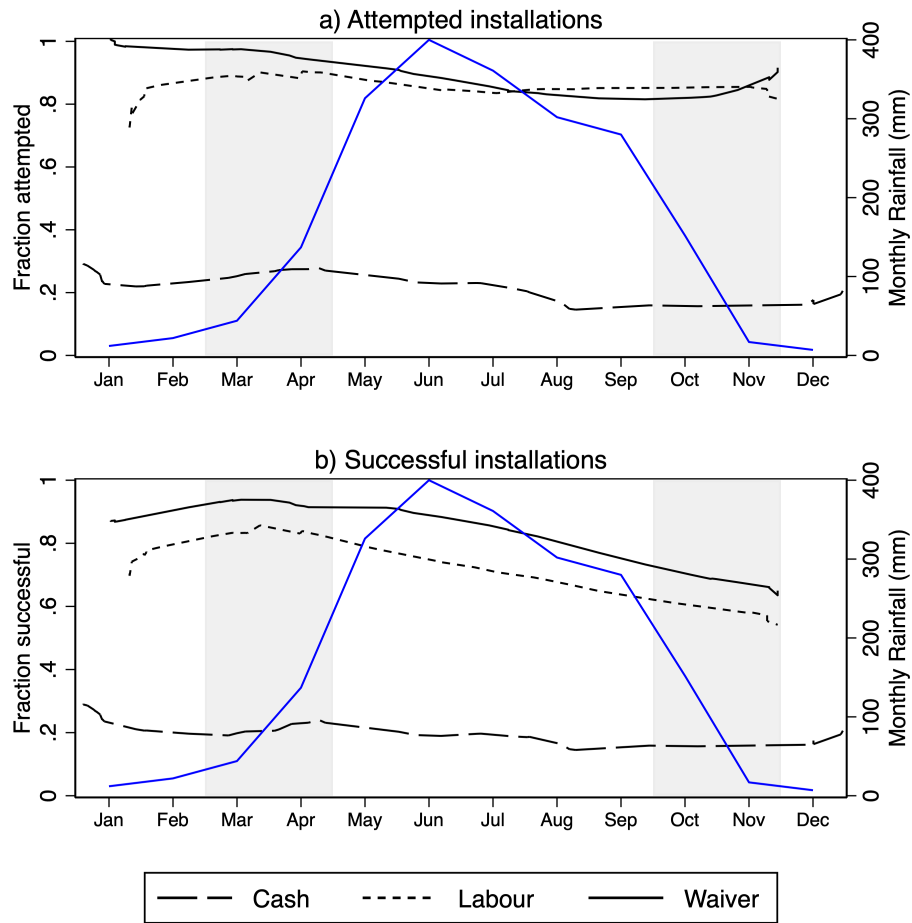
Notes Histograms show frequency of number of days between community meeting and first successful well installation. Sample includes communities that successfully install wells only. If the community held multiple meetings, the date of the first meeting is used.

Figure A4: Timing of community meetings and installation: by month of year



Notes Histograms show frequency of community meetings in panel a) and well installations in panel b). Monthly rainfall shown in blue, recovered from climate-data.org. “Lean seasons” shaded in grey, based on FAO classifications (GIEWS, 2021).

Figure A5: Take-up and successful installation: by month of year



Notes Shows results of locally linear regressions where the outcome variable is the fraction of offered wells for which the community completed project requirements and for which we attempted installation (panel a) or the fraction of offered wells successfully installed (panel b) and the explanatory variable is the Julian date of the first meeting. Monthly rainfall shown in blue, recovered from climate-data.org. “Lean seasons” shaded in grey, based on FAO classifications (GIEWS, 2021).

A2 Meeting attendance and participation

Appendix Tables A1 and A2 report attendance rates at community meetings, based on project records and retrospective survey data respectively.

Table A1: Participation in community meeting (project records)

	Cash	Labour	Waiver	C = W	L = W	C = L	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Attended	0.49 (0.02)	0.52 (0.02)	0.50 (0.02)	0.954	0.406	0.368	20 348
Female attended	0.27 (0.02)	0.28 (0.02)	0.26 (0.02)	0.690	0.468	0.886	20 348
Attended (lowest asset quintile)	0.44 (0.03)	0.46 (0.03)	0.51 (0.03)	0.124	0.444	0.656	861
Attended (highest asset quintile)	0.55 (0.03)	0.54 (0.03)	0.54 (0.03)	0.528	0.926	0.792	890

Notes Table reports means and clustered standard errors in parentheses. Attended means that field staff recorded that at least one member of the household, or one female member where specified, attended the meeting. Sample includes all listed households (rows 1 and 2) and households surveyed at baseline in the listed asset quintile (rows 3 and 4). Weights are applied so that each community counts equally in summary statistics. Significance levels obtained from randomization-based inference. Regressions control for centered strata indicators interacted with the treatment dummies, arbitrarily omitting one stratum. . p values from pairwise hypothesis tests in columns 4-6.

Table A2: Participation in community meeting (retrospective self-reports)

	Cash	Labour	Waiver	C = W	L = W	C = L	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Attended	0.55 (0.02)	0.59 (0.03)	0.58 (0.02)	0.276	0.990	0.370	5927
Female attended	0.13 (0.01)	0.09 (0.01)	0.10 (0.01)	0.036	0.888	0.054	6120
Attended (lowest asset quintile)	0.48 (0.03)	0.52 (0.04)	0.55 (0.03)	0.144	0.604	0.488	1193
Attended (highest asset quintile)	0.67 (0.02)	0.62 (0.03)	0.63 (0.03)	0.292	0.710	0.456	1221

Notes Table reports means and clustered standard errors in parentheses. Attended means that the household reported that at least one member, or one female member where specified, attended the meeting retrospectively at endline. Sample includes all randomly-sampled households surveyed at endline (rows 1 and 2) in the listed asset quintile (rows 3 and 4). Weights are applied so that each community counts equally in summary statistics. Significance levels obtained from randomization-based inference. Regressions control for centered strata indicators interacted with the treatment dummies, arbitrarily omitting one stratum. p values from pairwise hypothesis tests in columns 4-6.

B Recruitment, sampling, and data

The sample of communities and households was defined through several steps, illustrated in Figure 2 and in more detail in Appendix Figure B1.

B1 Community definition and eligibility

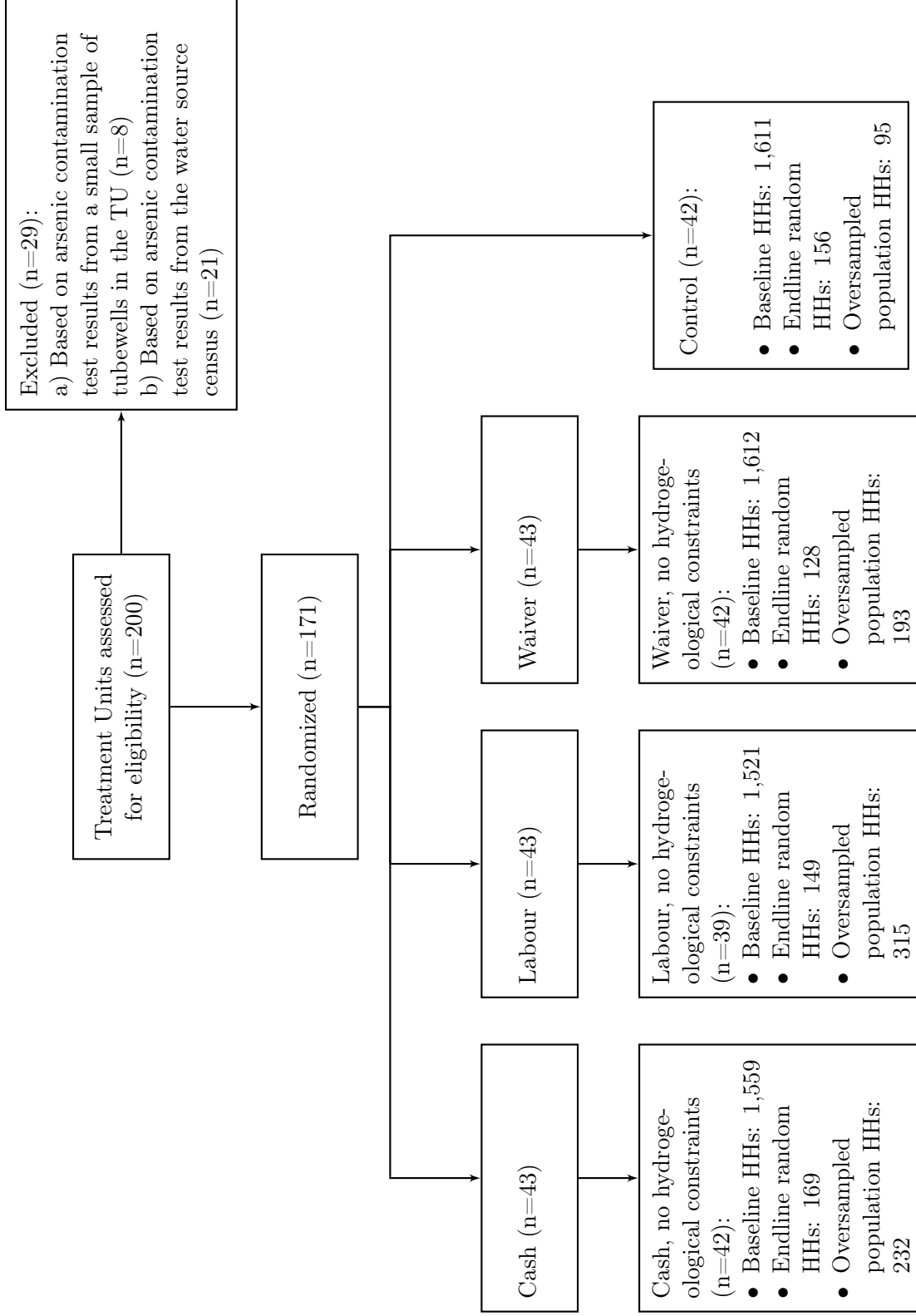
Our intervention is targeted towards communities exposed to arsenic contamination. Unfortunately, we had only limited data on arsenic contamination in the area when we began the study. We therefore used the data available to pre-select communities for inclusion and then used our own testing data to confirm eligibility.

We identified candidate villages for the intervention on the basis of contamination levels reported in the available sources of arsenic testing data. We had access to village-level data from the following data sources: (i) data from the Bangladesh Arsenic Mitigation Water Supply Project (BAMWSP), which included a large tubewell screening program conducted between 1999 and 2006; (ii) an assessment from the Department of Public Health Engineering (DPHE) on the most arsenic-contaminated villages in the Bogra region; (iii) data collected in 2008 by the Bangladesh Social Development Services (BSDS). We selected as candidate villages for the intervention all villages indicated by the DPHE or for which BAMWSP or BSDS data reported a share of arsenic contaminated tubewells equal or higher than 30%. We confirmed this initial selection by testing for arsenic contamination in a small sample of tubewells in the village.

For each of these villages, we then obtained the most recent list of resident households from administrative sources. For logistical reasons, we implement the program in geographically defined communities of between 50 and 250 households, which are smaller than many villages. To define these communities, we used the household lists to obtain village sizes, excluded villages with less than 50 households, and divided larger villages into several smaller communities along natural boundaries. Following this process, we identified 192 candidate communities in 103 villages, of which 51 were divided in two or more communities. We conducted a full census of existing sources of drinking water in these candidate communities.

We used the water source contamination data in order to finalize the selection of the communities eligible for receiving the arsenic mitigation program. Specifically, we excluded from the study all communities with less than 15% of arsenic contaminated water sources. We further screened communities with less than 25% of arsenic-contaminated water sources, including them in the program only if they presented a well defined cluster of contaminated water sources. To evaluate these communities with between 15% and 25% contamination, we reviewed the maps obtained from the water source census. We excluded communities where arsenic contaminated water sources were geographically scattered, because in these cases all households in the village already had a nearby source of arsenic-safe water.

Figure B1: Sample definition



The administrative data available to define communities was of limited quality. In particular, the available administrative lists sometimes did not correspond to the correct geographical location of households in the villages. As a result, we sometimes needed to redefine community boundaries after collecting household-level data. We did this by reviewing the geo-coordinates of households and water sources collected during the baseline survey and via field visits by field staff. After several rounds of refinements, we were largely successful in correctly defining communities with a comprehensive household list corresponding to a meaningful geographical entity. Communities were finalized before assignment to treatment. However, as a consequence of the redefinition process, the baseline household sample was not equally distributed, with a larger number of households interviewed in some communities and a smaller number of households interviewed in others. At endline, we corrected for these disparities by sampling additional randomly selected households in communities where at baseline we interviewed less than 40 households. Throughout, we correct for any disparities in the number of sampled households in each community using sampling weights that ensure that all communities count equally in summary statistics (Appendix B4).

B2 Household sample

B2.1 Baseline We used the available household administrative lists to randomly sample 40 households in each community for the baseline household survey. We accommodated cases when selected households were not available for the interview or refused to participate by providing enumerators with a list of potential replacement households, sorted in random order. Enumerators documented this replacement process on paper in the household list used by the enumerators and digitally using the survey form. Enumerators completed a form for all households that they attempted to locate and interview.

In 92% of cases the enumerators were able to conduct the interview with the originally-sampled household. When this was not possible, the reason was that the household was not found in 33% of cases, that no one was at home during the visit from our enumerator in 65% of cases, or that the respondent refused to participate in the survey in 2% of cases. Enumerators conducted the interview with the household head, their spouse, or another adult representative of the household. They always asked for the respondent’s informed consent, both for the interview and, separately, for the water testing. At baseline, 99.8% of contacted households agreed to the interview, and we successfully conducted the household survey in a total of 6,529 households across 171 eligible communities.

Occasionally, the number of households surveyed in a community was higher or lower than the targeted 40 households. This is because in some cases we had to revise the community definition after completing the household surveys and reviewing the locations of households: in some cases, the administrative lists had misassigned households to geographical clusters. We reassigned the households so that each community retained geographical consistency.

B2.2 Endline At endline, we attempted to survey all households interviewed at baseline plus two sets of additional households. First, because in some communities we interviewed fewer than 40 households at baseline, we randomly sampled additional households from the full household list to achieve a minimum representative sample of 40 households per community at endline. This added 616 households to the endline sample. Second, we oversampled

several specific groups at endline: households identified at baseline as community leaders, households with high baseline network centrality, and households that played a significant role during the program implementation (e.g., households that contributed cash or labour, donated land for tubewell installation, or were elected for management and maintenance of the new water source). We identified 866 of these households.

In total, we selected 8,011 households for the endline interview, of which 6,529 had been interviewed at baseline. Six of the oversampled households appeared to be a duplicate listing of another record. Otherwise, the enumerators successfully completed the interview with 99.6% of the households selected to participate in the endline household survey.

We were unable to complete the interview with 5 households that migrated, 4 households with no surviving household member, and 5 households that declined to participate. For 17 households, either the survey was not carried out or the data was lost. The attrition rate between the baseline and the endline survey is 0.21%. Among households that we were able to successfully contact at endline, 99.9% agreed to the interview.

B3 Matching households to water sources

Our data match households to the water sources they use. Matching households to decentralized infrastructure is not straightforward. We extensively piloted our matching procedures in the field and built a number of checks into the process.

At baseline, we conducted a full census of existing sources of drinking water. Enumerators visited all households residing in the community and asked for an exhaustive list of nearby water sources. We used the existing administrative household list to structure the water source census, and we collected information on households missing from that list during the census process. We included public water sources in the census.

We then conducted the baseline household survey in the randomly selected sample of households. Each household identified the water source(s) used to obtain water for drinking or cooking purposes, selecting water sources from the list established during the water source census. We showed the respondent a picture of each water source that he/she identified, to ensure that we correctly matched households to water sources. If the respondent reported using a water source not included in the water source census data, we collected the relevant information from this new source. This happened in only 2% of the household surveys, indicating good coverage in the census.

At endline, we did not repeat the water source census because of costs. Instead, we first conducted the household survey and then collected data from all the water sources that households reported using. Respondents could identify the water sources they used from the full baseline list of existing water sources, augmented, in treated communities, with the new project tubewell(s) installed by the project. To avoid resurveying and retesting water sources multiple times, we tagged each tested water source with a zip tie. If an enumerator visited a source that had already been registered and tested at endline, they took a photograph and recorded GPS coordinates, enabling us to confirm the match to the water source data already collected by another enumerator. If the respondent reported using a water source not included in the list, we collected the relevant information from this new source. This happened in 17% of households at endline. Most of these households were using new private wells that had been constructed between baseline and endline.

Most households consistently report using a single source, almost exclusively a shallow tubewell belonging to their own household or a close neighbour or relative. We asked households to list all water sources that they use in any season for drinking or cooking. We also coded follow-up questions to check that no source had been missed that was used in a different season. However, most households report only using one source throughout the year. Most households in the control group also report using the same source at baseline and endline, even though the surveys were not necessarily in the same season: only 13% of control households report using any source at endline that they did not use at baseline.

A measure of internal consistency in our data is whether households listed a project well in the water source roster that records all sources each household uses for drinking and cooking, and reported using a project well in a section on project implementation. Household reports were very consistent across these two different sections of the survey (Appendix Table B1).

Table B1: Internal consistency of self-reported data on user status

		Water source roster	
		Uses well	Does not use
a) Cash contribution requirement			
Implementation section	Uses well	0.43	0.02
	Does not use well	0.01	0.54
b) Labour contribution requirement			
Implementation section	Uses well	0.35	0.04
	Does not use well	0.00	0.61
c) Contribution waiver			
Implementation section	Uses well	0.35	0.03
	Does not use well	0.00	0.62

Notes Table compares whether household self-reported use of a project well in the water source roster and in the project implementation section. Communities in which we successfully installed at least one project well only.

The fidelity of our data depends on whether households accurately report which sources they use, and in particular, on whether households change how they report the sources they use in response to treatment. We test for evidence of differential reporting bias by examining the correlation between contamination in sources that households report using and contamination in household drinking water. While there are other reasons why this correlation could vary in response to treatment—for example, if households change whether and for how long they store drinking water in their homes—an increase in reporting bias should have a specific signature: a *decrease* in the correlation between source and household contamination and an *increase* in the share of household contamination that is unexplained by source contamination. Source and household drinking water contamination are positively correlated at baseline, and that these correlations do not vary systematically by treatment arm (columns 1 and 2). The same is largely true at endline (columns 3 and 4) and for changes in contamination between baseline and endline (columns 5 and 6). Any change in the relationship between source and household contamination between baseline and endline moves in the opposite direction to what we would expect if reporting bias were present.

Table B2: Testing for evidence of reporting bias

	Contamination in household drinking water					
	(1)	(2)	(3)	(4)	(5)	(6)
Source contamination	0.770*** (0.029)	0.687*** (0.048)	0.815*** (0.020)	0.729*** (0.028)	0.634*** (0.033)	0.557*** (0.038)
Source contamination x cash	-0.020 (0.044)	0.001 (0.060)	0.016 (0.030)	0.051 (0.045)	-0.100* (0.059)	-0.010 (0.056)
Cash	0.020 (0.023)	-0.000 (0.008)	0.015 (0.011)	0.005 (0.004)	0.012 (0.025)	0.024 (0.019)
Source contamination x labour	-0.050 (0.041)	-0.062 (0.057)	0.020 (0.028)	-0.014 (0.042)	-0.024 (0.047)	-0.052 (0.053)
Labour	0.046* (0.027)	0.027** (0.011)	0.025** (0.013)	0.015*** (0.004)	0.012 (0.024)	0.001 (0.020)
Source contamination x waiver	0.007 (0.044)	0.016 (0.058)	0.022 (0.028)	0.074* (0.039)	-0.034 (0.050)	-0.023 (0.053)
Waiver	-0.011 (0.023)	0.012 (0.010)	0.013 (0.009)	0.018*** (0.006)	0.020 (0.023)	0.013 (0.020)
Constant	0.088*** (0.015)	0.026*** (0.006)	0.016*** (0.006)	0.001 (0.001)	-0.056*** (0.018)	-0.016 (0.012)
N	6495	6495	6420	6420	6404	6404
Contaminant	Baseline	Baseline	Endline	Endline	Change	Change
Time	As > 10	As > 50	As > 10	As > 50	As > 10	As > 50

Notes Outcome variable is household-level contamination at the listed threshold. Results shown are for a regression of household-level contamination on source contamination, a dummy for each contribution treatment, and the interaction between source contamination and each contribution treatment. Sample includes households surveyed at baseline and follow-up with non-missing data for arsenic contamination. Weights applied so that each community counts equally in summary statistics. Standard errors clustered by community. Asterisks reflect statistical significance based on analytical standard errors:

* p<0.1, ** p<0.05, *** p<0.01.

B3.1 Water testing At both baseline and endline, we conducted field arsenic tests for all households surveyed and for all water sources that any interviewed household reported collecting water from for drinking and/or cooking, provided that the survey respondent consented to testing. We tested household water quality during the household surveys. At baseline, we tested water sources either during the census or during the household survey, if a household reported using a water source that was omitted from the census. At endline, we retested water quality in all water sources that any household reported using, irrespective of whether they were listed or unlisted, during the household survey. Before collecting samples from tubewells for arsenic testing, enumerators flushed the well by pumping the tubewell handle once for every ten feet of well depth. Respectively 99.7% and 99.2% agreed to water testing at baseline and at endline.

Among households interviewed at baseline, we successfully measured arsenic in household drinking water in 6,526 households (> 99.9%). We measured arsenic in the water source(s) used by the household for 6,515 households (99.8%). At endline, we measured arsenic in household drinking water for 7,910 households (99.2% of surveyed households) and in water source(s) used by the household for 7,913 households (99.2%). We focus primarily on whether arsenic contamination exceeds the WHO (10 parts per billion or ppb) or Bangladeshi (50 ppb) safe drinking water thresholds. The raw arsenic contamination data are highly skewed, as shown in Appendix Figures B2a and B2b.

Figure B2a: CDF of change in household contamination

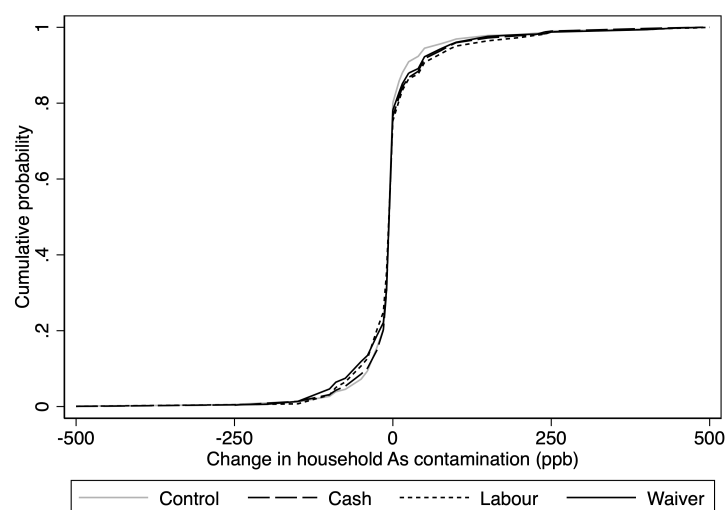
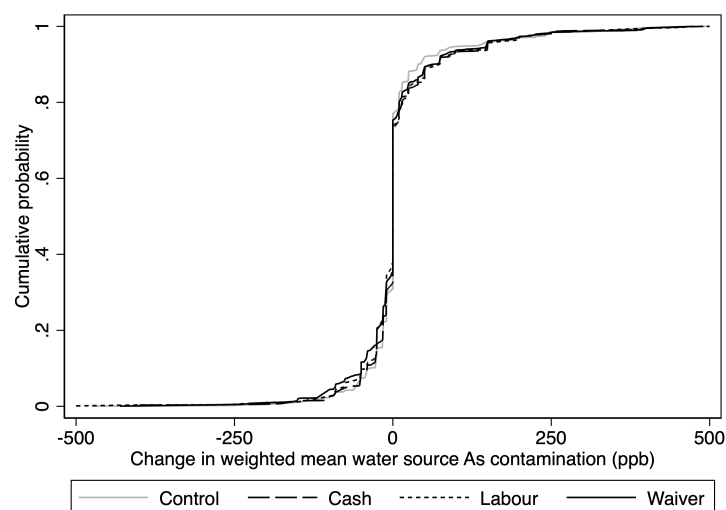


Figure B2b: CDF of change in mean water source contamination



Notes Figures show empirical cumulative distribution functions (CDF) for changes in arsenic contamination between baseline and endline.

We also collected data on fecal contamination to evaluate the broader impacts of the intervention on drinking water quality. The effect of new community water sources on fecal contamination in household drinking water is theoretically ambiguous, because of the potential for recontamination during transport and storage (see, e.g., Kremer et al., 2011). Also, while the wells we installed were essentially free from arsenic contamination, they had only slightly lower rates of fecal contamination than other wells in the same communities (Cocciolo et al., 2021). The program did not reduce exposure to fecal contamination on average (Cocciolo et al., 2021) nor under any treatment arm (Appendix Table B3). For this reason, we focus on effects on arsenic contamination in the main paper.

Table B3: Impact on fecal contamination

	Control Mean	Estimated impact			C = W	L = W	C = L	N
		Cash	Labour	Waiver				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fecal contamination index	-0.11 (0.04)	0.00 (0.05)	0.02 (0.05)	-0.05 (0.05)	0.462	0.302	0.828	13 589
Household drinking water	-0.04 (0.02)	-0.01 (0.02)	0.02 (0.02)	0.00 (0.02)	0.738	0.436	0.380	13 457
Water source	-0.04 (0.02)	0.01 (0.02)	-0.01 (0.02)	-0.03 (0.02)	0.116	0.394	0.558	13 427

Notes Table reports program impact relative to control group. The sample is all randomly-sampled households in study communities. Clustered standard errors are in parentheses. Regressions include baseline and endline data, community fixed effects, a post-treatment dummy, and centered strata indicators interacted with the post dummy and treatment arm-post dummy interactions, arbitrarily omitting one stratum. Column 1 reports mean change in control group. Weights are applied so that each community counts equally in summary statistics. Significance levels obtained from randomization-based inference. *p* values from pairwise hypothesis tests in columns 5-7.

Our original study design also included laboratory testing for arsenic in household drinking water for a subsample of households. Laboratory tests for arsenic are more accurate than field tests but much more expensive, and we did not have sufficient funds to complete all the intended laboratory tests at baseline. Additionally, correlation is weak between laboratory test results and field test results at endline, possibly reflecting a problem with tracking samples through the laboratory testing process, where we have incomplete control over tracking procedures. In the interests of brevity, we focus in the main paper only on arsenic contamination measured by field tests.

B4 Sampling weights

As described in Appendix B2, the household sample used for this study consists of a random sample of households drawn from household lists along with households from oversampled populations in some special categories of particular interest. The target number of randomly sampled households was 40 per community. Because of the errors in the administrative data used as a sampling frame, the baseline sample was not perfectly distributed across communities. At endline, we interviewed a minimum of 40 randomly sampled households per community, but some communities had more than 40 households in the sample. We use sampling weights throughout to ensure that each community counts equally in summary statistics. When we include oversampled households, we adjust sampling weights for these households to maintain representativeness and account for the share of the full population that these households represent.

B5 Monitoring data

We collected extensive project monitoring data, listed in Appendix Table B4.

Table B4: Monitoring of the intervention

Indicator	Mode of data collection
Treatment randomization	
Number of public lottery meetings organized	Researcher records
Preparatory visits, information gathering, community mobilization	
Number of visits carried out	Activity report
Number of landmarks identified	Electronic form
Number of information meetings organized	Electronic form
Community decision-making	
Number of community meetings organized	Electronic form/Activity report
Duration of the community meeting	Attendance sheet/Audio record
Attendance at the community meeting	Attendance sheet
Number of households actively participating in the discussion	Attendance sheet/Audio record
Number of locations discussed during the community meeting	Electronic form/Field staff report
Number of locations agreed on during the community meeting	Electronic form/Field staff report
(If cash treatment) Collection of cash contributions	
Number of visits carried out	Activity report
Total cash contributions collected	Electronic form
Number of contributing households	Electronic form
Installation of the pumpbody	
Number of installed deep tubewells	Electronic form/Office records
(If labour treatment) Number of contributing households	Electronic form
Days between agreement and construction of the pumpbody	Electronic form/Office records
Tubewell depth	Electronic form
Installed tubewell water quality testing	
Number of water samples collected	Electronic form
Arsenic contamination	Laboratory test
Iron contamination	Laboratory test
Manganese contamination	Laboratory test
Construction of the platform	
Days between agreement and construction of the platform	Electronic form/Office records
Selection of caretakers	
Days between agreement and caretakers selection	Electronic form/Office records
Number of meeting for caretakers selection	Attendance sheet/Activity report
Attendance at the meeting for caretakers selection	Attendance sheet
Number of caretakers selected	Electronic form
Caretakers training	
Number of caretakers training sessions organized	Office records
Attendance at the caretakers training	Office records
Monitoring visits	
Number of visits carried out	Electronic form/Activity report

Notes All indicators are at community or tubewell level, except the number of public lottery meetings organized and the number of caretakers training sessions organized.

C Randomization checks

We first confirm that the distribution of communities across treatment arms is geographically balanced (Appendix Figure C1). We then test formally for differences across treatment arms.

We first compare treated communities—that received the program under any contribution treatment—to communities assigned to the control group, which did not receive the program. We estimate the following equation:

$$Y_{ij} = \alpha + \beta_t \mathbb{1}_{tj} + \sum_{s=2}^S \delta_s (\mathbb{1}_{sj} - \overline{\mathbb{1}}_s) + \sum_{s=2}^S \phi_s (\mathbb{1}_{tj} \times (\mathbb{1}_{sj} - \overline{\mathbb{1}}_s)) + \epsilon_{ij} \quad (1)$$

where Y_{ij} is a baseline characteristic measured in household i in community j . The indicator variable $\mathbb{1}_{tj}$ takes the value one if community j is assigned to treatment and zero otherwise. We included stratification controls as a function of the indicator variables $\mathbb{1}_{sj}$, which take the value one if community j was assigned treatment status in strata s and zero otherwise. We arbitrarily omit one stratum; demeaning implies that which strata is omitted is irrelevant. These controls are centered and interacted with the treatment dummy (following Gibbons, Serrato and Urbancic, 2019; Imbens and Rubin, 2015; Lin, 2013). When using household-level data, we weight observations so that each community counts equally in summary statistics.

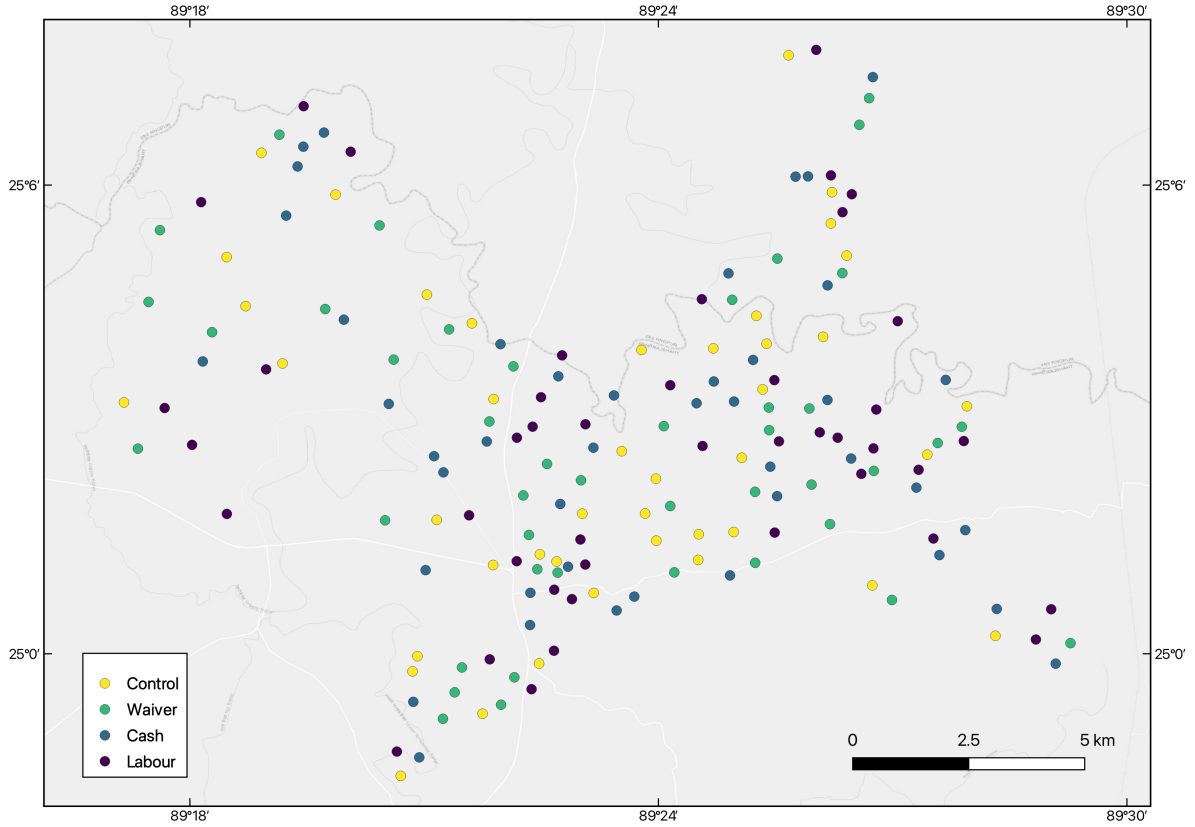
We evaluate the statistical significance of differences between the treated group and the control group using randomization-based inference. Specifically, we simulate the treatment assignment process 500 times and compare the observed differences between treated and control group to the distribution of differences observed under the sharp null of no difference between treated and control communities.

We then estimate versions of Equation 1 pairwise to compare i) each group of communities treated under contribution requirement m from the set $M = \{cash, labour, waiver\}$ to the control group and ii) communities treated under any pair of contribution treatments in M .

Appendix Tables C1 and C2 show randomization checks for household-level characteristics related to drinking water and socio-economic characteristics, respectively. For each table, column 1) reports the control group mean; column 2 reports the treated group mean; and columns 3 to 5 report the mean for each contribution treatment arm. Standard errors, in parentheses, account for clustering by community. The asterisks in columns 2-5 reflect statistical significance of the differences in means between the listed group and the control group for each characteristic separately, obtained by randomization-based inference as we describe above. The right hand column reports the number of observations in the full sample with non-missing data for the listed characteristic. Appendix Table C3 adapts the exercise to characteristics that are measured at the community level, obviating the need for weights and clustered standard errors.

Because Appendix Tables C1 to C3 draw many comparisons, the likelihood that some of the tests return statistically significant differences is relatively high, even if random assignment to treatment occurred as planned. Some of the characteristics are also highly correlated with one another. The starkest difference we report in these checks is that the control group has relatively lower exposure to arsenic at baseline than the treated groups; the treated groups are, relatively, more similar to one another than they are to the control group. This imbalance is of limited concern for our study design as our primary concern is

Figure C1: Map of treatment and control communities



with comparisons across treatment arms.

To shed further light on whether the observed differences across treatment arms are consistent with random assignment to treatment as designed, we report two sets of joint tests for equality on all the characteristics in each table for each pairwise comparison across treatment arms. The first is a joint F -test, in which we regress a dummy for treatment under one contribution requirement m on all the characteristics listed in each table, in a sample consisting of communities assigned to m and either communities assigned to the control group or to a different contribution requirement. We calculate the F statistic on a joint test with the null hypothesis that all the coefficients on the characteristics are equal to zero, and we compare this statistic to its distribution under randomization-based inference. We also calculate the Hotelling's T^2 statistic for each pairwise comparison and again compare this to its distribution under randomization-based inference. We report the p values from these joint tests in Appendix Tables C1 to C3. Across a total of 42 tests—two different joint tests; seven pairwise comparisons; and three groups of characteristics—we reject the null of joint equality only for one test at the 5% level and one test at the 10% level. These joint tests thus provide reassurance that differences across treatment arms are consistent with random assignment to treatment as designed.

Table C1: Balance checks: household water-related characteristics

	Control	Treated	Cash	Labour	Waiver	N
	(1)	(2)	(3)	(4)	(5)	(6)
As in drinking water (WHO)	0.57 (0.02)	0.64* (0.02)	0.65 (0.03)	0.66* (0.03)	0.62 (0.03)	6526
As in drinking water (BGD)	0.19 (0.02)	0.26** (0.02)	0.25 (0.03)	0.24 (0.02)	0.29*** (0.03)	6526
Fecal bacteria in drinking water	0.67 (0.02)	0.62* (0.01)	0.62 (0.02)	0.63 (0.02)	0.62 (0.02)	6503
As in water source (WHO)	0.63 (0.03)	0.71** (0.02)	0.72* (0.03)	0.72* (0.03)	0.69 (0.03)	6515
As in water source (BGD)	0.24 (0.02)	0.32** (0.02)	0.32* (0.03)	0.29 (0.03)	0.35*** (0.03)	6515
Fecal bacteria in water source	0.58 (0.02)	0.56 (0.01)	0.54 (0.02)	0.56 (0.01)	0.56 (0.02)	6505
Household stores drinking water	0.72 (0.02)	0.73 (0.01)	0.73 (0.01)	0.74 (0.02)	0.70 (0.02)	6526
Household treats drinking water	0.08 (0.01)	0.09 (0.01)	0.08 (0.01)	0.12 (0.02)	0.07 (0.01)	6515
Time to collect water (mins)	2.20 (0.08)	2.47 (0.24)	2.21 (0.05)	2.90 (0.57)	2.23 (0.07)	5879
Volume water collected (l)	58.7 (1.9)	58.8 (1.3)	58.7 (2.1)	56.0 (2.0)	60.3 (2.0)	6516
<i>p</i> value: joint F-test, vs control		0.364	0.248	0.364	0.028	
<i>p</i> value: joint F-test, vs cash				0.448	0.812	
<i>p</i> value: joint F-test, vs labour					0.062	
<i>p</i> value: Hotelling, vs control		0.788	0.922	0.406	0.742	
<i>p</i> value: Hotelling, vs cash				0.494	0.676	
<i>p</i> value: Hotelling, vs labour					0.258	

Notes Table reports means and clustered standard errors in parentheses. Weights are applied so that each community counts equally in summary statistics. Regressions include baseline data only, centered strata indicators, and centered strata indicators interacted with treatment. Significance levels obtained from randomization-based inference. Joint F-test from regressing treatment dummy on full set of variables and strata dummies and testing for joint significance. Asterisks in columns 2-5 reflect significance of differences with control group: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C2: Balance checks: household socio-economic characteristics

	Control	Treated	Cash	Labour	Waiver	N
	(1)	(2)	(3)	(4)	(5)	(6)
Household size	3.91 (0.03)	3.92 (0.02)	3.92 (0.04)	3.94 (0.03)	3.89 (0.05)	6529
Predicted income < threshold	0.77 (0.01)	0.75 (0.01)	0.76 (0.01)	0.74 (0.01)	0.75 (0.01)	6490
Muslim	0.95 (0.02)	0.95 (0.01)	0.95 (0.02)	0.93 (0.02)	0.96 (0.02)	6523
In-network size	1.07 (0.05)	1.02 (0.03)	1.07 (0.06)	1.02 (0.04)	0.95 (0.04)	6529
Out-network size	2.76 (0.06)	2.84 (0.04)	2.86 (0.05)	2.81 (0.06)	2.80 (0.05)	6529
Leader	0.099 (0.006)	0.086 (0.004)	0.080 (0.006)	0.096 (0.005)	0.082 (0.007)	6529
Share household uneducated	0.37 (0.01)	0.38 (0.01)	0.38 (0.01)	0.38 (0.01)	0.39 (0.01)	6529
Share household literate	0.54 (0.01)	0.54 (0.01)	0.54 (0.01)	0.54 (0.01)	0.52 (0.01)	6518
Owens livestock	0.74 (0.02)	0.77 (0.01)	0.78* (0.01)	0.76 (0.01)	0.77 (0.02)	6528
Owens land	0.53 (0.02)	0.53 (0.01)	0.52 (0.02)	0.56 (0.02)	0.51 (0.02)	6521
Household has toilet	0.84 (0.02)	0.84 (0.01)	0.82 (0.01)	0.85 (0.01)	0.84 (0.01)	6529
Number of rooms to sleep	1.90 (0.03)	1.86 (0.02)	1.87 (0.03)	1.87 (0.03)	1.83* (0.02)	6525
Metal roof	0.95 (0.01)	0.97* (0.00)	0.96 (0.01)	0.98** (0.01)	0.96 (0.01)	6528
Owens mobile phone	0.56 (0.02)	0.61** (0.01)	0.60 (0.02)	0.63** (0.02)	0.60 (0.02)	6529
Owens motorized vehicle	0.055 (0.005)	0.068 (0.004)	0.070 (0.007)	0.077** (0.007)	0.057 (0.006)	6525
Asset index	-0.04 (0.09)	0.00 (0.05)	0.00 (0.08)	0.06 (0.08)	-0.06 (0.07)	5800
Very high community trust	0.48 (0.01)	0.49 (0.01)	0.48 (0.02)	0.45 (0.02)	0.52 (0.02)	6523
Knows of a community group	0.27 (0.02)	0.30 (0.01)	0.28 (0.01)	0.29 (0.02)	0.30 (0.02)	6452
<i>p</i> value: joint F-test, vs control		0.130	0.480	0.600	0.662	
<i>p</i> value: joint F-test, vs cash				0.866	0.566	
<i>p</i> value: joint F-test, vs labour					0.368	
<i>p</i> value: Hotelling, vs control		0.506	0.826	0.368	0.534	
<i>p</i> value: Hotelling, vs cash				0.640	0.484	
<i>p</i> value: Hotelling, vs labour					0.170	

Notes Table reports means and clustered standard errors in parentheses. Weights are applied so that each community counts equally in summary statistics. Regressions include baseline data only, centered strata indicators, and centered strata indicators interacted with treatment. Significance levels obtained from randomization-based inference. Joint F-test from regressing treatment dummy on full set of variables and strata dummies and testing for joint significance. Note joint tests include additional variables not individually listed. Asterisks in columns 2-5 reflect significance of differences with control group: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C3: Balance checks: community characteristics

	Control	Treated	Cash	Labour	Waiver	N
	(1)	(2)	(3)	(4)	(5)	(6)
Number of households	160.1 (9.6)	160.6 (5.4)	153.3 (8.4)	170.1 (10.0)	161.1 (9.1)	171
Number of clusters	1.56 (0.13)	1.72 (0.09)	1.86 (0.13)	1.71 (0.14)	1.55 (0.17)	171
Number of wells offered	1.33 (0.08)	1.39 (0.04)	1.32 (0.06)	1.48 (0.08)	1.42 (0.08)	171
Maximum predicted well use	0.144 (0.006)	0.149 (0.004)	0.145 (0.006)	0.153 (0.006)	0.149 (0.008)	171
<i>p</i> value: joint F-test, vs control		0.676	0.206	0.596	0.736	
<i>p</i> value: joint F-test, vs cash				0.282	0.374	
<i>p</i> value: joint F-test, vs labour					0.878	
<i>p</i> value: Hotelling, vs control		0.982	0.496	0.492	0.960	
<i>p</i> value: Hotelling, vs cash				0.182	0.438	
<i>p</i> value: Hotelling, vs labour					0.558	

Notes Table reports means and standard errors in parentheses. Regressions include baseline data only, centered strata indicators, and centered strata indicators interacted with treatment. Significance levels obtained from randomization-based inference. Joint F-test from regressing treatment dummy on full set of variables and strata dummies and testing for joint significance. Asterisks in columns 2-5 reflect significance of differences with control group: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

D Simulations to support empirical specifications

Here we use simulations to demonstrate the advantages of the model with centered, interacted strata controls. The theoretical results that underpin this design are described in, e.g., Lin (2013). The simulations illustrate how the centered and interacted model has advantages in generalizability and precision, and, in some cases, delivers the sample average treatment effects when other plausible specifications do not. We also illustrate how the centered and interacted strata control approach is insensitive to which stratum is omitted.

D1 Cross-sectional data, comparison within treatment group only

This simulation mimics our comparison of community take-up rates across treatment arms, where data is by nature cross-sectional and only observed in treated units. We simulate $N = 180$ observations across three strata ($s = 1, 2, 3$) and three treatment arms ($T = 1, 2, 3$). Each stratum contains 60 observations, and within each stratum, 20 observations are assigned to each treatment arm. We simulate an outcome variable y_i where $s_i \in (1, 2, 3)$ denotes the strata number and $T_i \in (1, 2, 3)$ denotes the treatment arm number.

$$y_i = 0.1 s_i + 0.05 (T_i \times s_i) + \varepsilon_i, \quad \varepsilon_i \sim U(-0.15, 0.15).$$

This allows the average outcome to vary across strata and across treatment groups, and allows for heterogeneity of treatment effects across strata.

We then show the results of three alternative groups of specifications. Each specification regresses y_i on indicators for treatment under the three treatment arms (t_{1i}, t_{2i}, t_{3i}). The latter two groups include controls for strata indicators s_{ij} .

$$\begin{aligned} (1) \quad & y_i = \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_3 T_{3i} + \varepsilon_i, \\ (2) \quad & y_i = \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_3 T_{3i} + \gamma_1 s_{1i} + \gamma_2 s_{2i} + \varepsilon_i, \\ (3) \quad & y_i = \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_3 T_{3i} + \sum_{k,j} \delta_{kj} T_{ik} (s_{ij} - \bar{s}_j) + \varepsilon_i. \end{aligned}$$

Specification 1) is a simple comparison across treatment arms. Specification group 2) includes strata fixed effects. Since both the treatment dummies and the strata dummies sum to one, one stratum must be omitted in each case. The equation above includes strata dummies s_{1i} and s_{2i} , but we show results omitting each strata dummy sequentially. Specification group (3) centers (demeans) the strata dummies and interacts them with the treatment indicators. A similar collinearity problem applies to specification group 3, so we also omit one strata dummy's interactions in each version of the specification.

Table D1 shows the results. Column 1 shows the simple comparison of means across treated groups. Columns 2-4 shown results when including strata dummies as controls, sequentially omitting each of the three strata fixed effects. The results are sensitive to which strata dummy is excluded, and the reported coefficients no longer in general capture mean outcomes in each treatment arm. Column 3, which omits strata 2, happens to produce estimates close to the true means, but this is an artifact of the particular parameter values we simulated rather than a general property. Note that the estimated magnitude of the difference between coefficients is however unaffected. Columns 5-7 show results from the centered and

interacted specifications. In contrast to the results in columns 2-4, the reported coefficients now capture the mean outcomes in each treatment arm, and the results are insensitive to which strata dummy's interactions are omitted from the control vector. Compared to column 1, the advantage is that accounting for differences across strata increases precision.

Table D1: Cross-sectional data, all units treated

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
t1	0.30 (0.03)	0.51 (0.02)	0.09 (0.02)	0.31 (0.02)	0.30 (0.01)	0.30 (0.01)	0.30 (0.01)
t2	0.41 (0.03)	0.61 (0.02)	0.20 (0.02)	0.41 (0.02)	0.41 (0.01)	0.41 (0.01)	0.41 (0.01)
t3	0.47 (0.03)	0.68 (0.02)	0.26 (0.02)	0.48 (0.02)	0.47 (0.01)	0.47 (0.01)	0.47 (0.01)
Strata F.E.	No	S1, S2	S2, S3	S1, S3	No	No	No
Centered strata controls	No	No	No	No	S1, S2	S2, S3	S1, S3

Notes All specifications include treatment arm fixed effects and no constant. Columns 2-4 each include two strata dummies as controls, as listed. Columns 5-7 each include two strata dummies, as listed, centered and interacted with the three treatment arm dummies as controls.

D2 Panel data, balanced assignment to treatment across strata

We now expand the simulation described in Appendix Section D1 to panel data with two periods ($t = 0, 1$) and four treatment arms ($T = 1, 2, 3, 4$), where we treat $T = 1$ as the control group. This now mimics our evaluation of impact on arsenic contamination. Here one can think of each observation as capturing a community-level mean outcome, such as the share of households using arsenic-contaminated sources. Later we show the results are identical with data that corresponds to individual-level observations. The simulated sample includes 120 unique units ($N_i = 120$), each observed in both periods, resulting in $N = 240$ observations in total. Each of the three strata contains 40 units, and within each stratum, 10 units are assigned to each treatment arm.

We simulate an outcome variable y_{it} as follows:

$$y_{it} = \theta_i + 0.05 s_i + 0.05 (s_i \times post_t) + 0.05 ((T_i - 1) \times s_i \times post_t) + \varepsilon_{it},$$

$$\theta_i \sim U(0, 0.1), \quad \varepsilon_{it} \sim U(-0.05, 0.05).$$

This approach allows for baseline heterogeneity across units and across strata, heterogeneous trends across strata, and heterogeneous treatment effects across strata.

We compare the following specifications:

$$\begin{aligned}
(1) \quad & y_{it} = \alpha + \theta post_t + \sum_k (\beta_k T_k + \eta_k T_k \times post_t) + \varepsilon_{it}, \\
(2) \quad & y_{it} = \alpha_i + \theta post_t + \sum_k \eta_k T_k \times post_t + \varepsilon_{it}, \\
(3) \quad & y_{it} = \alpha_i + \sum_j \lambda_j (s_{ij} \times post_t) + \sum_k \eta_k T_k \times post_t + \varepsilon_{it}, \\
(4) \quad & y_{it} = \alpha_i + \theta post_t + \sum_k \eta_k T_k \times post_t + \sum_j \delta_j (s_{ij} - \bar{s}_j) \times post_t \\
& \quad + \sum_{k,j} \phi_{kj} T_k \times (s_{ij} - \bar{s}_j) \times post_t + \varepsilon_{it}.
\end{aligned}$$

Specification (1) includes only a constant, treatment dummies, a post indicator, and their interactions. Specification (2) replaces the treatment dummies with unit fixed effects. Specification (3) further replaces the post indicator with a full set of strata $\times$ post interactions. Specification (4) replaces the strata $\times$ post interactions with centered (demeaned) strata controls interacted with the post dummy and the post $\times$ treatment interactions (excluding one strata control arbitrarily).

We also estimate sample average treatment effects directly. We first run a stacked, fully saturated regression that includes a full set of treatment-by-strata and treatment-by-strata-by-post indicators, specifically:

$$y_{it} = \sum_{j=1}^J \pi_{0j} s_{ij} + \sum_{j=1}^J \pi_{1j} (s_{ij} \cdot post_t) + \sum_{k=2}^K \sum_{j=1}^J \pi_{2jk} (T_{ik} \cdot s_{ij}) + \sum_{k=2}^K \sum_{j=1}^J \beta_{kj} (T_{ik} \cdot s_{ij} \cdot post_t) + \varepsilon_{it}$$

The collection $\{\beta_{kj}\}$ provides stratum-specific treatment effects, which we then average using the share of observations in each strata—in this case, these are equal shares—to estimate the sample average treatment effects. We further estimate a variant of this equation that includes unit fixed effects instead of the strata-treatment specific intercepts:

$$y_{it} = \alpha_i + \sum_{j=1}^J \pi_{1j} (s_{ij} \cdot post_t) + \sum_{k=2}^K \sum_{j=1}^J \beta_{kj} (T_{ik} \cdot s_{ij} \cdot post_t) + \varepsilon_{it}$$

We cluster throughout at the primary unit, which in the simulation in this section and the following section is observed twice.

The results are shown in Table D2. For this case, where treatment has equal variance in each strata, all specifications give the same point estimates, and these point estimates correspond to the average weighted effects. The advantage of the interacted model here is that it yields the most precise point estimates, which is expected behavior: the regression adjustment with centered and interacted controls cannot hurt precision and may improve it (Lin, 2013).

As with the simulation described in Appendix Section D1, the centering or demeaning process means that the results with centered strata interactions are insensitive to which of the three strata fixed effects is omitted. Table D3 confirms that the estimated treatment effects are identical regardless of which of the three strata controls we omit from the centered and interacted controls.

Table D2: Panel data with control group, balanced treatment across strata

	Regression results				ATE	
	(1)	(2)	(3)	(4)	(5)	(6)
t2_post	0.088 (0.014)	0.088 (0.014)	0.088 (0.012)	0.088 (0.007)	0.088 (0.007)	0.088 (0.007)
t3_post	0.200 (0.017)	0.200 (0.017)	0.200 (0.012)	0.200 (0.007)	0.200 (0.007)	0.200 (0.007)
t4_post	0.304 (0.023)	0.304 (0.023)	0.304 (0.015)	0.304 (0.007)	0.304 (0.007)	0.304 (0.007)
Unit F.E.	No	Yes	Yes	Yes	No	Yes
Strata-post dummies	No	No	Yes	No	-	-
Centered strata-post controls	No	No	No	Yes	-	-

Notes All specifications include treatment $\times$ post interactions. Column 1) includes a constant, an uninteracted post dummy, and the uninteracted treatment arm fixed effects. Column 2) replaces the constant with unit fixed effects. Column 3) keeps unit fixed effects and replaces the uninteracted post dummy with strata interactions. Column 4) keeps unit fixed effects and replaces the strata interactions with centered and interacted strata controls. Columns 5 and 6 report sample average treatment effects, estimated using a fully interacted model with each strata and then averaging treatment effects across strata. Column 6 includes unit fixed effects in each strata-specific model.

Table D3: Regressions with centered and interacted strata controls, insensitivity to excluded stratum

	(1)	(2)	(3)
t2_post	0.088 (0.007)	0.088 (0.007)	0.088 (0.007)
t3_post	0.200 (0.007)	0.200 (0.007)	0.200 (0.007)
t4_post	0.304 (0.007)	0.304 (0.007)	0.304 (0.007)
Centered strata-post controls	S1, S2	S2, S3	S1, S3

Notes All regressions include treatment $\times$ post interactions, unit fixed effects, and centered and interacted strata controls. Each column omits a different strata dummy from the interacted controls.

D3 Panel data, unbalanced assignment to treatment across strata

We next repeat the steps reported in Appendix Section D2, with the one difference that instead of assigning equal shares of each strata to each treatment arm, we allow these fractions to vary as shown in Table D4. The exact shares are not important, but what is important is that the variance of each of the treatment variables varies across strata. All other steps are identical to those described in Appendix Section D2, including the underlying data generating process.

Table D4: Unbalanced assignment to treatment across strata

Treatment:	Strata:		
	1	2	3
1	31	23	11
2	24	12	8
3	12	12	36
4	13	33	25

Notes Number of units assigned by treatment arm and strata in Appendix Section D3.

We now replicate the same specifications as in Appendix Section D2. The results are shown in Table D5. Unlike the case where treatment arms have the same proportions in each strata, the regression specifications do not yield consistent estimates across specifications. In general, this is because the regression coefficients in the specifications in columns 1-3 represent averages of stratum-specific effects that depend on the variance of each treatment indicator in each stratum, unlike the simple average across strata that defines the sample average treatment effect (see, e.g. Gibbons, Serrato and Urbancic, 2019). Among the four regression specifications, only the specification with the centered and interacted strata controls (column 4) recovers the sample average treatment effects. As in Appendix Section D2, this specification also yields the most precisely estimated coefficients.

Note that although the estimates in columns 4-6 here should capture the same parameter as the corresponding columns in Table D2, the actual point estimates differ slightly given the random variance in the simulated error terms. The true treatment effect parameters in the simulated data generating process are 0.1, 0.2, and 0.3.

Table D5: Panel data with control group, unbalanced treatment across strata

	Regression results				ATE	
	(1)	(2)	(3)	(4)	(5)	(6)
t2_post	0.101 (0.015)	0.101 (0.015)	0.107 (0.011)	0.120 (0.008)	0.120 (0.008)	0.120 (0.008)
t3_post	0.292 (0.018)	0.292 (0.018)	0.202 (0.012)	0.219 (0.009)	0.219 (0.009)	0.219 (0.009)
t4_post	0.365 (0.018)	0.365 (0.018)	0.306 (0.012)	0.316 (0.007)	0.316 (0.007)	0.316 (0.007)
Unit F.E.	No	Yes	Yes	Yes	No	Yes
Strata-post dummies	No	No	Yes	No	-	-
Centered strata-post controls	No	No	No	Yes	-	-

Notes All specifications include treatment $\times$ post interactions. Column 1) includes a constant, an uninteracted post dummy, and the uninteracted treatment arm fixed effects. Column 2) replaces the constant with unit fixed effects. Column 3) keeps unit fixed effects and replaces the uninteracted post dummy with strata interactions. Column 4) keeps unit fixed effects and replaces the strata interactions with centered and interacted strata controls. Columns 5 and 6 report sample average treatment effects, estimated using a fully interacted model with each strata and then averaging treatment effects across strata. Column 6 includes unit fixed effects in each strata-specific model.

D4 “Household-level” data

Lastly we confirm that our conclusions here are not sensitive to switching to household-level data. We expand the simulated dataset from Appendix Section D3 to include 100 individual observations so that each observation is equivalent to a household, and we generate a binary outcome variable y_d with the same average value in each primary unit and period as the outcome variable y in Appendix Section D3. We then replicate the same specifications as in Appendix Sections D2 and D3, clustering at the primary unit. As in Appendix Section D3, among the four regression specifications, only the specification with the centered and interacted strata controls (column 4) recovers the sample average treatment effects. The sample average treatment effects from this simulation are identical to the equivalent simulation with “village-level” data in Appendix Section D3.

Table D6: Panel data with control group, unbalanced treatment across strata, individual data

	Regression results				ATE	
	(1)	(2)	(3)	(4)	(5)	(6)
t2_post	0.101 (0.015)	0.101 (0.015)	0.107 (0.011)	0.120 (0.007)	0.120 (0.007)	0.120 (0.007)
t3_post	0.292 (0.018)	0.292 (0.018)	0.202 (0.012)	0.219 (0.009)	0.219 (0.009)	0.219 (0.009)
t4_post	0.365 (0.018)	0.365 (0.018)	0.306 (0.012)	0.316 (0.007)	0.316 (0.007)	0.316 (0.007)
Unit F.E.	No	Yes	Yes	Yes	No	Yes
Strata-post dummies	No	No	Yes	No	-	-
Centered strata-post controls	No	No	No	Yes	-	-

Notes Observations are “households” in primary units equivalent to the observations in Appendix Sections D2 and D3. Outcomes are binary variables with the same primary unit and period specific averages as in Appendix Section D3. All specifications include treatment $\times$ post interactions. Column 1) includes a constant, an uninteracted post dummy, and the uninteracted treatment arm fixed effects. Column 2) replaces the constant with unit fixed effects. Column 3) keeps unit fixed effects and replaces the uninteracted post dummy with strata interactions. Column 4) keeps unit fixed effects and replaces the strata interactions with centered and interacted strata controls. Columns 5 and 6 report sample average treatment effects, estimated using a fully interacted model with each strata and then averaging treatment effects across strata. Column 6 includes unit fixed effects in each strata-specific model. Standard errors are clustered by primary unit.

E Using machine learning to understand selection

This appendix section lays out our approach to using machine learning to understand selection into, respectively, successful well installation at the community level, and well use and contributions at the household level. Appendix E1 describes how we implement the approach of Chernozhukov et al. (2025) to understand heterogeneity in selection across treatment arms. Appendix E2 describes how we adapt this approach to describe the characteristics that best predict selection on average (i.e., pooling data across the three treatment arms).

E1 Heterogeneous selection across treatment arms

We use machine learning approaches designed to evaluate heterogeneous treatment effects to study selection: into successfully installing wells at the community level and into contributing towards a well and using a well at the household level. We adapt the algorithm described in Chernozhukov et al. (2025). The inputs are data from units of observation $i \in [N] = \{1, \dots, N\}$. Units of observation may be communities or households. The vector Z_i contains variables that are either measured at baseline or predetermined, so cannot plausibly be affected by the intervention. Appendix Table E1 provides a full list of all variables in Z_i .

We use Y_i to denote a selection variable. In our case, this will be either the share of wells for which communities successfully fulfilled all project requirements, an indicator for contributing towards a well in cash or in labour, or the share of household water obtained from a project well.

We implement the algorithm separately for each pair of study arms. We arbitrarily label them reference (R) and comparison (C) and define D_i so that $D_i = 0$ if $i \in R$ and $D_i = 1$ if $i \in C$. Note that in the original formulation in Chernozhukov et al. (2025), R is the control group and C is the treated group. In our case, R is one of the contribution arms and C is another contribution arm. Our objective is to learn about the Conditional Average Difference in Treatment Effects (CADTE), the difference between the conditional average expected outcome in C and in R . We define $r(Z)$ and $c(Z)$ as proxy predictors for the conditional average expected outcome when $D_i = 0$ and when $D_i = 1$, respectively.

We then split the data many times into main (M) and auxiliary (A) samples. We follow Chernozhukov et al. (2025) and fix the number of splits at 100 for all analyses. The steps of the algorithm are as follows:

1. We compute a propensity score $p(Z_i)$ for $i \in [N]$. In our case, because assignment to treatment was random at each lottery, the propensity score is just the mean value of D_i in each group of communities assigned to treatment at the same lottery, which in practice is close to 0.5 in each of the analyses.
2. We consider 100 splits of half the data into M and half the data into A . Specifically, to split the data each time, we sort communities by D_i , then by a random number generated at the lottery level and then by a second random number at the community level. We then assign every other community to M and to A . This approach ensures that M and A are balanced in terms of D_i and approximately balanced across space. For each split s we:

Table E1: Variables included in selection analyses

	Community variable	Household variable
Demand	Mean As exposure index Share with As in drinking water (WHO) Share with As in drinking water (BGD) Share with As in water source (WHO) Share with As in water source (WHO) Mean WTP for new safe source (cash) Mean WTP for new safe source (time) Predicted well use at optimal location	As exposure index As in drinking water (WHO) As in drinking water (BGD) As in water source (WHO) As in water source (WHO) WTP for new safe source (cash) WTP for new safe source (time) -
Income	Predicted share below poverty threshold Share self-reported poor Mean household asset index Share self-reported high income Mean savings Share loaning money Share borrowing money Asset inequality As-weighted share below poverty threshold As-weighted share self-reported poor As-weighted mean household asset index As-weighted share self-reported high income	Predicted income below poverty threshold Self-reported poor Household asset index Self-reported high income Savings Loaned money Borrowed money - - - - -
Social	Share Muslim Share knowing of community association Share with very high community trust Mean social network size Share identified as leaders Mean household size Community size Clusters per 100 households	Muslim Knows of community association Reports very high community trust Social network size Identified as leader Household size - -
Value of time	Mean reservation wage Mean inferred value of time	Reservation wage Inferred value of time
Education	Share of household heads uneducated Share of household heads illiterate	Household head uneducated Household head illiterate

Notes All variables are either measured at baseline or are plausibly pre-determined.

- (a) We use random forests to predict $r(\cdot)$ and $c(\cdot)$ using A , bootstrapping 1000 times and using the default tuning parameters to reduce run time. For each i in M we predict $r(Z_i)$ and $c(Z_i)$.
- (b) We estimate the Best Linear Predictors for the CADTE by weighted OLS in M , i.e.:

$$Y_i = \hat{\alpha}_0 + \hat{\alpha}_1 r(Z_i) + \hat{\alpha}_2 (c(Z_i) - r(Z_i)) + \hat{\beta}_1 (D_i - p(Z_i)) + \hat{\beta}_2 (D_i - p(Z_i)) ((c(Z_i) - r(Z_i)) - \mathbb{E}_{N,M}((c(Z_i) - r(Z_i)))) + \hat{\epsilon}_i, i \in M \quad (2)$$

The weights $w(Z_i)$ are set equal to $\{p(Z_i)(1 - p(Z_i))\}^{-1}$, multiplied by any survey weights included in standard regressions. Standard errors are clustered by community.

- (c) We estimate the Sorted Group Average Differential Treatment Effects (GADTES), averages of the CADTE by quartiles ranked by $c(Z) - r(Z)$, by weighted OLS in M , i.e.:

$$Y_i = \hat{\alpha}_0 + \hat{\alpha}_1 r(Z_i) + \hat{\alpha}_2 (c(Z_i) - r(Z_i)) + \sum_{k=1}^4 \hat{\gamma}_k (D_i - p(Z_i)) \mathbb{1}(c(Z_i) - r(Z_i) \in I_k) + \hat{\nu}_i, i \in M \quad (3)$$

where $I_k = [l_{k-1}, l_k)$ and l_k is the k th quartile of $(c(Z_i) - r(Z_i))_{i \in M}$. Weights and standard errors are as in step b).

- (d) We conduct a classification analysis (CLAN) in which we describe the average characteristics in the most and least affected quartiles. We estimate the CLAN parameters in M by:

$$\hat{\delta}_k = \mathbb{E}_{N,M}(Z_i | c(Z_i) - r(Z_i) \in I_k) \quad (4)$$

where weights and standard errors are estimated as in step b).

3. We take the median point estimates, 95% confidence intervals, and p values for all parameters of interest across the 100 splits. We monotize the GADTES parameters and their confidence intervals following Chernozhukov, Fernández-Val and Galichon (2009).
4. We adjust confidence intervals to account for uncertainty created by the splitting process, using the median 95% confidence interval values to construct 90% confidence intervals and multiplying the p values by 2.

In the concluding three subsections of this appendix, we report or discuss the BLP estimates and GADTES parameters. We also plot the CLAN parameters for all variables for which we reject the null hypothesis that the most and least differentially affected quartiles have the same characteristics at the 10% level.

E2 Characteristics that are correlated with selection on average

We adapt this approach to provide a descriptive account of which characteristics are on average correlated with selection, i.e., which characteristics are correlated with selection when we pool all treatment arms together. Our objective is to learn about the Conditional Average Value (CAV) of the selection variables Y_i . We define $y(Z)$ as a proxy predictor for the conditional average expected value of Y_i .

Again, we split the data 100 times into main (M) and auxiliary (A) samples. The steps of the adapted algorithm are as follows:

1. We consider 100 splits of half the data into M and half the data into A . Specifically, to split the data each time, we sort communities by a random number generated at the lottery level and then by a second random number at the community level. We then assign every other community to M and to A . This approach ensures that M and A are balanced across space. For each split s we:
 - (a) We use random forests to predict $y(\cdot)$ using A , bootstrapping 1000 times and using the default tuning parameters to reduce run time. For each i in M we predict $y(Z_i)$.
 - (b) We estimate the Best Linear Predictors for the CAV by weighted OLS in M , i.e.:

$$Y_i = \hat{\alpha} + \hat{\beta} (y(Z_i) - \mathbb{E}_{N,M}(y(Z_i))) + \hat{\epsilon}_i, i \in M \quad (5)$$

The weights $w(Z_i)$ are survey weights included in standard regressions. Standard errors are robust or clustered by community.

- (c) We estimate the Sorted Group Average Values (GAV), averages of the CAV by quartiles ranked by $y(Z)$, by weighted OLS in M , i.e.:

$$Y_i = \sum_{k=1}^4 \hat{\gamma}_k \mathbb{1}(y(Z_i) \in I_k) + \hat{\nu}_i, i \in M \quad (6)$$

where $I_k = [l_{k-1}, l_k)$ and l_k is the k th quartile of $(y(Z_i))_{i \in M}$. Weights and standard errors are as in step b).

- (d) We conduct a classification analysis (CLAN) in which we describe the average characteristics in the most and least affected quartiles. We estimate the CLAN parameters in M by:

$$\hat{\delta}_k = \mathbb{E}_{N,M}(Z_i | y(Z_i) \in I_k) \quad (7)$$

where weights and standard errors are as in step b).

2. We take the median point estimates, 95% confidence intervals, and p values for all parameters of interest across the 100 splits. We monotone the GAV parameters and their confidence intervals following Chernozhukov, Fernández-Val and Galichon (2009).
3. We adjust confidence intervals to account for uncertainty created by the splitting process, using the median 95% confidence interval values to construct 90% confidence intervals and multiplying the p values by 2.

F Extended data on contributions

This section provides extended data on contributions. Appendix F1 describes ex ante commitments to contribute. Appendix F2 explores differences in the number of contributors across treatment arms. Appendix F3 compares project records of the number of contributors to self-reported data in the endline survey. Appendix F4 explores the distribution of contributions among contributing households. Appendix F5 describes data on which household members took decisions about contributing and contributed cash or labour.

F1 Commitment to contribute

Appendix Table F1 shows the share of households who initially committed either cash or labour, as recorded by field staff during or after the community meeting or retrospectively by households at endline. This share is very similar across the two contribution treatment arms. Many more households later reported that they had committed to contribute than our field staff recorded, but this pattern is also very similar across the two treatment arms.

Table F1: Commitment to contribute at or after community meeting

	Cash	Labour	C = L	N
	(1)	(2)	(3)	(4)
a) Project records				
Household committed to contribute	0.096 (0.012)	0.090 (0.008)	0.740	13 677
b) Self-reported data				
Household committed to contribute	0.356 (0.019)	0.376 (0.016)	0.540	2534

Notes Table reports means and clustered standard errors in parentheses. Data in panel a) record commitments to contribute as reported to field staff at or after the community meeting; sample is all listed households in the community. Data in panel b) record self-reported, retrospective commitments to contribute, recorded at endline; sample is all randomly-sampled households at endline. Regressions control for centered strata indicators interacted with the treatment dummies, arbitrarily omitting one stratum. Weights are applied so that each community counts equally in summary statistics. Significance levels obtained from randomization-based inference. p values from pairwise hypothesis tests in column 3.

F2 Differences in numbers of contributors across treatment arms

In this subsection, we describe differences in the numbers of contributors between cash and labour treatment arms. There are multiple ways to compare these data. Appendix Table F2 reports the number of contributors per offered well (panel a) and the number of contributors per successfully installed well (panel b). Appendix Table F3 repeats this exercise for the number of contributors per community (panel a), per community in which we successfully installed at least one well (panel b), and per community in a subsample of communities that either successfully installed a well under the cash contribution requirement or were assigned to the labour contribution requirement but are matched to the communities that successfully raised cash on baseline characteristics (panel c). Appendix Table F4 reports the

share of households contributing in all communities (panel a) and communities successfully raising contribution requirements (panel b). All three analyses yield similar conclusions: the number of contributors per well installed is greater under cash than labour. This offsets the lower rate at which cash contributions were raised, attenuating the difference between labour and cash in the number of contributors per community or offered well. Accounting for the baseline differences between communities does not close the difference in contributor numbers in communities that successfully installed wells (Appendix Table F3, panel c).

Table F2: Number of contributors per well

	Cash	Labour	C = L	N
	(1)	(2)	(3)	(4)
a) All offered wells				
No. of households contributing	3.5 (1.1)	4.2 (0.3)	0.558	119
b) All wells successfully installed				
No. of households contributing	12.1 (0.8)	4.9 (0.4)	<0.001	58

Notes Table reports means and clustered standard errors in parentheses. Sample includes all offered wells (panel a) and all successfully-installed wells (panel b). Regressions control for centered strata indicators interacted with the treatment dummies, arbitrarily omitting one stratum. Weights are applied so that each community counts equally in summary statistics. Significance levels obtained from randomization-based inference. p values from pairwise hypothesis tests in column 3.

Table F3: Number of contributors per community

	Cash	Labour	C = L	N
	(1)	(2)	(3)	(4)
a) All communities				
No. of households contributing	4.2 (1.3)	8.2 (1.0)	0.008	86
b) All communities successfully installing at least one well				
No. of households contributing	15.0 (1.0)	9.0 (1.0)	0.008	43
c) Successful cash communities and matched labour communities				
No. of households contributing	17.7 (2.5)	12.3 (0.7)	0.164	18

Notes Table reports means and clustered standard errors in parentheses. “Matched labour communities” are those that most closely resemble communities that successfully installed wells under the cash treatment arm, based on a random forest model trained on baseline characteristics. Regressions control for centered strata indicators interacted with the treatment dummies, arbitrarily omitting one stratum. Significance levels obtained from randomization-based inference. p values from pairwise hypothesis tests in column 3.

Table F4: Share of households contributing

	Cash	Labour	C = L	N
	(1)	(2)	(3)	(4)
a) All communities				
Contributed	0.030 (0.009)	0.035 (0.003)	0.752	13 804
b) All communities meeting contribution requirement				
Contributed	0.086 (0.014)	0.038 (0.004)	0.028	7485

Notes Table reports means and clustered standard errors in parentheses. Contributed means that field staff recorded that at least one member of the household contributed in cash or in labour. Sample includes all listed households in indicated communities. Regressions control for centered strata indicators interacted with the treatment dummies, arbitrarily omitting one stratum. Weights are applied so that each community counts equally in summary statistics. Significance levels obtained from randomization-based inference. *p* values from pairwise hypothesis tests in column 3.

F3 Comparing project records to self-reported data

Households also reported whether they contributed to a project well retrospectively during the endline survey. Appendix Table F5 compares these self-reported data to our project records. Our project records are accurate in that most households that we recorded contributing confirm this later on, with only 1% of households being recorded as contributing and not later on confirming this. However, a substantial share of households later report contributing when our project records did not show this to be the case. This share is larger under the labour contribution arm than under the cash contribution arm.

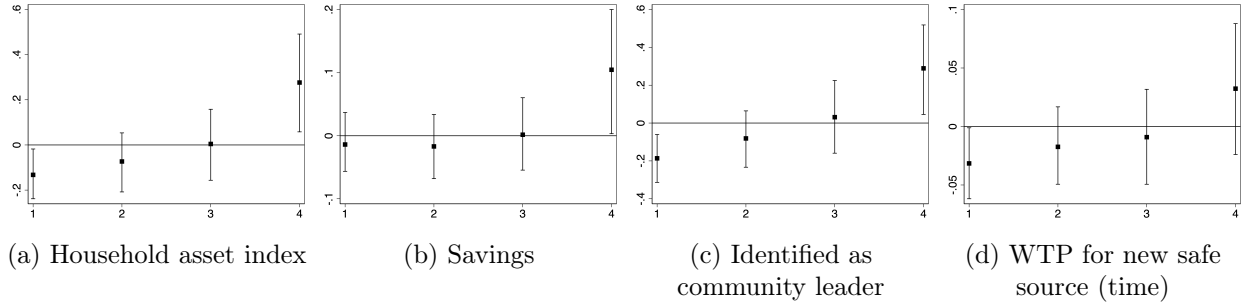
Table F5: Contributing households: project records and survey data

		Project records	
		Contributed	Did not contribute
a) Cash contribution requirement			
Self-reported data	Contributed	0.32	0.10
	Did not contribute	0.01	0.57
b) Labour contribution requirement			
Self-reported data	Contributed	0.14	0.17
	Did not contribute	0.01	0.68

Notes “Project records” use field staff records to identify contributing households. “Self-reported data” uses contributions reported by respondent household during the endline survey. Communities in which at least one well was successful installed only.

We use machine learning to evaluate which households are most likely to later report contributing despite not being recorded as contributing by the project, conducting a classification analysis following the approach in Appendix E2. Those who reported contributing later despite not being originally recorded as contributing are wealthier, as measured by assets and savings, more likely to be identified by other households as a leader, and express higher willingness to contribute time for a new safe source (Appendix Figure F1).

Figure F1: CLAN, difference between self-reported contributions and project records, pooled



Notes Graphs plot the normalized value of the listed characteristic by quartile defined by heterogeneity score. x axis shows quartiles defined by predicted heterogeneity. y axis shows normalized value of listed characteristics. Only characteristics for which the top and bottom quartiles differ significantly ($p < 0.1$) are shown.

When we use survey records of households that contributed, the differential reporting behaviour in labour treatment arm communities tends to offset the differences we observed when we used data based on project records. Appendix Table F6 repeats the analysis in Appendix Table F4 using self-reported data on whether households contributed or not. Note that the sample size and precision changes: we observe contribution status according to project records for all households in each community, whereas we only observe self-reported status in surveyed households. The estimated share of households contributing is higher when we use self-reported data both because of the households that reported contributing despite not being recorded as contributing in our project records and also because the full household lists for each community may include households that have migrated or passed away, whereas we only surveyed households we could locate within the community. When we use self-reported data, the share of households contributing is now considerably higher on average in communities assigned to the labour treatment arm than in communities assigned to the cash treatment arm. In communities that successfully raised contributions, the share of households contributing is still higher for cash than for labour, but the difference is no longer statistically significant.

F4 Contribution amounts

In this section we consider how the total contributions are shared between contributing households. Appendix Figure F2a illustrates the distribution of contributions across households. The greater the number of contributors to each well, and the more equally the contribution is shared between them, the more convex this graph should be. The graph shows that the cash contributions are more evenly shared across a greater number of contributors, with a large share of contributors contributing small shares of the total requirement. More households contributed larger shares under the labour contribution requirement. One way to interpret this difference is that successfully raising cash contributions reveals relatively widespread demand for new wells, as manifested by a large number of contributors. The difference between treatment arms partially reduces, but does not disappear, when we narrow the comparison

Table F6: Share of households contributing (survey data)

	Cash	Labour	C = L	N
	(1)	(2)	(3)	(4)
a) All communities				
	0.061 (0.016)	0.148 (0.013)	<0.001	3474
b) All communities meeting contribution requirement				
	0.203 (0.026)	0.166 (0.012)	0.232	1771

Notes Table reports means and clustered standard errors in parentheses. Contributed means that the household self-reported at endline that at least one member of the household contributed in cash or in labour. Sample contains all randomly-sampled households in indicated communities. Weights are applied so that each community counts equally in summary statistics. Significance levels obtained from randomization-based inference with stratification controls. p values from pairwise hypothesis tests in column 3.

to labour treatment arm communities matched to communities that successfully raised cash contributions. The differences between treatment arms are attenuated, however, if we rely on self-reported data about contributions (Appendix Figure F2b).

Appendix Figure F3 plots the amount households contributed against the value of their assets, both on a logarithmic scale. Panel a) shows the results for our project records; panel b) shows the results for self-reported data. Only households surveyed at endline are shown. All four relationships shown display weakly positive associations between assets and amounts contributed, somewhat stronger for cash than for labour. In simple log-log regressions with robust standard errors, the relationship between assets and amounts contributed is statistically significant for cash when we use project records and for labour when we use self-reported data. The relationship is marginally insignificant ($p = 0.156$) for cash when we use self-reported data. The relationship is weak for labour when we use project records.

F5 Within-household data on contributions

Households reported which household members decided whether to contribute and which provided contributions. We first confirm that contributing households are demographically similar in the cash and labour treatment arms (Appendix Figure F4a). We then compare the shares of household members who had decision-making power over contributions (Appendix Figure F4b) and provided the contributions (Appendix Figure F4c). Men are overwhelming more likely both to take decisions about contributions and to contribute. The vast majority of male household members above the age of 35 reported taking decisions about contributing and contributing themselves. Men below 35 are more likely to contribute labour than cash; above 35, the reverse is true. Few women of any age report taking decisions about contributions or contributing. Very young women (below 25) and the relatively old (above 55) are more likely to provide labour than cash, while for the middle-aged the reverse is true, but mean contribution levels are in any case low for women of all ages.

Figure F2a: Distribution of contributions: project records

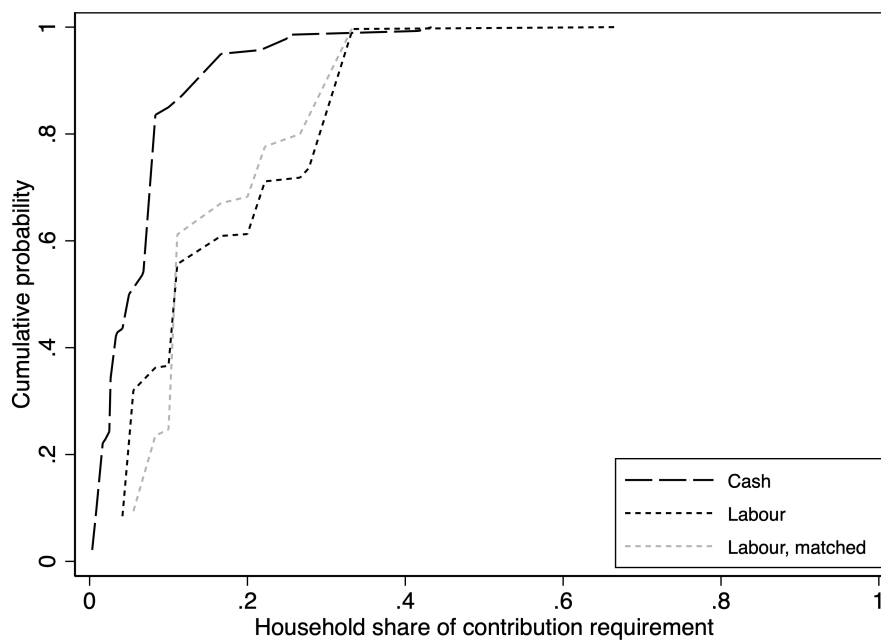
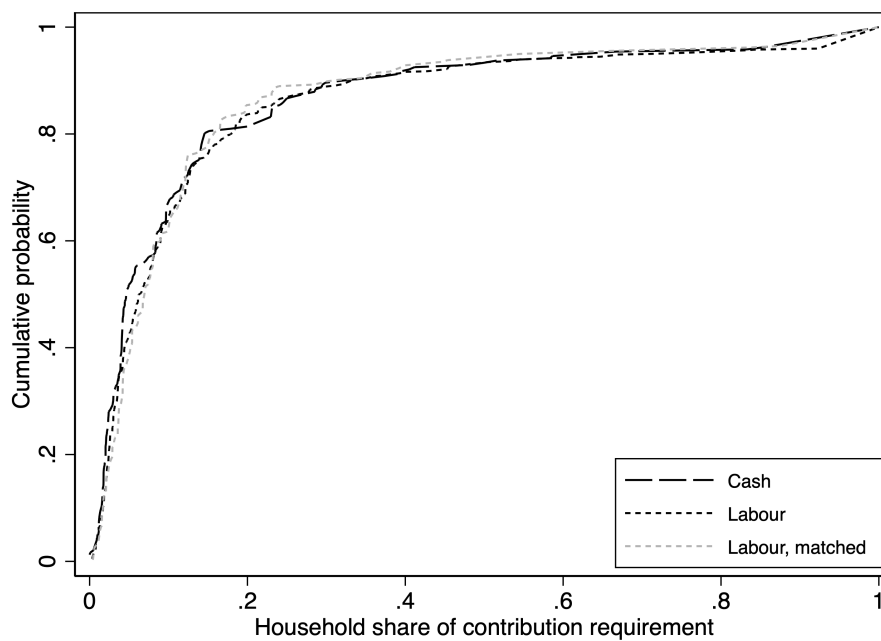
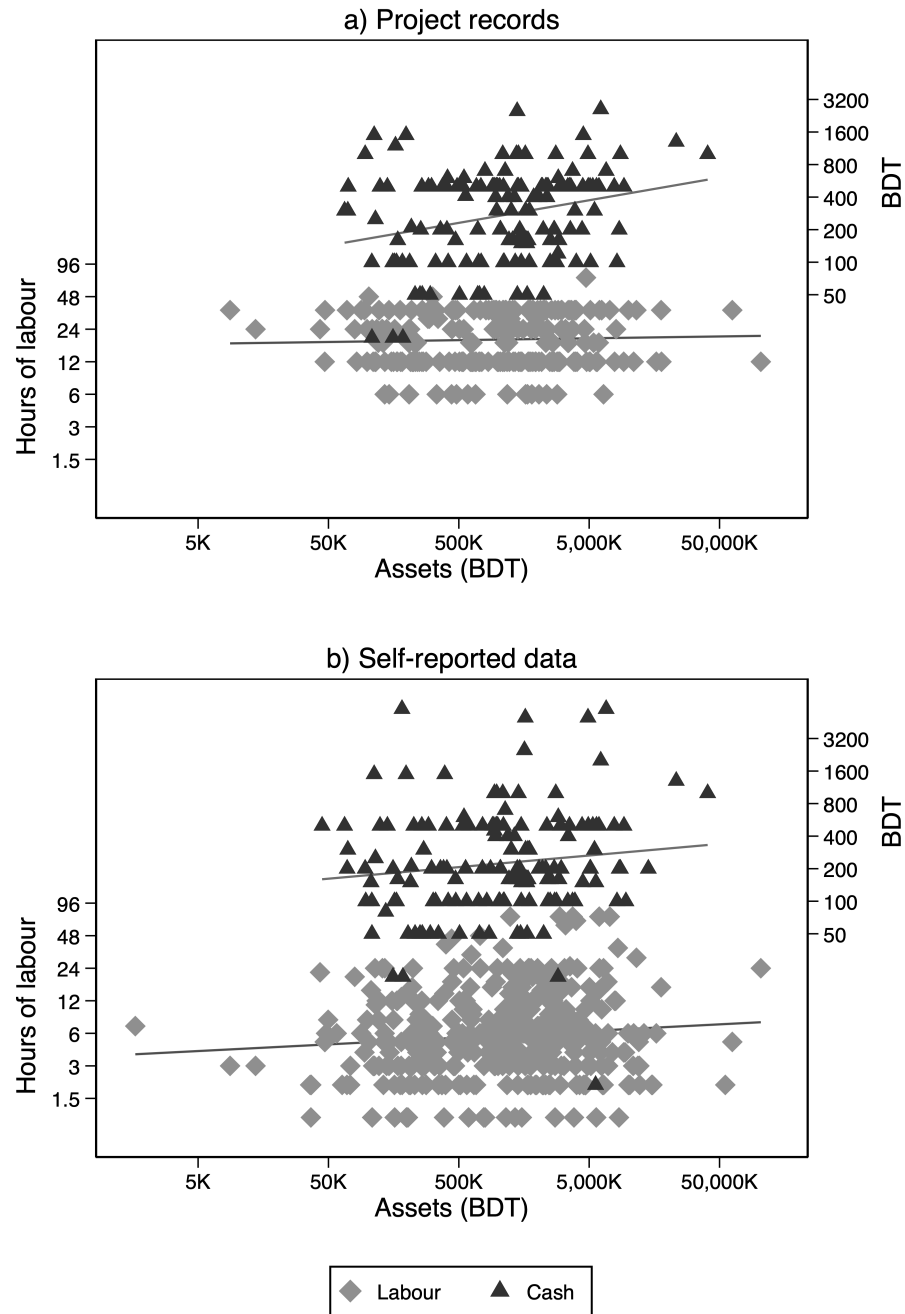


Figure F2b: Distribution of contributions: self-reported data



Notes Graph shows empirical cumulative distribution function (CDF) of shares of total community contribution requirement borne by individual contributing households. Sample includes households contributing to successfully installed wells only.

Figure F3: Relationship between assets and contributions



Notes Scatterplot shows individual contributing households based on project records (panel a) and self-reported survey data (panel b). Assets measured at endline. Both x and y axes have log scales. Best fit lines shown in grey.

Figure F4a: Demographics of contributing households

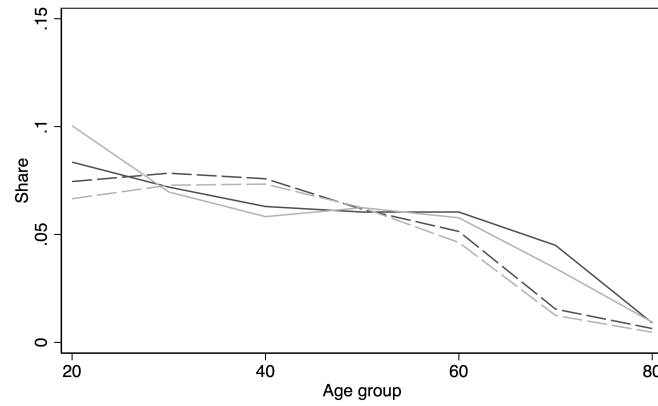


Figure F4b: Within-household distribution of decision-making over contributions

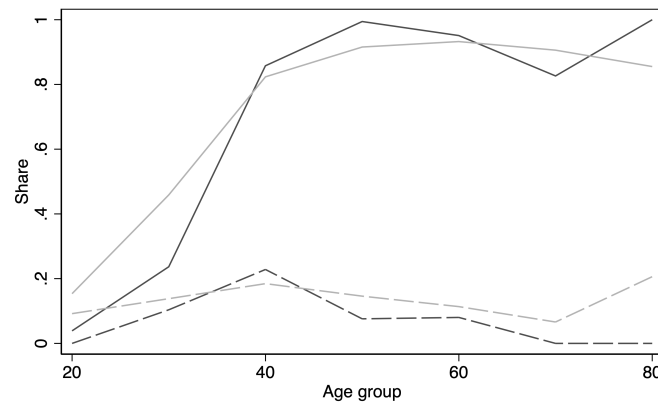


Figure F4c: Within-household distribution of contributions



Notes Graph shows share of household members by age and gender (panel a), and share of household members of listed age and gender who took decisions over household contributions (panel b) and provided contributions (panel c). Sample includes households who provided contributions.

G Follow-up experiment

During a follow-up experiment, we offered an additional well in 34 communities. These wells were offered to incentivize households to truthfully reveal their relative valuations of different decision-making processes, the primary goal of the follow-up experiment (Cocciolo, Habib and Tompsett, 2019). We randomly selected the communities that would receive additional wells, and for each community, randomly extracted a choice between two different decision-making processes, each paired with a community contribution requirement in cash. The three decision-making processes we studied were the consensus-based approach used in this study; unrestricted decision-making by local institutions; and delegation of decision-making to external agents, specifically, our field staff. For each offered choice, one decision-making process was paired with a contribution requirement of 3000 BDT and one with a contribution requirement ranging between 1000 and 5000 BDT. We then implemented majority rule over the randomly-extracted choice. A quirk of the preferences we observe is that in almost all communities, the median voter chooses the option associated with the lowest contribution requirement, implying that in almost all cases (all but one) the community does the same. The result is that these 34 communities are randomly exposed to cash contribution requirements of between 1000 and 3000 BDT.

This allows us to estimate take-up rates for each of those cash contribution requirements. Specifically, we regress whether or not the community successfully raised the contribution for the additional well on the required cash contribution, instrumenting with the randomly assigned minimum requirement in the one case where the community endogenously selected an option associated with a higher contribution rate. We control for the decision-making process paired with the lowest cash contribution to account for differences in how communities took decisions in this follow-up study. The estimated take-up rates are shown in Appendix Figure G1. The wider confidence intervals for intermediate take-up rates reflect the smaller sample sizes in the follow-up experiment than in the main experiment.

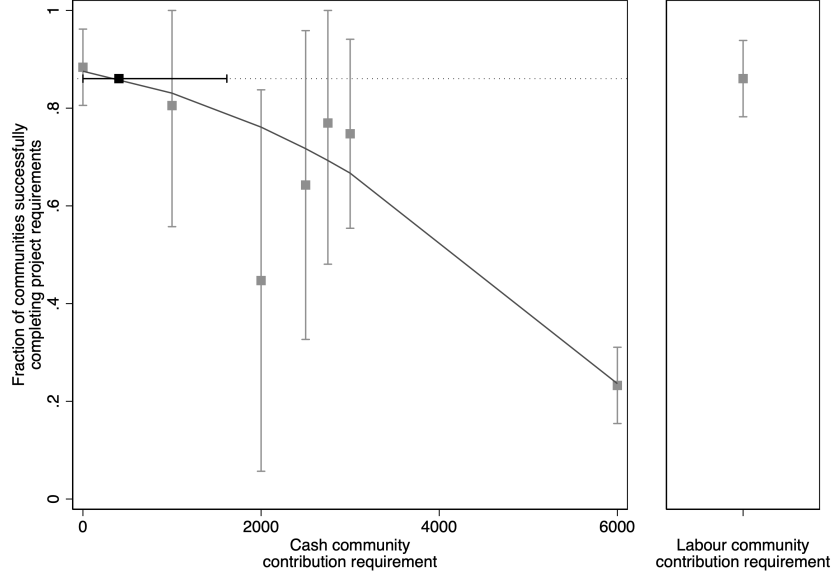
We also estimate a quadratic demand curve, pooling the data from the main and follow-up experiments, after partialling out the effects of the different decision-making processes to which communities in the follow-up experiments were eventually exposed. The estimated demand curve, also shown in Appendix Figure G1, suggests a take-up equal to the labour contribution requirement for a cash contribution requirement of 404 BDT.

To describe the uncertainty in this estimate, we bootstrap the estimation of the demand curve, take-up under the labour contribution requirement, and the level of cash contribution that yields the same take-up as the labour contribution requirement. Specifically, we use a Bayesian bootstrap process in which we generate 500 vectors of Dirichlet weights by drawing random variables from a Gamma distribution and normalizing by their sum, and repeat the estimation process 500 times, weighting in each instance by the Dirichlet weights. If the demand curve lies below the estimated take-up rate under labour for all non-negative contribution requirements, we code the intersection point as zero. We obtain the 90% confidence interval of $[0, 1643]$ from the distribution of estimated intersection points obtained from the Bayesian bootstrap.

Table G1 reports take-up of wells in the follow-up experiment by treatment assignment in the main experiment. Take-up of the additional wells was very high in the labour and waiver treatment arms, moderate in the control group, and low in the cash treatment arm.

Lower take-up in the cash treatment arm could reflect discouragement effects. As we note in the main text, this pattern of results seems inconsistent with communities assigned to cash contributions withholding contributions in anticipation of future preferred offers.

Figure G1: Take-up with contribution requirement



Notes Graph shows take-up at varying levels of cash contribution requirement. Estimated take-up at community contribution requirement equal to zero (waiver treatment arm) and 6000 BDT (cash treatment arm) from this study. Take-up at intermediate prices from follow-up experiment. Take-up under the labour contribution requirement shown in the right-hand panel. Grey bars shown 90% confidence intervals for take-up rates. A quadratic best-fit line is shown in dark grey. The dotted line shows the intersection between predicted take-up and estimated take-up under the labour contribution requirement. Black confidence interval shows the bootstrapped 90% confidence interval for the intersection point.

Table G1: Take-up in follow-up study, by contribution arm

	Control	Waiver	Labour	Cash
Communities offered wells	15	8	3	8
Communities raising contributions	10	8	3	3
Percentage	67%	100%	100%	38%

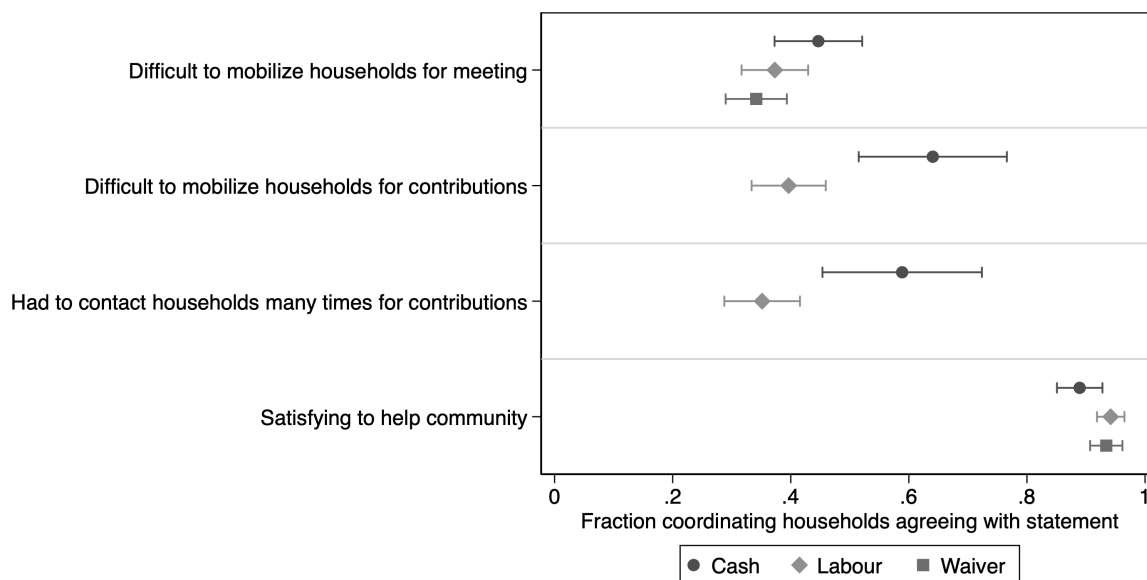
Notes Table reports take-up rates of wells in follow-up project, by contribution requirement treatment. Note that control villages were oversampled by design although respondents were not informed this would be the case during the endline survey.

H Self-reported evaluations and motivations

In this section, we collate information reported by survey participants at endline about aspects of the project operations. We use the full sample at endline, including over-sampled populations such as contributing households. Survey weights account for oversampling so that in the full sample, each community counts equally in summary statistics. Appendix Figures H1 to H6 report simple mean comparisons across treatment arms.

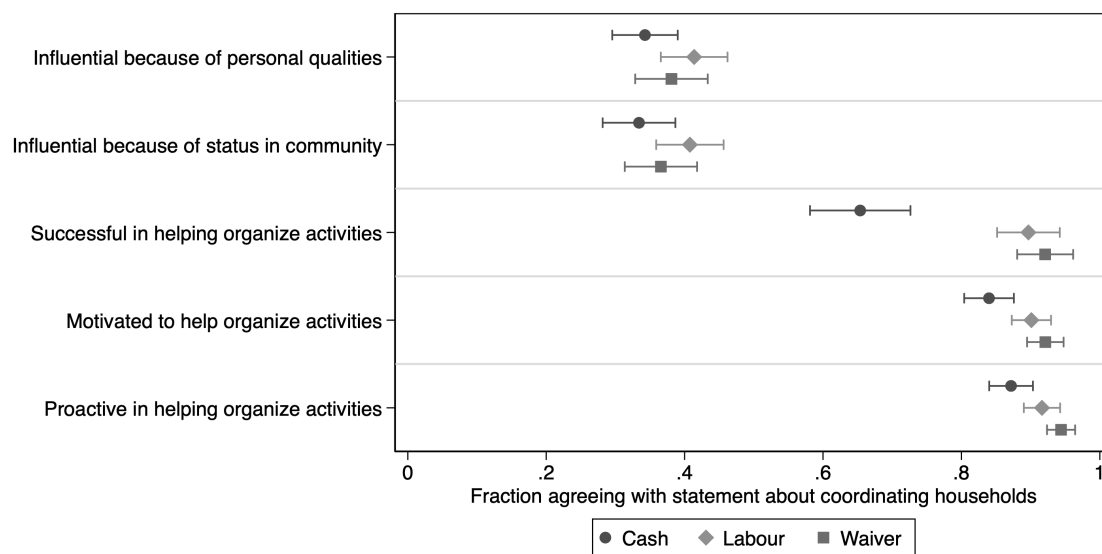
Appendix Figure H1 summarizes what households who self-identified as contact or coordinating households reported about the difficulty of coordinating collective actions. Appendix Figure H2 summarizes what other households reported about the contact or coordinating households in their communities, for the sample of households who knew of at least one household who played this role. Appendix Figure H3 summarizes the reasons given by households who had initially committed to contributed for later not contributing; figure H4 more generally summarizes reasons given for not contributing. For the cash treatment arm, we report these results separately for communities that successfully raised contributions for at least one well and communities that did not. Appendix Figure H5 summarizes the reasons why households who did contribute contributed. Finally, Appendix Figure H6 summarizes the reasons that households who did not use a project well gave for not using the well.

Figure H1: Self-identified coordinating households' evaluations of difficulty of coordination



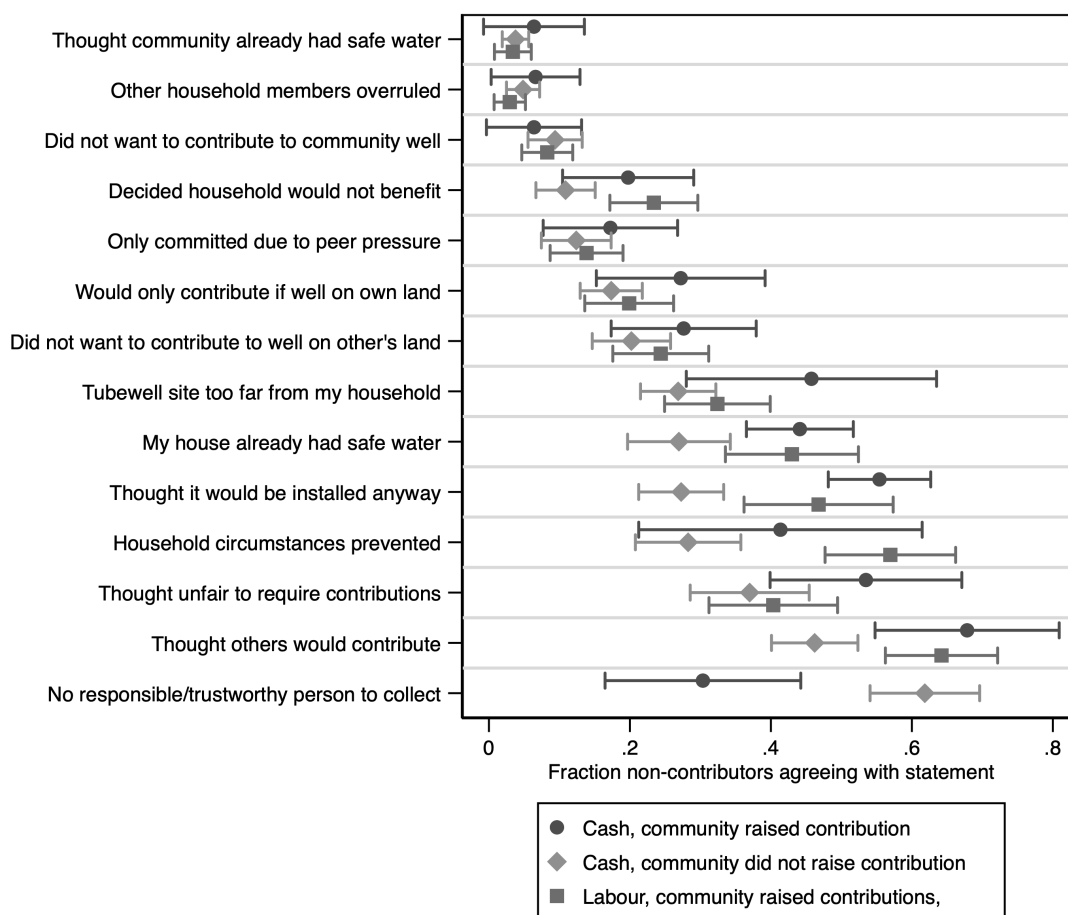
Notes Coefficients from regressions of dummy for agreeing or strongly agreeing with statement on contribution treatment indicators. Sample is all self-identified contact coordinating households. Sampling weights applied. Bars show 95% confidence intervals. Standard errors clustered by community.

Figure H2: Household evaluations of other coordinating households



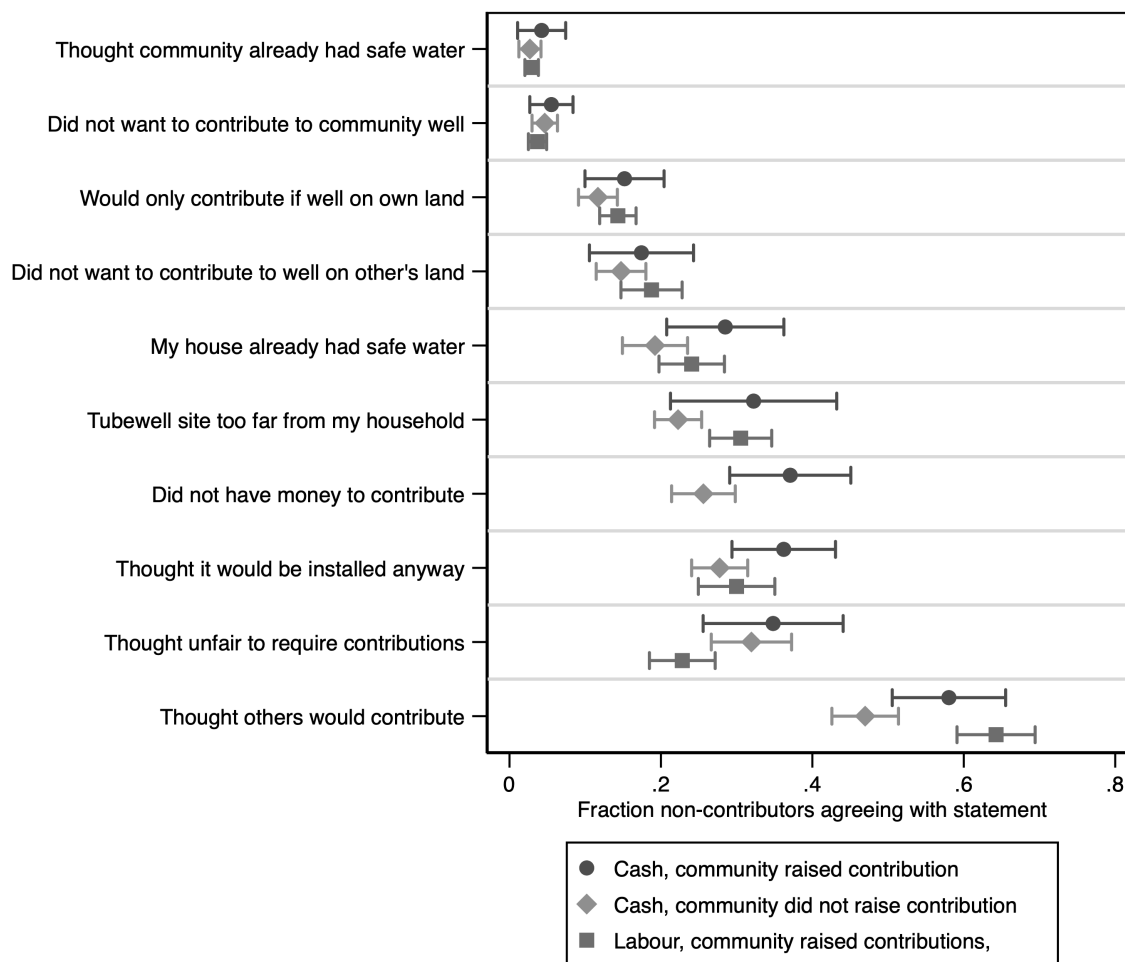
Notes Coefficients from regressions of dummy for agreeing or strongly agreeing with statement on contribution treatment indicators. Sample is all households who knew the identity of at least one coordinating household. Sampling weights applied. Bars show 95% confidence intervals. Standard errors clustered by community.

Figure H3: Reasons given by non-contributing households who had previously committed to contribute at community meetings



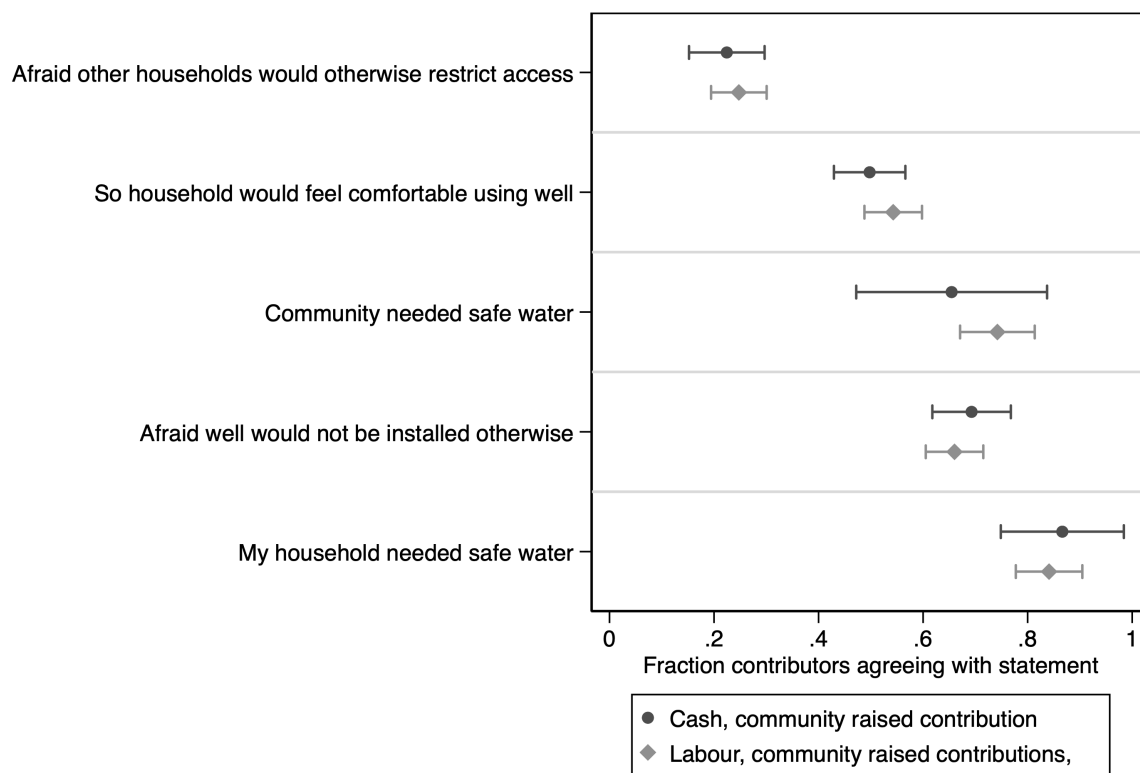
Notes Coefficients from regressions of dummy for agreeing or strongly agreeing with statement on contribution treatment indicators interacted with success raising contributions. Sample is all households who committed to contribute at the meeting but later did not contribute in communities that agreed on a site. Sampling weights applied. Bars show 95% confidence intervals. Standard errors clustered by community.

Figure H4: Reasons given by non-contributing households



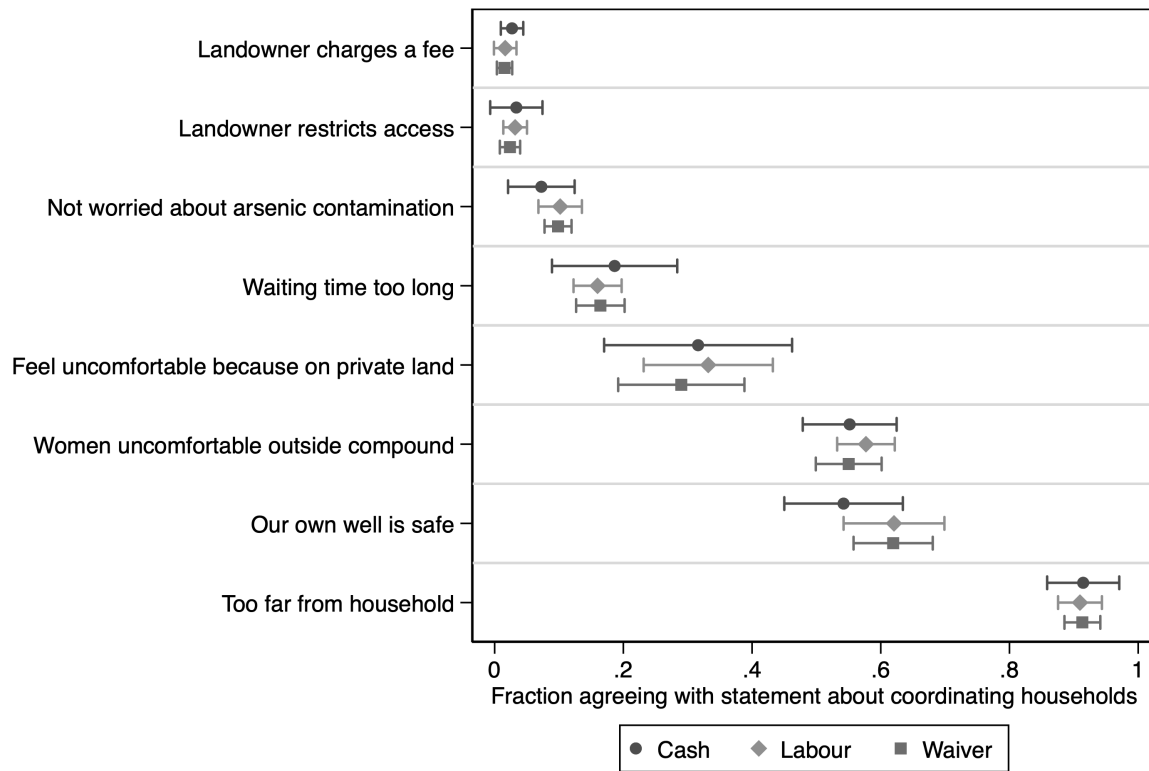
Notes Coefficients from regressions of dummy for agreeing or strongly agreeing with statement on contribution treatment indicators interacted with success raising contributions. Sample is all households who did not contribute in communities that agreed on a site. Sampling weights applied. Bars show 95% confidence intervals. Standard errors clustered by community.

Figure H5: Reasons given by contributing households



Notes Coefficients from regressions of dummy for agreeing or strongly agreeing with statement on contribution treatment indicators. Sample is all households who contributed in communities that successfully raised contributions. Sampling weights applied. Bars show 95% confidence intervals. Standard errors clustered by community.

Figure H6: Reasons given by non-user households



Notes Coefficients from regressions of dummy for agreeing or strongly agreeing with statement on contribution treatment indicators. Sample is all households who reported not using their nearest constructed project well. Sampling weights applied. Bars show 95% confidence intervals. Standard errors clustered by community.

I Robustness of effects on program impact

I1 Declines in arsenic exposure in the control group

Our estimated program impacts would be biased if there were spatial spillovers, in particular, if households used project wells installed in different communities. We observe a decline in arsenic exposure in the control group which could be interpreted as indicative of spatial spillovers. To shed light on how this decline emerges, we examine trends in measured arsenic contamination in a large sample of wells which we tested at both baseline and endline. These wells show, on average, declines in arsenic, especially at low levels, that line up almost exactly with the declines in exposure to arsenic in control households. These secular trends are very similar to the control group in all three treatment arms (Appendix Table I1).

Table I1: Changes in arsenic contamination in wells used at both baseline and endline

	Control Mean	Estimated impact			C = W	L = W	C = L	N
	(1)	Cash (2)	Labour (3)	Waiver (4)	(5)	(6)	(7)	(8)
Arsenic contamination (WHO)	−0.06 (0.02)	−0.01 (0.03)	−0.04 (0.03)	−0.01 (0.02)	0.856	0.308	0.560	12 056
Arsenic contamination (BGD)	−0.01 (0.02)	0.02 (0.02)	0.00 (0.02)	−0.01 (0.02)	0.590	0.884	0.530	12 056

Notes Table estimates differences with respect to control group. The sample is all wells for which we have both baseline and endline data on arsenic contamination, comprising wells used by at least one household in our sample at both baseline and endline. Clustered standard errors are in parentheses. Regressions include baseline and endline data, well fixed effects, a post-treatment dummy, and centered strata indicators interacted with the post dummy and treatment arm-post dummy interactions, arbitrarily omitting one stratum. Significance levels obtained from randomization-based inference. p values from pairwise hypothesis tests in columns 5-7.

I2 Specification tests

Appendix Table I2 evaluates robustness of the effects on program impact across other specification choices. Panel a) reports the main estimates, from analysis that includes all randomly-sampled households, community fixed effects, and stratification controls—comprising centered strata indicators interacted with the post dummy and treatment arm-post dummy interactions, arbitrarily omitting one stratum—using randomization-based inference. Estimates shown in panel b) use analytical standard errors for inference, which is considerably less conservative than randomization-based inference, yielding smaller p values. Estimates in panel c) restrict to households for whom we have both baseline and endline data, and estimates in panel d) include household fixed effects instead of community fixed effects in the same sample as panel c). Both of these yield very similar results to the main estimates. Estimates in panel e) use treatment arm fixed effects, interacted with centered strata indicators, instead of community fixed effects. This results in very similar point estimates but sacrifices efficiency because it does not absorb baseline difference across communities that are orthogonal to treatment status. Estimates in panel f) use the full sample, weighted to account for oversampling of populations of special interest, resulting in very similar, fractionally more precise estimates. Panel g) reports results from randomly-sampled households with commu-

nity fixed effects, but omits the strata controls, giving similar but less precisely estimated comparisons across treatment arms. Panel h) reports results from randomly-sampled households with community fixed effects, replacing the centered stratification controls with strata-post interactions, again giving similar results. Estimates in panel i) are from a subsample weighted by tubewell feasibility, excluding communities in which installation attempts failed and weighting communities in which we did not attempt installation by their probability of installation success. The comparison across treatment arms remains very similar.

Table I2: Impact on arsenic contamination: Robustness tests

	Control Mean	Estimated impact			C = W	L = W	C = L	N
	(1)	Cash (2)	Labour (3)	Waiver (4)	(5)	(6)	(7)	(8)
a) Main estimates								
Arsenic exposure index	-0.10 (0.02)	0.01 (0.03)	-0.10** (0.04)	-0.11** (0.04)	0.014	0.992	0.026	13 646
b) Analytical standard errors								
Arsenic exposure index	-0.10 (0.02)	0.01 (0.03)	-0.10** (0.04)	-0.11*** (0.04)	0.003	0.971	0.005	13 646
c) Panel sample								
Arsenic exposure index	-0.10 (0.02)	0.01 (0.03)	-0.11** (0.04)	-0.10* (0.04)	0.036	0.750	0.020	13 014
d) Household fixed effects								
Arsenic exposure index	-0.10 (0.02)	0.01 (0.03)	-0.11** (0.04)	-0.10* (0.04)	0.036	0.750	0.020	13 014
e) Treatment arm fixed effects								
Arsenic exposure index	-0.10 (0.02)	0.01 (0.03)	-0.10 (0.04)	-0.12 (0.04)	0.172	0.996	0.220	13 646
f) All households								
Arsenic exposure index	-0.10 (0.02)	0.02 (0.03)	-0.10** (0.04)	-0.11** (0.04)	0.014	0.992	0.010	14 497
g) Omitting strata controls								
Arsenic exposure index	-0.10 (0.03)	0.01 (0.04)	-0.08 (0.05)	-0.10** (0.05)	0.016	0.788	0.048	13 646
h) Strata-post dummies								
Arsenic exposure index	- (.)	0.01 (0.03)	-0.10** (0.04)	-0.10** (0.04)	0.014	0.902	0.026	13 646
i) Weighted by tubewell feasibility								
Arsenic exposure index	-0.09 (0.03)	0.00 (0.03)	-0.13*** (0.04)	-0.10* (0.04)	0.050	0.466	0.008	12 568

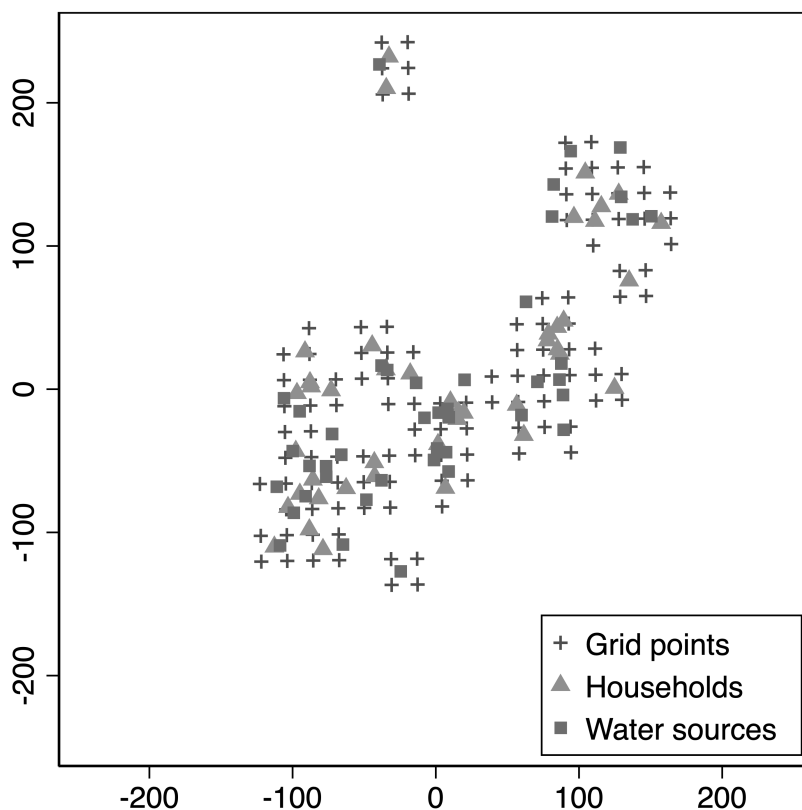
Notes Table reports program impact relative to control group. Clustered standard errors are in parentheses. Regressions include baseline and endline data. *p* values from pairwise hypothesis tests in columns 5-7. Asterisks reflect significance of impacts with respect to the control group: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

J Measuring location quality

We measure location quality by evaluating how the selected location compares to other potentially feasible locations within the community. We first predict well use across a grid of potentially feasible locations within each community. We then construct measures of location quality by comparing predicted well use at a selected location to the distribution of predicted use across all potentially feasible locations within the community.

The grid of potentially feasible locations is at 20m resolution. We define a potentially feasible location as a point on the grid that lies within 20m of a household, a community water source, or a project well. The logic is that communities are typically densely populated quasi-islands situated on relatively high ground and surrounded by relatively low-lying ground that frequently floods. Land that is suitable for housing or existing water sources is also potentially suitable for a new, safe well. Not all grid locations will, in practice, be feasible. Some of the grid locations may lack drainage, be too close to potential groundwater contaminants such as pit latrines, or lack sufficient space or overhead clearance for well installation. However, all feasible locations should be close to a point on the grid of potentially feasible locations. Appendix Figure J1 illustrates.

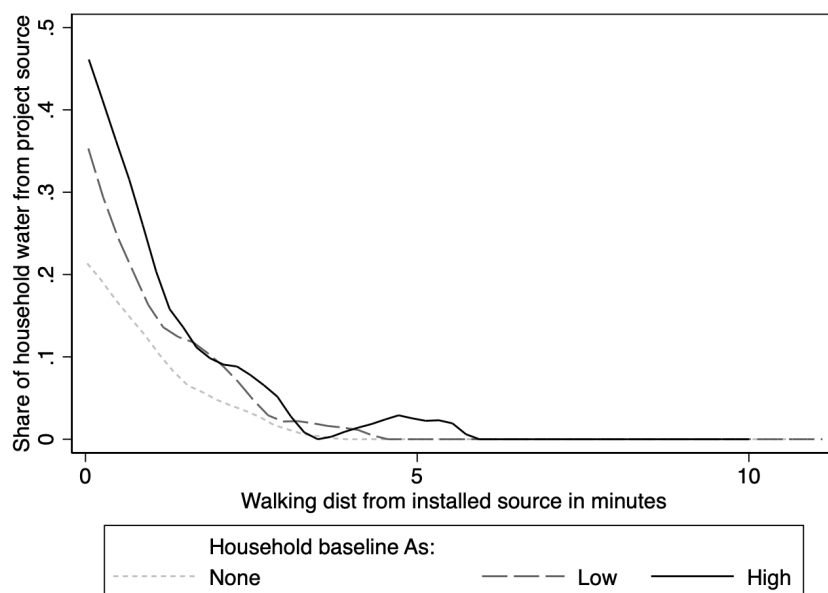
Figure J1: Example grid



Notes Example of grid of potential locations. Coordinate system in metres, centered at village centroid.

For a given well location or pair of locations, we can predict well use in the baseline household sample using a simple model in which well use depends only on baseline arsenic contamination and distance between the household and the well location. Household-level well use is highly sensitive to distance, and households exposed to higher levels of arsenic at baseline are more likely to use the well at any walking distance (Appendix Figure J2). Specifically, to predict well use for a given well location or pair of locations, we assign every household in the baseline sample into one of 33 bins, defined by interacting 11 bins of distance (<40m; 40-80m;...;>400m) between the household and the well location, or nearest well location if there are two wells, with 3 bins of baseline arsenic contamination in household drinking water (<10ppb; 10-50ppb ; >50ppb). We assign households in each bin the average well use rate or share of water collected from a project well from households with these characteristics in the sample of communities that installed at least one program well. For communities offered one well, we predict well use at every potentially feasible location on the grid. For communities offered two wells, we predict well use for every pair of grid points. The resulting predictions estimate the average amount of water a household would obtain from project wells installed in a given a location or pair of locations, if use depended only on distance to the well and baseline arsenic contamination. Appendix Figure J3 shows the distribution of predicted well use for each community, sorted by maximum predicted use.

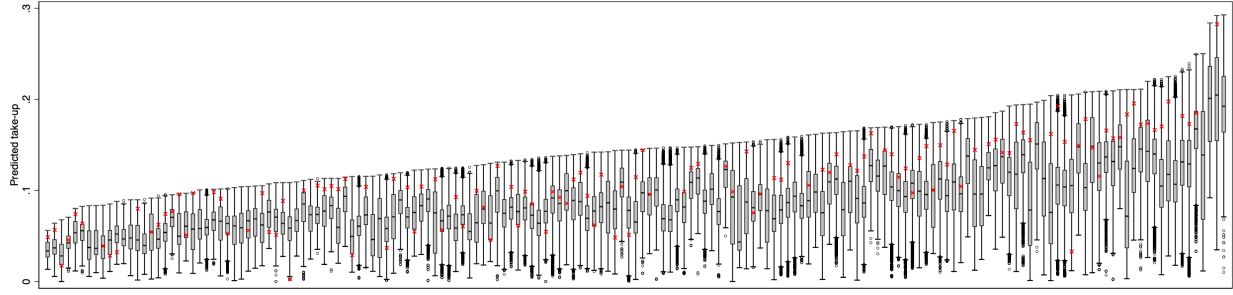
Figure J2: Household-level take-up, baseline arsenic contamination, and distance from well



Notes Reproduced with permission from Cocciolo et al. (2021). Graph shows mean share of drinking and cooking water obtained from project well by distance to well and baseline arsenic contamination in household drinking water. Data from communities in which we successfully installed wells.

Maximum predicted well use is also a useful measure of potential impact. Maximum predicted well use measures the use rates we would observe were a well (or wells) constructed in the location (or pair of locations) with highest predicted well use, holding use rates constant

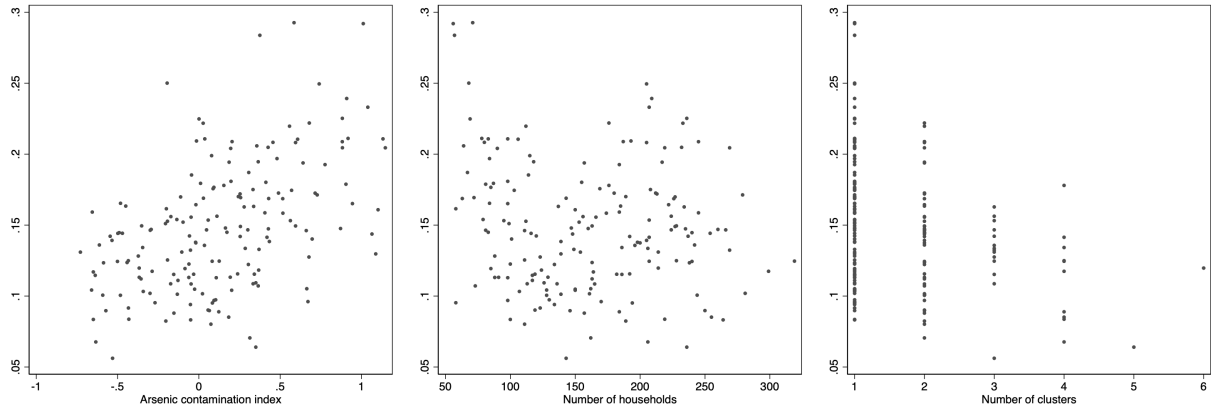
Figure J3: Predicted well use by location



Notes Box plot shows distribution of predicted well use across potentially feasible locations or pairs of locations. Communities (N=171) sorted by maximum predicted well use. Red crosses show selected locations in treated communities that chose locations (N=127).

as a function of distance between households and wells and baseline arsenic contamination. Maximum predicted well use is highest in compact, densely populated communities, with high arsenic contamination that is spatially clustered, and where we offered more wells per household. Appendix Figure J4 shows the correlations between maximum predicted well use and baseline arsenic contamination, community size, and compactness (proxied by the number of clusters, where a larger number of clusters indicates a less compact community). Note that the non-monotonic relationship between community size and maximum predicted well use arises because of differences in whether we offered one or two safe wells.

Figure J4: Correlates with maximum predicted well use



Notes Each point is a community. N = 171.

To compare location quality across treatment arms, we construct three measures of location quality: predicted well use at the selected location, normalized by the maximum predicted well use across potential locations within the community; raw predicted use at the selected location; and a rank measure that takes the value one if communities choose the location that maximizes predicted use and zero if communities choose the location that min-

imizes predicted use (Appendix Table J1). For all three measures, communities choose the best locations under the labour contribution requirement. However, the differences between treatment arms are only statistically significant for the comparison between cash and labour using the rank measure. One possible way in which differences in decision quality might emerge is via improved deliberation. Under the labour contribution requirement, field staff recorded the highest number of active meeting participants and active female participants, communities discussed the largest number of candidate locations without reducing the quality of proposed locations, and communities chose the largest share of sites on public land (Appendix Tables J2 and J3). However, as with the differences in location quality, most of these differences are relatively small and not statistically significant.

Our measure of location quality is naturally imperfect. We use an “as the crow flies” distance between households and potential well sites to predict use. Although distance is a powerful predictor of use, our measure does not capture dimensions of distance that are a function of specific local geography. As an extreme example, it may be impossible to walk between two nearby locations because of surface water between them. However, we show in Appendix K that measures we obtain from this exercise strongly predict well use.

Table J1: Predicted well use at selected locations

	Cash	Labour	Waiver	C = W	L = W	C = L	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Normalized	0.70 (0.03)	0.76 (0.02)	0.72 (0.03)	0.534	0.310	0.218	127
Raw	0.104 (0.007)	0.117 (0.006)	0.109 (0.007)	0.658	0.316	0.262	127
Rank	0.66 (0.04)	0.78 (0.03)	0.70 (0.05)	0.508	0.140	0.076	127

Notes Table reports means and clustered standard errors in parentheses. The sample is all communities that held a meeting and selected a location. Regressions control for centered strata indicators interacted with the treatment dummies, arbitrarily omitting one stratum. Significance levels obtained from randomization-based inference. p values from pairwise hypothesis tests in columns 4-6. Normalized: predicted use at the selected location(s) normalized by the maximum predicted use over a grid of potential locations. Raw: predicted use at the selected location(s). Rank: takes the value 1 if selected location chosen has the highest predicted use and 0 if selected location has lowest predicted use.

Table J2: Deliberation during community meetings

	Cash	Labour	Waiver	C = W	L = W	C = L	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
No. of participants active	4.9 (0.4)	6.5 (0.7)	5.0 (0.5)	0.872	0.082	0.030	129
No. of female participants	0.29 (0.10)	0.40 (0.19)	0.36 (0.11)	0.670	0.854	0.668	129
No. of sites discussed/well	1.42 (0.10)	1.65 (0.11)	1.50 (0.11)	0.644	0.378	0.134	129
Fraction agreed sites private	0.65 (0.07)	0.57 (0.06)	0.64 (0.07)	0.838	0.576	0.446	127
Meeting duration (minutes)	84.7 (3.3)	81.2 (3.4)	85.6 (4.4)	0.820	0.330	0.542	128

Notes Table reports means and robust standard errors in parentheses. The sample is all treated communities with non-missing data. Regressions control for centered strata indicators interacted with the treatment dummies, arbitrarily omitting one stratum. Significance levels obtained from randomization-based inference. p values from pairwise hypothesis tests in columns 4-6.

Table J3: Predicted well use at sites proposed at meetings

	Cash	Labour	Waiver	C = W	L = W	C = L	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Normalized	0.70 (0.03)	0.72 (0.02)	0.73 (0.02)	0.480	0.850	0.568	272
Raw	0.104 (0.006)	0.111 (0.006)	0.108 (0.006)	0.744	0.594	0.492	272
Rank	0.66 (0.03)	0.73 (0.03)	0.71 (0.03)	0.366	0.524	0.306	272

Notes Table reports means and clustered standard errors in parentheses. Sample is all proposals for locations discussed at community meetings. For communities assigned two wells, each proposal is a pairwise combination between locations discussed. Weights are applied so that each community counts equally in summary statistics. Regressions control for centered strata indicators interacted with the treatment dummies, arbitrarily omitting one stratum. Significance levels obtained from randomization-based inference. p values from pairwise hypothesis tests in columns 4-6.

K Decomposition of well use

Well use, measured as the average share of a household’s water drawn from a project source, can be mechanically multiplicatively decomposed into five terms. The decomposition is as follows:

$$\frac{\text{use}_j}{\overline{\max_{l_j \in L_j} \widehat{\text{use}}_{l_j}} \Big|_{j \in J}} = \underbrace{\frac{\max_{l_j \in L_j} \widehat{\text{use}}_{l_j}}{\overline{\max_{l_j \in L_j} \widehat{\text{use}}_{l_j}} \Big|_{j \in J}}}_{\text{geography}} \times \underbrace{\frac{\widehat{\text{use}}_{l_j^*}}{\max_{l_j \in L_j} \widehat{\text{use}}_{l_j}}}_{\text{location choice}} \times \underbrace{\frac{\text{attempted}_j}{\text{offered}_j}}_{\text{community take-up}} \times \underbrace{\frac{\text{installed}_j}{\text{attempted}_j}}_{\text{geology}} \times \underbrace{\frac{\text{use}_j}{\widehat{\text{use}}_{l_j^*} \times \frac{\text{installed}_j}{\text{offered}_j}}}_{\text{household take-up}}$$

where use_j is the mean share of household water from a project well in community $j \in J$, the set of communities; $\widehat{\text{use}}_{l_j}$ is predicted well use at location $l_j \in L_j$, the set of potentially feasible locations in j ; offered_j , attempted_j and installed_j are the number of wells offered, attempted, and installed in j , respectively; and l_j^* is the chosen location in j from among the set L_j . We rescale use_j by the mean maximum predicted use across communities $\left(\overline{\max_{l_j \in L_j} \widehat{\text{use}}_{l_j}} \Big|_{j \in J} \right)$, so that we can interpret well use relative to what it is feasible, on average on across communities, to achieve.

The first term, *geography*, is the component of impact that is predetermined by the distribution of households and arsenic contamination across space. We quantify the influence of geography using the maximum predicted well use across a grid of potentially feasible locations within each community, normalized by the average of this value across communities. As described in Appendix J, we predict use of wells using baseline arsenic contamination and the distance between households and each potential location. All else equal, use of project wells is higher in communities with higher baseline arsenic contamination, because households with higher arsenic contamination are more likely to adopt a new arsenic-safe well at any distance (Cocciolo et al., 2021). Use is also sensitive to distance. Potential impact is thus higher when arsenic-affected households live close to one another, allowing many of them to be served by a single well, than when arsenic-affected households are geographically scattered. Our measure of geography captures the share of a community’s drinking water that would be collected from a project well or wells if well locations were chosen to maximize predicted well use.

The second term, *location choice* measures how the selected location compares to the other potentially feasible locations within the community. Our primary measure is predicted well use at the selected location, normalized by the maximum predicted well use across potential locations within the community. This measure takes the value one if the community chooses the location that maximizes predicted use.

The third term, *community take-up* captures the share of offered wells for which the community met all the project requirements, resulting in an attempt to install a well by the contractor.

The fourth term, *geology* determines whether or not attempted installations were successful. On average, as we discuss in section I.B of the main paper, only 89% of attempted installations were successful. The attempts that failed did so because of local geological

conditions—an impenetrable rock layer or an unstable sandy layer—that are only revealed when installation is attempted. We measure the influence of geology by the share of attempted installations that were successful.

The fifth term is a residual measure of *household take-up*. This residual measure captures how use of successfully installed wells varies after accounting for both geography and location choice, i.e., the extent to which use is higher or lower than predicted given the local geography of population density and baseline arsenic contamination. When this residual measure is high, it means that households report using the well at higher rates than those we would expect given their distance to the well location and baseline arsenic contamination. When it is low, the reverse is true. This residual measure could be low because landowners restrict access to the site or otherwise dissuade households from using the well, because local households have more pressing priorities than exposure to arsenic, because a site is less physically accessible than predicted by our model of location quality, or for other reasons.

Some terms are undefined or unmeasurable in some cases, for example, when no installations are attempted we do not observe the share that would have been successful (the *geology* term). However, in all these cases, the value of the term is irrelevant to the multiplicative decomposition because another term is always equal to zero, and well use is zero. In the example, whether or not an attempted installation would have been successful is irrelevant if no attempt is made to install a well.

Appendix Table K1 shows how the components correlate with one another. Geography, or potential impact, is positively correlated with both location choice—communities choose relatively better locations when there is more at stake—and community take-up—communities are more likely to select into well installation when potential impact is greater. Geology, or the probability of successfully installing a well if installation is attempted, is uncorrelated with other terms in the decomposition, as anticipated. However, the correlations are relatively weak.

Table K1: Correlation matrix

	(1)	(2)	(3)	(4)	(5)
Geography (1)	1.00				
Location choice (2)	0.17*	1.00			
Community take-up (3)	0.15*	0.01	1.00		
Geology (4)	-0.07	-0.01	-0.03	1.00	
Household take-up residual (5)	0.11	-0.14	0.01	0.03	1.00

Notes Pairwise correlation between components. Undefined terms assigned their mean value.

* $p < 0.10$

Appendix Table K2 shows how each individual component separately predicts well use. All five terms are positively and significantly correlated with well use. In ascending order of importance, location choice explains 3% of the variance in well use; geology, 8%; geography, 19%; community take-up, 29%; and household take-up, 30%. Because this last term is a residual, any measurement error in any of the other components loads onto this term, so 30% is likely an upper bound for the share of variance that is unexplained by the four measurable components.

Table K2: Regression results

	(1)	(2)	(3)	(4)	(5)
Geography	0.88*** (0.22)				
Location choice		0.51** (0.23)			
Community take-up			0.71*** (0.07)		
Geology				0.77*** (0.12)	
Household take-up residual					0.56*** (0.10)
Constant	-0.41** (0.20)	0.10 (0.15)	0.00 (0.01)	-0.23** (0.09)	-0.12 (0.09)
R^2	0.19	0.03	0.29	0.08	0.30

Notes Results from pairwise regressions of normalized takeup on individual components. Undefined terms set to their mean value. Robust standard errors in parentheses. ** $p < 0.05$ *** $p < 0.01$

Appendix Tables K3a to K3c complete this exercise by comparing how each of the components varies across the three treatment arms. This is not necessarily the most efficient way to test for differences across treatment arms but it is useful to put all these mechanisms on a level playing field in terms of sample size. Appendix Table K3a shows results for all treated communities. Only the community take-up component varies significantly across treatments, and the differences are very stark. As we show in Appendix J, location choices yield highest relative predicted use for the labour treatment arm, but in this sample with this measure of location quality, the differences are not statistically significant. The exogenously-determined terms, geography and geology, are balanced across treatment arms, as we would expect to be the case on average given random assignment to treatment.

Because we set undefined terms to their mean value for this exercise, one might worry that we are biased against finding differences between these components across treatment arms. We evaluate whether this affects our results by conditioning on whether communities completed program requirements for at least one well, resulting in at least one attempted installation (Appendix Table K3b), and whether we installed at least one well (Appendix Table K3c). Because we are conditioning on community take-up, we no longer report results for this component. Neither the exogenously determined terms, geography and geology, nor household take-up show differences in these selected subsamples either. Differences in location quality are statistically significant for one comparison between models in one of the two selected subsamples. That household take-up does not vary across treatment arms is also confirmed by comparing the household take-up residual across treatment arms in household-level data (Appendix Table K4).

Table K3a: Components of well use decomposition: all treated communities

	Cash	Labour	Waiver	C = W	L = W	C = L	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Geography	0.97 (0.04)	1.02 (0.04)	0.99 (0.05)	0.562	0.688	0.376	129
Location quality	0.70 (0.03)	0.76 (0.02)	0.72 (0.03)	0.554	0.260	0.218	129
Community take-up	0.22 (0.06)	0.85 (0.04)	0.88 (0.05)	0.001	0.562	0.001	129
Geology	0.90 (0.03)	0.87 (0.03)	0.93 (0.03)	0.406	0.114	0.452	129
Household take-up (residual)	1.05 (0.02)	1.06 (0.09)	1.00 (0.10)	0.938	0.816	0.838	129

Table K3b: Components of well use decomposition: all communities with at least one attempted installation

	Cash	Labour	Waiver	C = W	L = W	C = L	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Geography	0.98 (0.05)	1.02 (0.04)	1.03 (0.05)	0.388	0.662	0.934	88
Location quality	0.63 (0.03)	0.75 (0.03)	0.72 (0.03)	0.394	0.392	0.086	88
Geology	0.92 (0.08)	0.88 (0.03)	0.94 (0.03)	0.786	0.196	0.794	88
Household take-up (residual)	0.80 (0.02)	1.03 (0.09)	1.03 (0.11)	0.270	0.868	0.488	88

Table K3c: Components of well use decomposition: all communities with at least one successful installation

	Cash	Labour	Waiver	C = W	L = W	C = L	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Geography	1.13 (0.07)	1.02 (0.04)	1.03 (0.04)	0.780	0.502	0.610	82
Location quality	0.74 (0.00)	0.73 (0.03)	0.72 (0.03)	0.714	0.486	0.824	82
Geology	1.00 (.)	0.96 (0.01)	0.97 (0.02)	0.278	0.708	0.928	82
Household take-up (residual)	0.75 (0.02)	1.02 (0.11)	1.09 (0.11)	0.288	0.922	0.550	82

Notes Tables report means and robust standard errors in parentheses. Undefined terms set to their mean value. Regressions control for centered strata indicators interacted with the treatment dummies, arbitrarily omitting one stratum. Significance levels obtained from randomization-based inference. p values from pairwise hypothesis tests in columns 4-6.

Table K4: Residual household take-up: household-level data

	Cash	Labour	Waiver	C = W	L = W	C = L	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Residualized share of water	0.01 (0.03)	0.00 (0.01)	0.00 (0.01)	0.444	0.798	0.560	3115

Notes Table reports means and clustered standard errors in parentheses. Sample includes households surveyed at baseline in communities in which we constructed at least one well. “Residualized share of water” measures the share of water obtained from project wells at endline after nonparametrically parsing out predicted use based on i) baseline household arsenic contamination and ii) distance to a constructed well. Weights are applied so that each community counts equally in summary statistics. Regressions include endline data only and control for centered strata indicators interacted with the treatment dummies, arbitrarily omitting one stratum. Significance levels obtained from randomization-based inference. p values from pairwise hypothesis tests in columns 4-6.

Examining patterns of well use yields insights into mechanisms that drive overall impact if the intervention primarily works through providing new safe wells in all treatment arms. We find no evidence for other potential channels via which the project could affect exposure to arsenic contamination. For example, project wells may displace a small amount of private well construction (Appendix Table K5). But most new private wells are shallow and not tested for arsenic, and are no more or less likely to be arsenic contaminated than other wells in the same community. New private wells therefore do little to reduce exposure to arsenic (Cocciolo et al., 2021). Across communities, impact correlates strongly and fairly linearly with use of project wells (Appendix Table K6).

Table K5: Use of new sources

	Control Mean	Estimated impact			C = W	L = W	C = L	N
	(1)	Cash (2)	Labour (3)	Waiver (4)	(5)	(6)	(7)	(8)
Share from newly-adopted sources	0.09 (0.01)	0.01 (0.01)	0.06*** (0.01)	0.04*** (0.01)	0.014	0.224	<0.001	7121
Share from unlisted sources at endline	0.16 (0.02)	-0.03 (0.02)	-0.02 (0.02)	-0.04 (0.02)	0.372	0.290	0.766	7122

Notes Table reports program impact relative to control group. The sample is all randomly-sampled households in study communities. Clustered standard errors are in parentheses. “Share from newly-adopted sources” measures the share of water obtained from sources that the household self-reports adopting since baseline. “Share from unlisted sources at endline” measures the share of water obtained from sources that were neither listed at baseline nor project wells, most of which reflect non-project sources that were constructed between baseline and endline. Weights are applied so that each community counts equally in summary statistics. Regressions include endline data only, centered strata indicators, and centered strata indicators interacted with the treatment dummies. Significance levels obtained from randomization-based inference. p values from pairwise hypothesis tests in columns 5-7. Asterisks in columns 2-4 reflect significance of impacts with respect to the control group: *** $p < 0.01$.

Table K6: Correlation between use of project wells and impact on arsenic contamination

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treated	-0.05 (0.04)	0.01 (0.04)	0.00 (0.04)					
Share		-0.91*** (0.27)	-0.54 (0.64)		-0.81*** (0.27)	-0.04 (0.72)		
Share ²			-1.27 (2.15)			-2.52 (2.28)		
Cash				0.02 (0.04)	0.04 (0.04)	0.03 (0.04)	0.01 (0.04)	0.00 (0.04)
Labour				-0.08 (0.05)	-0.01 (0.05)	-0.04 (0.06)	0.01 (0.06)	-0.05 (0.07)
Waiver				-0.09* (0.05)	-0.02 (0.05)	-0.05 (0.05)	0.03 (0.05)	-0.02 (0.06)
Cash x share							0.06 (0.38)	1.48* (0.79)
Labour x share							-1.02** (0.51)	0.73 (1.38)
Waiver x share							-1.27*** (0.38)	-0.08 (1.16)
Cash x share ²								-3.81** (1.75)
Labour x share ²								-7.11 (5.79)
Waiver x share ²								-3.86 (3.17)
Constant	-0.11*** (0.03)	-0.11*** (0.03)	-0.11*** (0.03)	-0.11*** (0.03)	-0.11*** (0.03)	-0.11*** (0.03)	-0.11*** (0.03)	-0.11*** (0.03)
R ²	0.01	0.10	0.11	0.04	0.11	0.12	0.14	0.17
AIC	-11.6	-26.6	-25.2	-13.6	-24.1	-24.0	-26.5	-26.3
BIC	-5.4	-17.2	-12.6	-1.1	-8.4	-5.2	-4.5	5.1

Notes Results from OLS regression of community mean change in arsenic contamination index between baseline and endline. *Share* is the mean share of household drinking water collected from a project well. Robust standard errors in parentheses. N = 171. * $p < 0.1$ ** $p < 0.05$ *** $p < 0.01$

L Selection

Appendices L1, L2 and L3 report detailed results for each of the three types of selection we study: successful well installation, well use, and contributions. We provide additional information on the naïve comparisons we draw in Figures 4, 5, and 6. We additionally report detailed results from the machine learning approaches described in Appendix E. For all three machine learning exercises, a common feature is that we cannot statistically reject the null of no heterogeneity in selection with respect to the baseline characteristics we include. However, in all three cases, we do observe systematic differences in which characteristics are correlated with predicted selection.

L1 Describing community-level selection into well installation

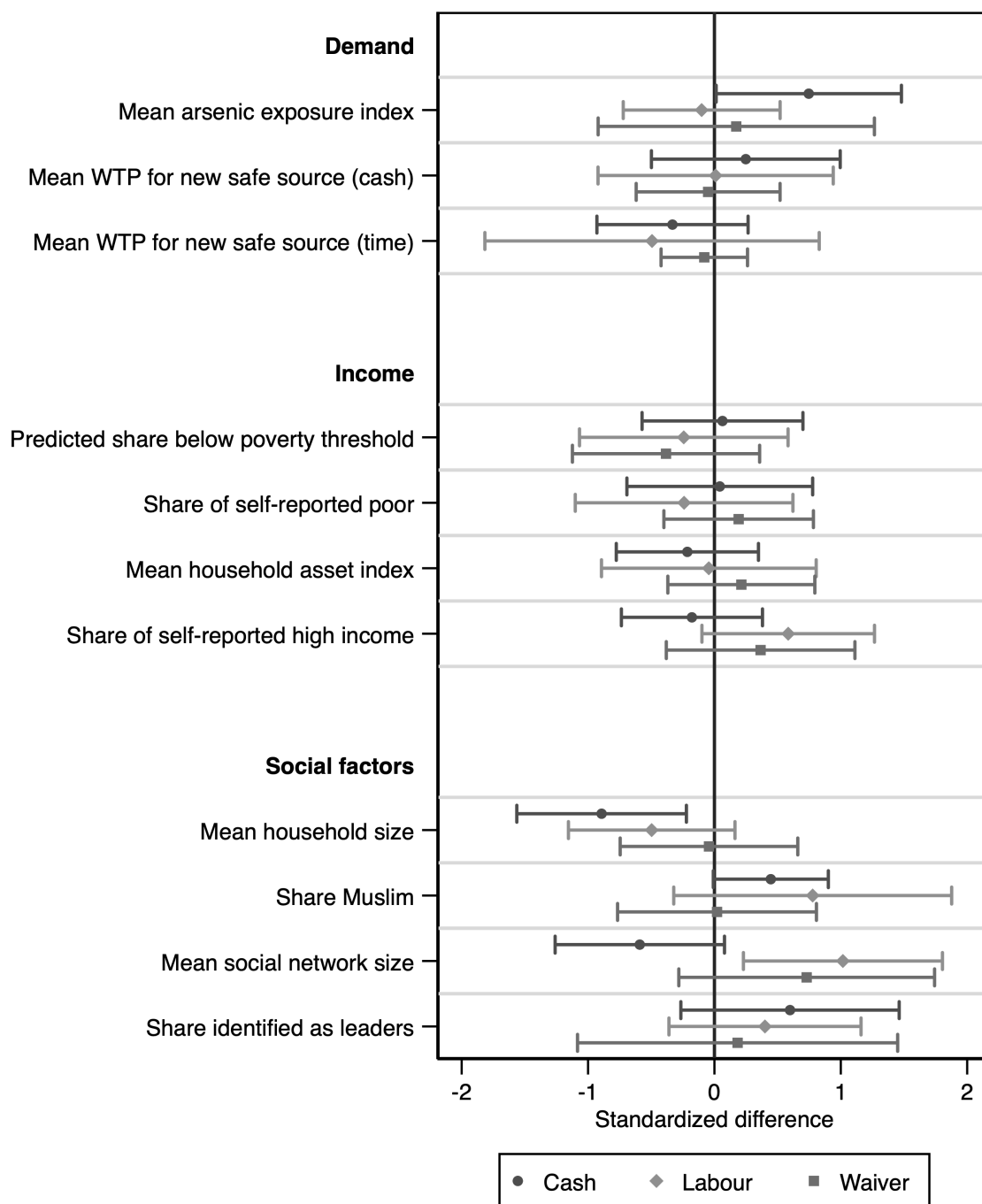
We focus primarily on selection into attempted installation (“community take-up”), because the effectively random variation in whether or not an attempted well installation is successful just introduces noise into the selection process. Appendix Table L1 provides a tabular version of the comparisons shown in Figure 4, showing that arsenic exposure at baseline is most strongly correlated with selection into attempted installation under the cash treatment arm. Appendix Figure L1 and Appendix Table L2 repeat the exercise for selection into successful installation, confirming that patterns are similar but less precisely estimated.

Table L1: Selection into attempted installation

	Cash x Attempt	Labour x Attempt	Waiver x Attempt	C = W	L = W	C = L	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
a) Demand							
Mean arsenic exposure index	0.96*** (0.37)	0.12 (0.36)	0.72 (0.69)	0.620	0.440	0.070	129
Mean WTP for new safe source (cash)	0.12 (0.37)	0.91*** (0.29)	0.25 (0.23)	0.772	0.102	0.144	129
Mean WTP for new safe source (time)	-0.36 (0.29)	-1.09 (1.10)	-0.18 (0.18)	0.766	0.860	0.212	129
Predicted well use at optimal location	0.64 (0.47)	-0.04 (0.45)	0.61 (0.42)	0.956	0.340	0.120	129
b) Income							
Predicted share below poverty threshold	0.27 (0.33)	0.25 (0.60)	-0.17 (0.44)	0.326	0.652	0.968	129
Share of self-reported poor	-0.16 (0.37)	0.01 (0.65)	0.12 (0.30)	0.494	0.922	0.710	129
Mean household asset index	-0.19 (0.27)	0.07 (0.63)	0.31 (0.32)	0.276	0.804	0.548	129
Share self-reported high income	-0.26 (0.28)	0.23 (0.42)	0.26 (0.47)	0.250	0.952	0.312	129
c) Social factors							
Mean household size	-0.84** (0.32)	-0.01 (0.41)	0.32 (0.27)	0.034	0.534	0.066	129
Share Muslim	0.43* (0.24)	0.72 (0.88)	0.23 (0.45)	0.710	0.826	0.576	129
Mean social network size	-0.50 (0.32)	0.83 (0.52)	0.40 (0.37)	0.072	0.630	0.004	129
Share identified as leaders	0.48 (0.41)	-0.34 (0.39)	-0.08 (0.78)	0.238	0.780	0.044	129

Notes Each row reports results from a separate regression of community mean baseline or predetermined characteristics on treatment arm dummies and treatment arm dummies interacted with the fraction of wells offered to a given community for which the community successfully completed all project requirements. The unit of observation is the community. Robust standard errors are shown in parentheses. p values reported in columns 4-6 test the null hypothesis that the difference between successful and unsuccessful communities is equal under each pair of contribution requirement treatments, using randomization-based inference. Asterisks shown in columns 1-3 reflect significance of the difference between successful and unsuccessful communities under the respective treatment arm, based on analytical standard errors: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure L1: How did communities in which we successfully installed wells differ from those in which we did not?



Notes Figure shows regression-estimated difference between communities in which the project installed all offered wells and those in which we installed no offered well. Regressions include treatment arm fixed effects. 95% confidence intervals reflect robust standard errors. All variables are standardized so that differences are in standard deviations.

Table L2: Selection into successful installation

	Cash x Success	Labour x Success	Waiver x Success	C = W	L = W	C = L	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
a) Demand							
Mean arsenic exposure index	0.75** (0.37)	-0.10 (0.31)	0.17 (0.55)	0.230	0.646	0.076	129
Mean WTP for new safe source (cash)	0.25 (0.38)	0.01 (0.47)	-0.05 (0.29)	0.524	0.926	0.652	129
Mean WTP for new safe source (time)	-0.33 (0.30)	-0.49 (0.67)	-0.08 (0.17)	0.656	0.968	0.748	129
b) Income							
Predicted share below poverty threshold	0.06 (0.32)	-0.24 (0.42)	-0.38 (0.37)	0.308	0.810	0.520	129
Share of self-reported poor	0.04 (0.37)	-0.24 (0.44)	0.19 (0.30)	0.722	0.496	0.522	129
Mean household asset index	-0.21 (0.28)	-0.04 (0.43)	0.21 (0.29)	0.342	0.650	0.714	129
Share self-reported high income	-0.18 (0.28)	0.58* (0.35)	0.37 (0.38)	0.238	0.674	0.134	129
c) Social factors							
Mean household size	-0.89*** (0.34)	-0.50 (0.33)	-0.04 (0.36)	0.104	0.356	0.368	129
Share Muslim	0.45* (0.23)	0.78 (0.56)	0.02 (0.40)	0.416	0.410	0.496	129
Mean social network size	-0.59* (0.34)	1.02** (0.40)	0.73 (0.51)	0.006	0.650	<0.001	129
Share identified as leaders	0.60 (0.44)	0.40 (0.38)	0.18 (0.64)	0.350	0.772	0.640	129

Notes Each row reports results from a separate regression of community mean baseline or predetermined characteristics on treatment arm dummies and treatment arm dummies interacted with the fraction of wells offered to a given community that were successfully installed. The unit of observation is the community. Robust standard errors are shown in parentheses. p values reported in columns 4-6 test the null hypothesis that the difference between successful and unsuccessful communities is equal under each pair of contribution requirement treatments, using randomization-based inference. Asterisks shown in columns 1-3 reflect significance of the difference between successful and unsuccessful communities under the respective treatment arm, based on analytical standard errors: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Because the lack of selection on income is striking, we added additional measures of income to the exercise. Appendix Figure L2a repeats the exercise in Figure 4 for measures of access to credit or savings, possibly a better proxy for the availability of resources in cash. Appendix Figure L2a repeats the exercise for arsenic-exposure-weighted income, which might better capture the availability of cash among those affected by arsenic. We find no evidence for systematic selection on income in these dimensions either.

Figure L2a: Communities that did and did not complete requirements: Credit and savings

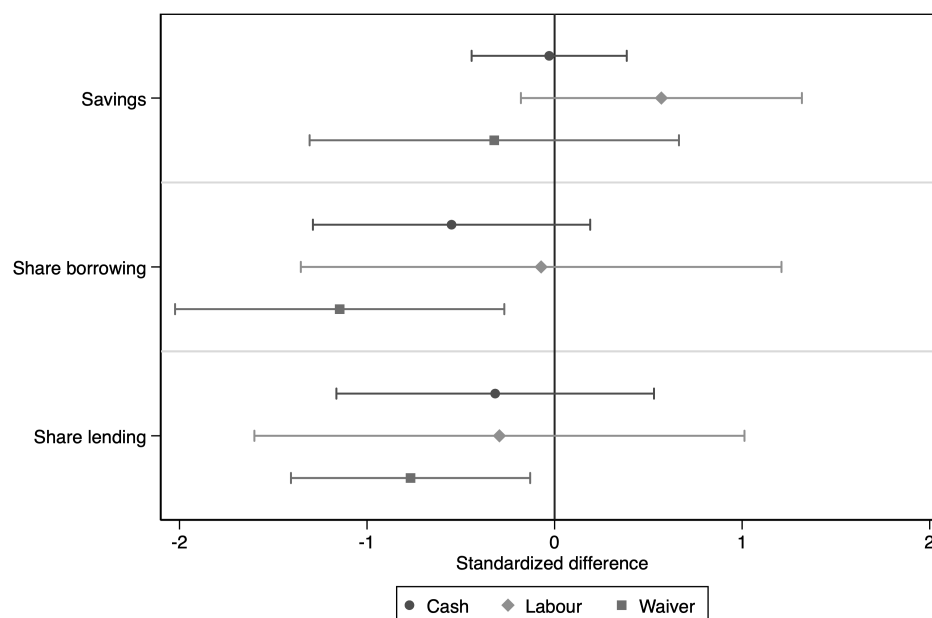
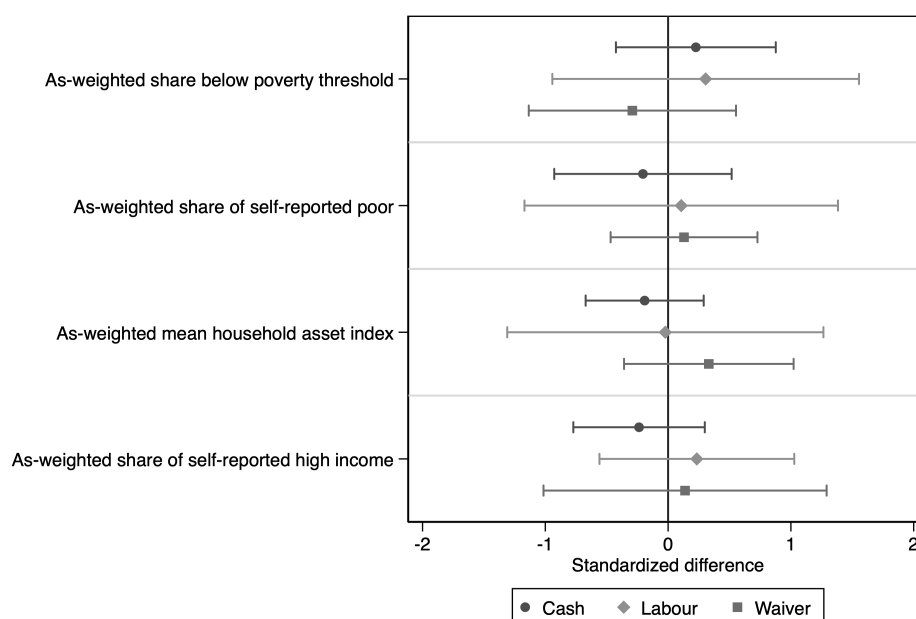


Figure L2b: Communities that did and did not complete requirements: As-weighted income



Notes Graph shows regression-estimated difference between communities that successfully completed project requirements for all offered wells and those that did not complete project requirements for any offered well. Regressions include treatment arm fixed effects. 95% confidence intervals reflect robust standard errors. All variables are standardized so that differences are in standard deviations.

Turning to results from machine learning analyses, Appendix Tables L3 and L4 show that we have limited ability to detect heterogeneity, either in average take-up rates or across treatment arms (note the wide confidence intervals on the heterogeneity parameters, $\hat{\beta}$ in Equation 5 and $\hat{\beta}_2$ in Equation 2). Appendix Figures L3 to L4 show similarly that we cannot reject the null of equality of take-up across quartiles with the highest predicted take-up and lowest predicted take-up, or quartiles with the highest and lowest differential take-up across treatment arms, respectively. However, note that the average treatment effects on take-up reported in Appendix Table L4 are, as in Table 2, large and strongly significant for the comparison between the cash treatment arm and either of the other two treatment arms.

Table L3: Attempted installation: average characteristics, best linear predictor

Treated vs control	
AVE	HET
0.65	-0.01
(0.54, 0.77)	(-0.79, 0.76)
[<0.001]	[1.000]

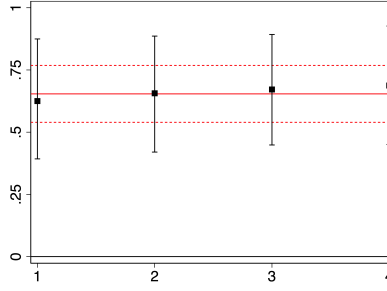
Notes Best linear predictor of conditional average value. Medians over 100 splits. 90% confidence interval in parentheses. p values for the hypothesis that the parameter is equal to zero in brackets.

Table L4: Attempted installation: heterogeneity across treatment arms, best linear predictor

Cash vs waiver		Labour vs waiver		Cash vs labour	
ATE	HET	ATE	HET	ATE	HET
-0.65	-0.05	-0.03	-0.04	-0.63	0.52
(-0.88, -0.42)	(-1.22, 1.16)	(-0.25, 0.17)	(-1.39, 1.31)	(-0.86, -0.40)	(-0.64, 1.73)
[<0.001]	[1.000]	[1.000]	[1.000]	[<0.001]	[0.786]

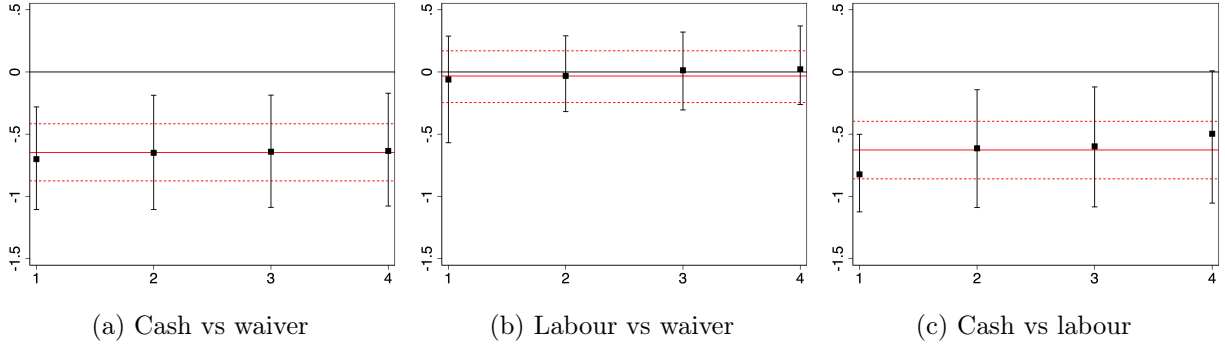
Notes Best linear predictor of conditional average treatment effect. Medians over 100 splits. 90% confidence interval in parentheses. p values for the hypothesis that the parameter is equal to zero in brackets.

Figure L3: GAV, attempted installations



Notes Graph plots the estimated value of listed outcome variable by quartile defined by heterogeneity score. x axis shows quartiles defined by predicted heterogeneity. y axis shows predicted value of listed variable for the listed quartile. Average value shown in red.

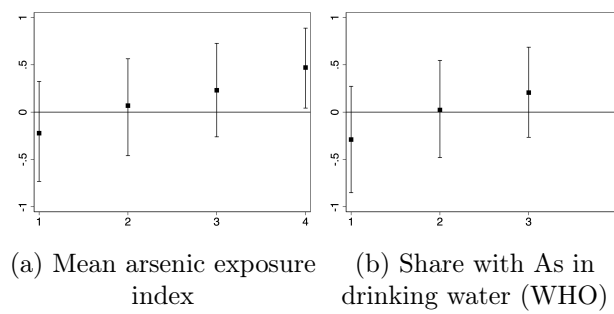
Figure L4: GADTES, attempted installations



Notes Graphs plot the estimated differential treatment effect by quartile defined by heterogeneity score. x axis shows quartiles defined by predicted heterogeneity. y axis shows estimated differential effect of treatment under the first listed contribution requirement compared to the second-listed contribution requirement. Average differential treatment effect shown in red.

Appendix Figure L5 shows which variables are on average correlated with predicted take-up. We plot only variables for which the quartile with the highest predicted take-up differs from the quartile with the lowest predicted take-up (p 0.1). On average, pooling across all three treatment arms, only measures of arsenic contamination—specifically, the index and the share with arsenic contamination in their drinking water at the WHO threshold—are correlated with predicted take-up. In the pairwise classification analysis, no variables are significantly differentially correlated with differences in predicted take-up under any pair of treatment arms.

Figure L5: CLAN, attempted installations, pooled



Notes Graphs plot the normalized value of the listed characteristic by quartile defined by heterogeneity score. x axis shows quartiles defined by predicted heterogeneity. y axis shows normalized value of listed characteristics. Only characteristics for which the top and bottom quartiles differ significantly ($p < 0.1$) are shown.

L2 Describing household-level selection into well use

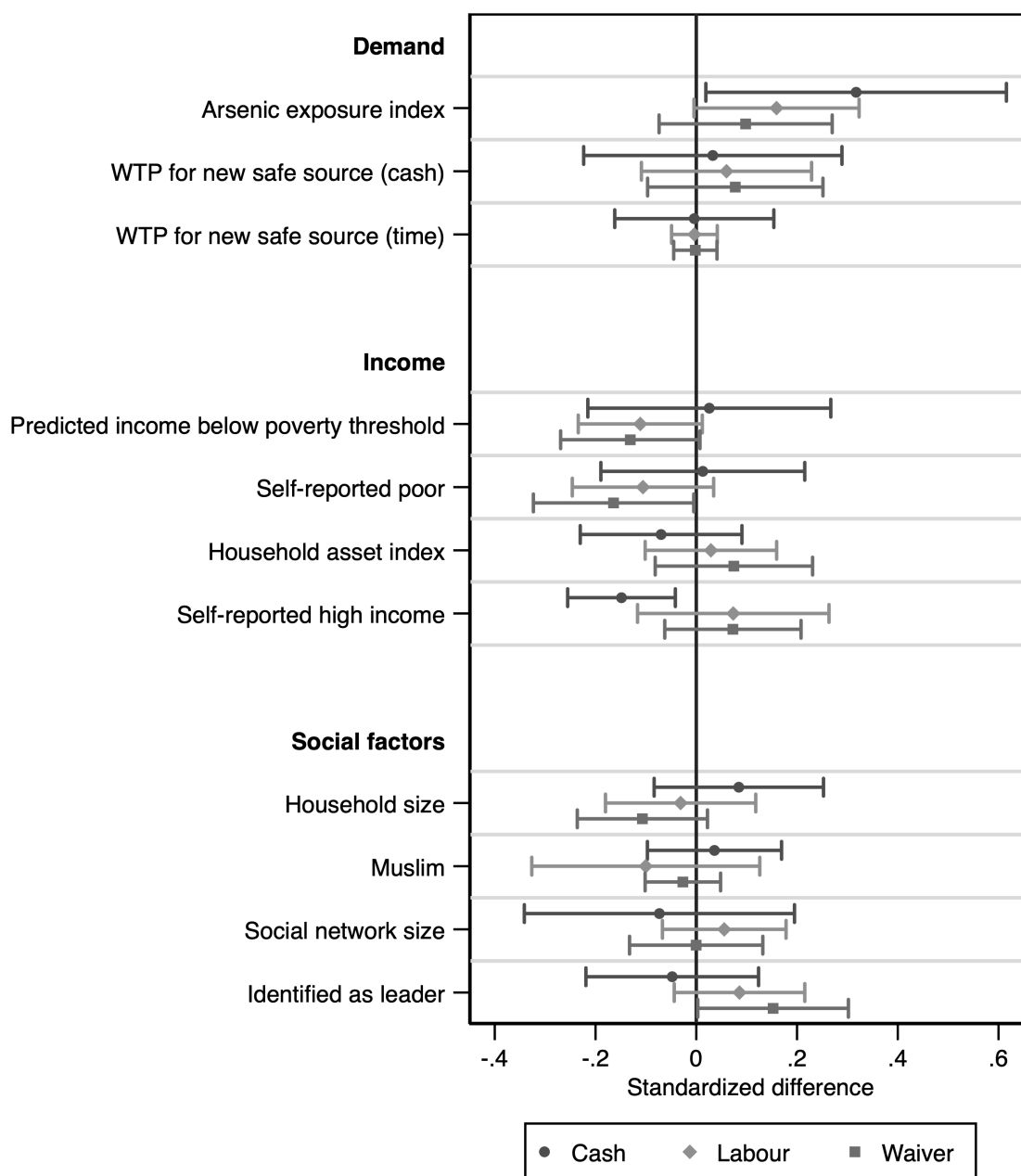
Appendix Table L5 provides a tabular version of the comparisons shown in Figure 5. Appendix Figure L6 and Appendix Table L6 repeat the same exercise for a simpler comparison of well users to non-users, without accounting for intensity of use. The patterns are very similar.

Table L5: Selection into greater use of wells

	Cash x Use (1)	Labour x Use (2)	Waiver x Use (3)	C = W (4)	L = W (5)	C = L (6)	N (7)
a) Demand							
Arsenic exposure index	0.57*** (0.19)	0.36** (0.15)	0.13 (0.14)	0.066	0.326	0.448	4779
Mean WTP for new safe source (cash)	-0.29 (0.21)	0.20 (0.21)	0.14 (0.17)	0.164	0.860	0.176	4779
Mean WTP for new safe source (time)	0.03 (0.13)	0.01 (0.04)	0.00 (0.04)	0.696	0.918	0.776	4779
b) Income							
Predicted income below poverty threshold	0.12 (0.26)	-0.17 (0.14)	-0.12 (0.14)	0.374	0.790	0.256	5841
Self-reported poor	0.01 (0.18)	-0.10 (0.14)	-0.16 (0.14)	0.476	0.704	0.656	5860
Household asset index	-0.37*** (0.14)	-0.16 (0.14)	-0.07 (0.11)	0.144	0.632	0.358	5842
Self-reported high income	-0.20** (0.08)	-0.05 (0.13)	0.04 (0.11)	0.186	0.600	0.570	5860
c) Social factors							
Household size	-0.23 (0.16)	-0.38** (0.16)	-0.27* (0.15)	0.880	0.636	0.600	4779
Muslim	0.10 (0.07)	-0.17 (0.20)	-0.09 (0.11)	0.324	0.766	0.664	5868
Social network size	-0.23 (0.24)	0.04 (0.12)	-0.16 (0.16)	0.774	0.328	0.232	5868
Identified as leader	-0.22 (0.25)	-0.02 (0.09)	0.30** (0.13)	0.018	0.036	0.238	5868

Notes Each row reports results from a separate regression of household mean baseline or pre-determined characteristics on community dummies and treatment arm dummies interacted with the volumetric share of household water obtained from a project well. The unit of observation is the household. Clustered standard errors are shown in parentheses. *p* values reported in columns 4-6 test the null hypothesis that the difference between full adopters and non-adopters is equal under each pair of contribution requirement treatments, using randomization-based inference. Asterisks shown in columns 1-3 reflect significance of the difference between full adopters and non-adopters under the respective treatment arm, based on analytical standard errors: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure L6: Users compared to non-users



Notes Graph shows within-community regression-estimated difference between households that collected any water from a project well and households that collected no water from a project well at endline. Regressions include community fixed effects. Survey weights applied. 95% confidence intervals reflect clustered standard errors. All variables are standardized so that differences are in standard deviations.

Table L6: Selection into use of wells

	Cash x User	Labour x User	Waiver x User	C = W	L = W	C = L	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
a) Demand							
Arsenic exposure index	0.32** (0.15)	0.16* (0.08)	0.10 (0.09)	0.120	0.654	0.280	4779
Mean WTP for new safe source (cash)	0.03 (0.13)	0.06 (0.09)	0.08 (0.09)	0.776	0.878	0.856	4779
Mean WTP for new safe source (time)	0.00 (0.08)	0.00 (0.02)	0.00 (0.02)	0.956	0.950	>0.999	4779
b) Income							
Predicted income below poverty threshold	0.03 (0.12)	-0.11* (0.06)	-0.13* (0.07)	0.210	0.840	0.244	5841
Self-reported poor	0.01 (0.10)	-0.11 (0.07)	-0.16** (0.08)	0.212	0.540	0.328	5860
Household asset index	-0.07 (0.08)	0.03 (0.07)	0.07 (0.08)	0.280	0.668	0.412	5842
Self-reported high income	-0.15*** (0.05)	0.07 (0.10)	0.07 (0.07)	0.072	0.996	0.144	5860
c) Social factors							
Household size	0.08 (0.08)	-0.03 (0.08)	-0.11 (0.07)	0.076	0.468	0.370	4779
Muslim	0.04 (0.07)	-0.10 (0.11)	-0.03 (0.04)	0.442	0.620	0.654	5868
Social network size	-0.07 (0.14)	0.06 (0.06)	0.00 (0.07)	0.538	0.568	0.322	5868
Identified as leader	-0.05 (0.09)	0.09 (0.07)	0.15** (0.08)	0.156	0.500	0.254	5868

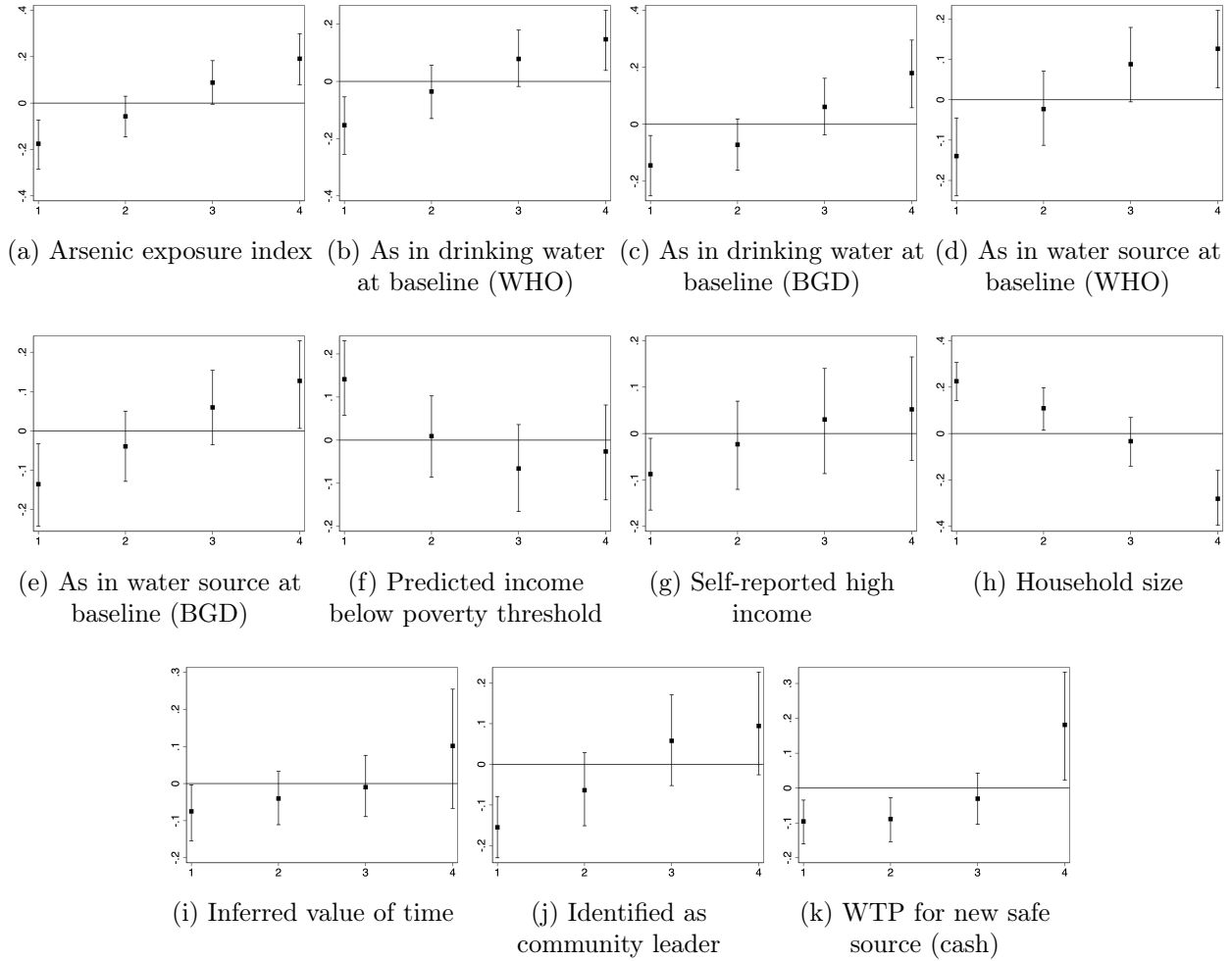
Notes Each row reports results from a separate regression of household mean baseline or pre-determined characteristics on community dummies and treatment arm dummies interacted with a dummy that takes the value one if a household uses a project well. The unit of observation is the household. Clustered standard errors are shown in parentheses. *p* values reported in columns 4 to 6 test the null hypothesis that the difference between users and non-users is equal under each pair of contribution requirement treatments, using randomization-based inference. Asterisks shown in columns 1 to 3 reflect significance of the difference between users and non-users under the respective treatment arm, based on analytical standard errors: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Turning to results from machine learning analyses, we predict use relative to community means and residualize all household characteristics with respect to their community means. We restrict the sample to communities in which we successfully installed wells. We cannot reject the null that there is no heterogeneity in well use with respect to baseline household-level characteristics. We also cannot reject the null hypotheses that use rates are the same across the quartiles with highest and lowest relative propensity to use the well on average. Nor can we reject the corresponding null hypotheses for differences across treatment arms. As also shown elsewhere, average use rates also do not differ across treatment arms in communities in which we successfully installed wells. In the interests of brevity, we do not report the best linear predictor or quartile results in this appendix.

Appendix Figure L7 shows which variables are on average correlated with predicted use. We plot only variables for which the quartile with the highest predicted use differs from the quartile with the lowest predicted use with a p value of less than 0.1. On average, pooling across communities that successfully installed wells in all three treatment arms, measures of arsenic contamination are strongly correlated with predicted use, i.e., those households who have higher arsenic contamination compared to their community are more likely to adopt the wells. Users appear slightly wealthier than non-users. Leaders and those with a high willingness to pay for a new safe source more likely to use wells. Also, smaller households appear more likely to use the well. Smaller households also have fewer children as a share of household size, so they may have fewer time constraints, allowing them to incur the additional time costs of using a community well.

Appendix Figures L8a to L8c report the variables that are most strongly associated with larger predicted differences in use of the wells between the treatment arms. Again, we plot only variables for which we reject the null hypothesis that the variables are equal in the quartiles with the highest and lowest predicted differences in well use. Exposure to arsenic contamination more strongly predicts use under the cash treatment arm, compared to either of the other two treatment arms. Poverty positively predicts use under the cash treatment arm in contrast to negatively predicting use under the waiver or labour treatment arms. Willingness to pay for a new safe source also correlates with higher predicted use in the labour treatment arm more strongly than in the waiver treatment arms: curiously, in cash rather than time. Smaller households have higher predicted use under the labour treatment arm than under the other treatment arms.

Figure L7: CLAN, well use, pooled



Notes Graphs plot the normalized value of the listed characteristic by quartile defined by heterogeneity score. x axis shows quartiles defined by predicted heterogeneity. y axis shows normalized value of listed characteristics. Only characteristics for which the top and bottom quartiles differ significantly ($p < 0.1$) are shown.

Figure L8a: CLAN, well use, cash vs waiver

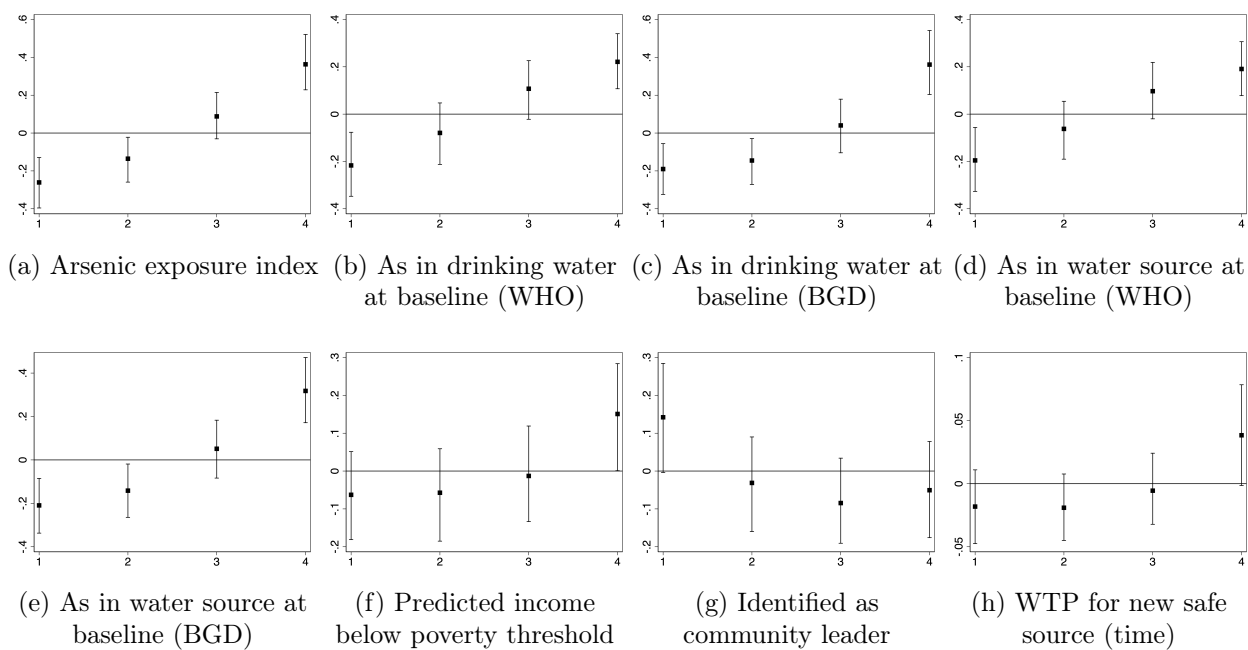


Figure L8b: CLAN, well use, labour vs waiver

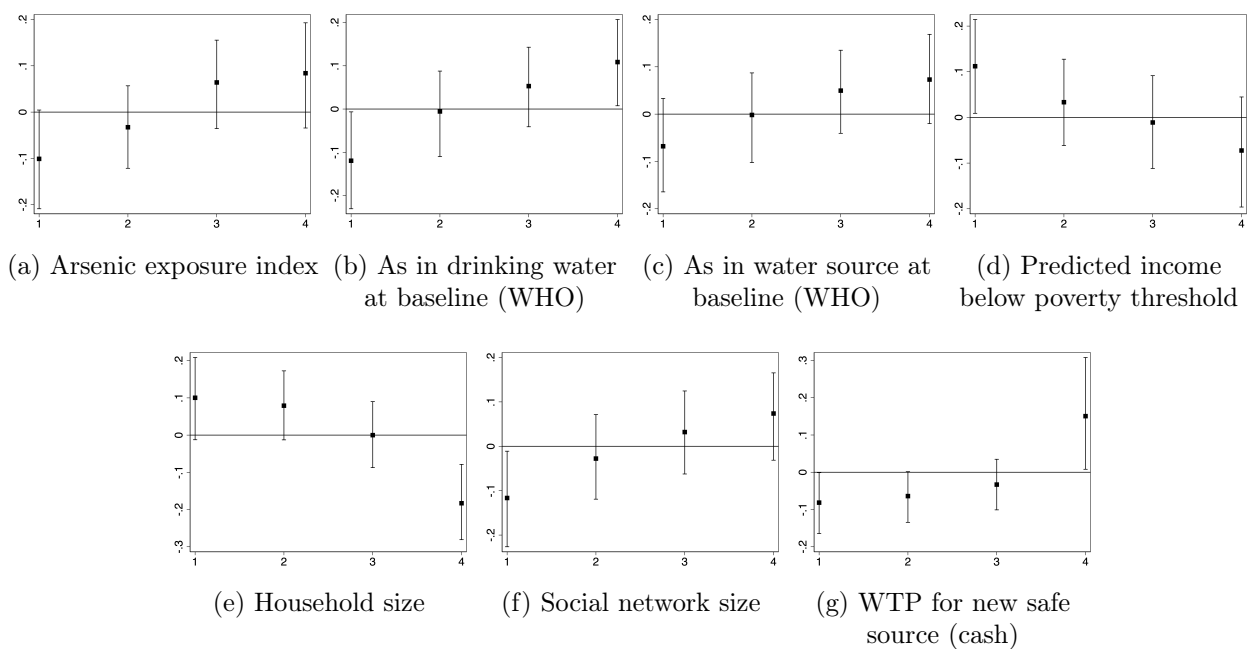
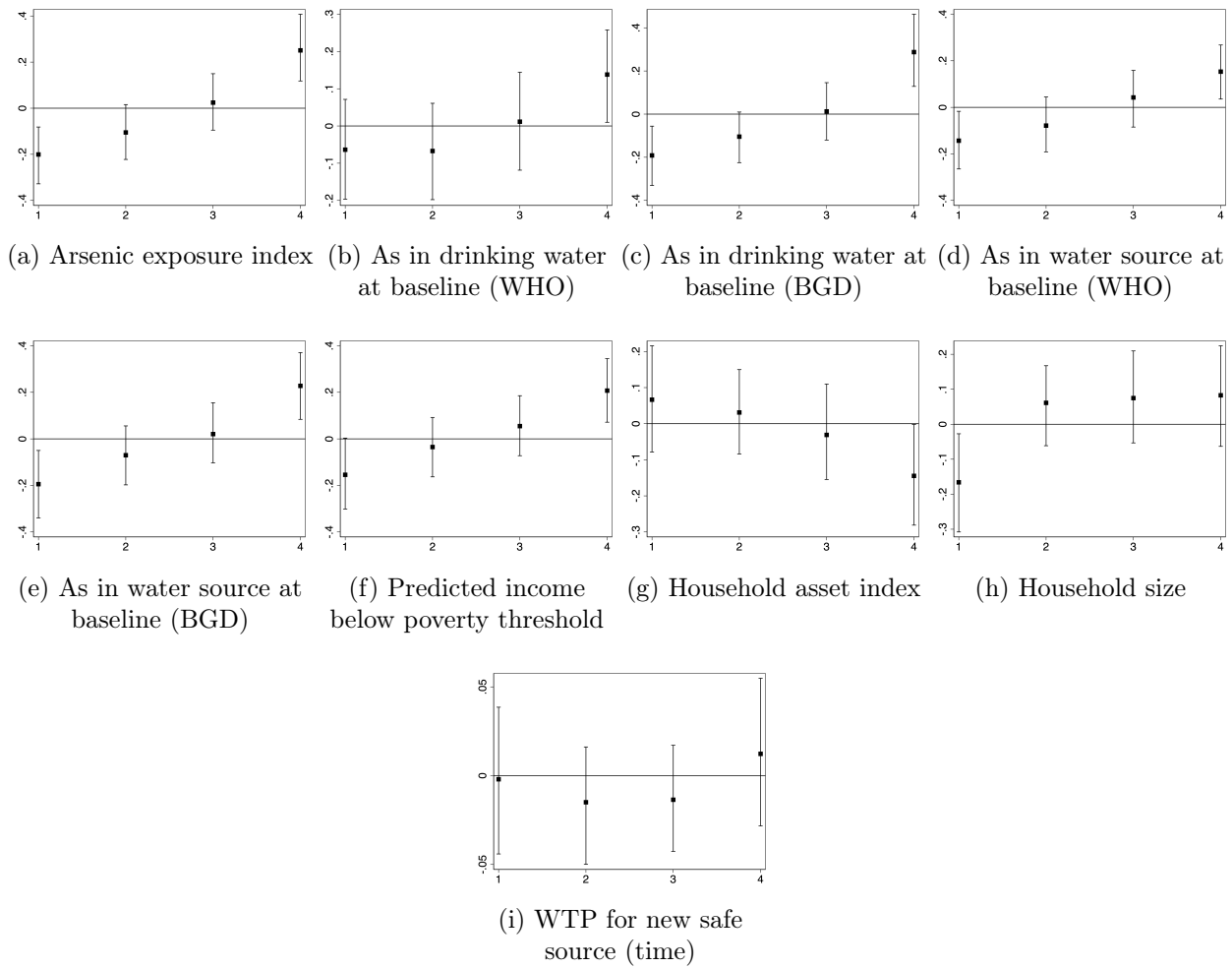


Figure L8c: CLAN, well use, cash vs labour



Notes Graphs plot the normalized value of the listed characteristic by quartile defined by heterogeneity score. x axis shows quartiles defined by predicted heterogeneity. y axis shows normalized value of listed characteristics. Only characteristics for which the top and bottom quartiles differ significantly ($p < 0.1$) are shown.

L3 Describing household-level selection into contribution

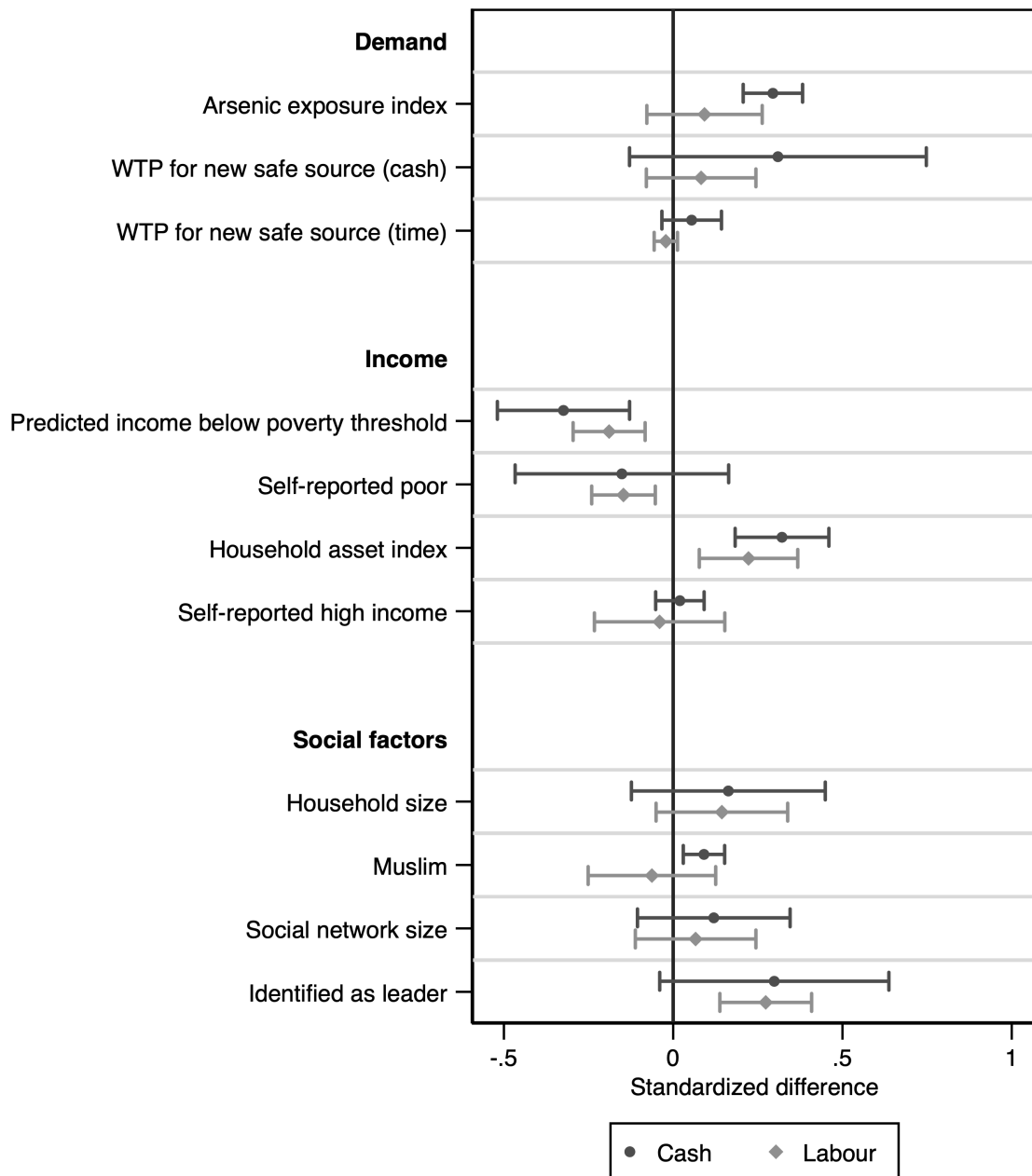
Appendix Table L7 provides a tabular version of the comparisons shown in Figure 6. Both these exercises use information on whether households contribute or not from project records. Appendix Figure L9 and Appendix Table L8 repeat the same exercise for self-reported data on contributions. Because households who reported contributing during the surveys despite not being recorded as contributing in project records are wealthier and more likely to be identified as leaders, using self-reported data intensifies the correlation between contributing and measures of wealth or leadership status. Correlation with baseline arsenic contamination remains very similar.

Table L7: Characteristics of contributing households: project records

	Cash x Contributor	Labour x Contributor	C = L	N
	(1)	(2)	(3)	(4)
a) Demand				
Arsenic exposure index	0.39*** (0.11)	0.19 (0.12)	0.056	1687
Mean WTP for new safe source (cash)	0.66 (0.42)	0.07 (0.16)	0.042	1687
Mean WTP for new safe source (time)	-0.01 (0.06)	0.03 (0.03)	0.770	1687
b) Income				
Predicted income below poverty threshold	-0.39*** (0.13)	-0.15 (0.11)	0.044	2201
Self-reported poor	-0.15 (0.16)	-0.17* (0.09)	0.570	2210
Household asset index	0.37** (0.14)	0.07 (0.07)	0.004	2201
Self-reported high income	-0.02 (0.07)	-0.18*** (0.05)	0.998	2210
c) Social factors				
Household size	0.31* (0.15)	0.15 (0.16)	0.208	1687
Muslim	0.09** (0.04)	-0.08 (0.10)	0.338	2216
Social network size	0.09 (0.10)	-0.02 (0.10)	0.214	2216
Identified as leader	0.41** (0.19)	0.20 (0.12)	0.104	2216

Notes Each row reports results from a separate regression of household baseline or predetermined characteristics on community dummies and treatment arm dummies interacted with a dummy that takes the value one if project staff recorded that the household contributed in cash or labour. Sampled includes all households surveyed at endline or, when the characteristic is measured at baseline, all households surveyed at both baseline and endline, in communities that successfully installed at least one well. Standard errors clustered at community level and shown in parentheses. Weights applied so that each community counts equally in summary statistics. p values reported in column 3 test the null hypothesis that the difference between contributing and non-contributing households is equal under each pair of contribution requirement treatments, using randomization-based inference. Asterisks shown in columns 1-2 reflect significance of the difference between contributing and non-contributing households under the respective treatment arm, based on analytical standard errors: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure L9: Characteristics of contributing households: self-reported data



Notes Graph shows regression-estimated difference between households that report contributing cash or labour and other households in the same communities. Regressions include community fixed effects. Survey weights applied. 95% confidence intervals reflect clustered standard errors. All variables are standardized so that differences are in standard deviations.

Table L8: Characteristics of contributing households: self-reported data

	Cash x Contributor	Labour x Contributor	C = L	N
	(1)	(2)	(3)	(4)
a) Demand				
Arsenic exposure index	0.29*** (0.04)	0.09 (0.09)	0.170	3243
Mean WTP for new safe source (cash)	0.31 (0.22)	0.08 (0.08)	0.188	3243
Mean WTP for new safe source (time)	0.05 (0.04)	-0.02 (0.02)	0.024	3243
b) Income				
Predicted income below poverty threshold	-0.32*** (0.10)	-0.19*** (0.05)	0.136	3992
Self-reported poor	-0.15 (0.16)	-0.15*** (0.05)	0.962	4005
Household asset index	0.32*** (0.07)	0.22*** (0.07)	0.396	3995
Self-reported high income	0.02 (0.04)	-0.04 (0.10)	0.720	4005
c) Social factors				
Household size	0.16 (0.14)	0.14 (0.10)	0.898	3243
Muslim	0.09*** (0.03)	-0.06 (0.09)	0.374	4013
Social network size	0.12 (0.11)	0.07 (0.09)	0.684	4013
Identified as leader	0.30* (0.17)	0.27*** (0.07)	0.884	4013

Notes Each row reports results from a separate regression of household baseline or predetermined characteristics on community dummies and treatment arm dummies interacted with a dummy that takes the value one if the household self-reported contributing in cash or labour. Sampled includes all households surveyed at endline or, when the characteristic is measured at baseline, all households surveyed at both baseline and endline, in communities that successfully installed at least one well. Standard errors clustered at community level and shown in parentheses. Weights applied so that each community counts equally in summary statistics. p values reported in column 3 test the null hypothesis that the difference between contributing and non-contributing households is equal under each pair of contribution requirement treatments, using randomization-based inference. Asterisks shown in columns 1-2 reflect significance of the difference between contributing and non-contributing households under the respective treatment arm, based on analytical standard errors: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

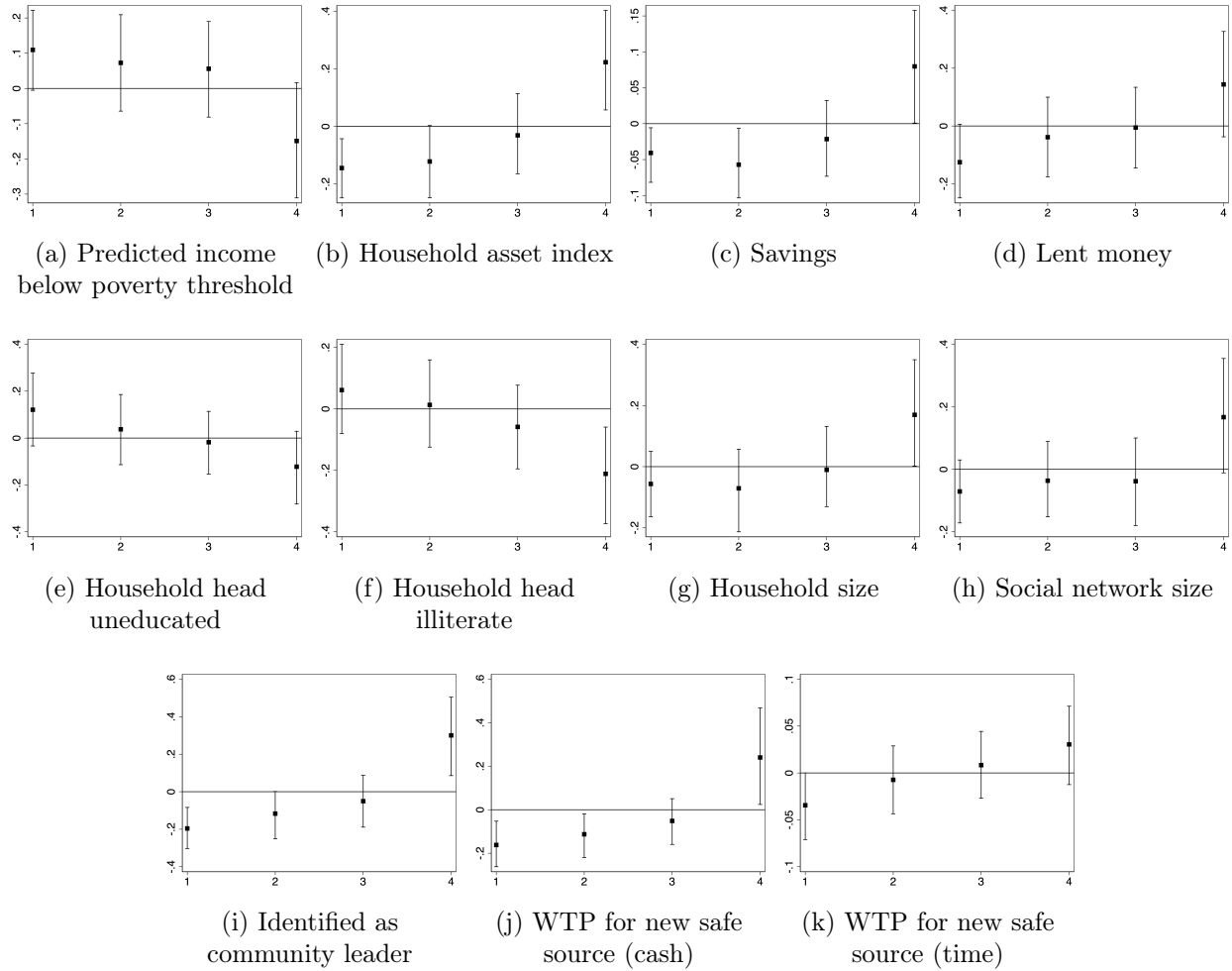
We now turn to results from machine learning. Similarly to well use, we predict contributions relative to community average means, residualize all household characteristics with respect to their community means, and restrict the sample to communities in which we successfully installed wells. We repeat the exercise twice, first using self-reported contributions and then using contributions as recorded by field staff. In practice, the results are very similar for the two measures.

We again cannot reject the null of no heterogeneity in contribution rates according to baseline characteristics, either on average or differentially across treatment arms. We also cannot reject the null hypotheses that contribution rates are the same across the quartiles with highest and lowest relative propensity to contribute on average or the null that the same characteristics are associated with well use across all pairs of treatment arms. Again, in the interests of brevity, we do not report the best linear predictor or quartile results in this appendix.

Appendix Figures L10 and L11 plot the characteristics that are on average correlated with predicted contribution behaviour. Again, we plot only variables for which the quartile with the highest predicted contribution rate differs from the quartile with the lowest predicted contribution rate with a p value of less than 0.1. Again, we repeat the exercise for both self-reported contributions and contributions based on project records. On average, predicted contributions are higher for households with higher incomes across a range of metrics and higher WTP for a new safe source, and those who are identified as either network focal points or community leaders. For self-reported contributions, higher arsenic contamination is associated with higher predicted contribution rates. Other variables that correlate with income—savings, education, lending, and household size—also differ across groups with the highest and lowest predicted contribution rates. Results are broadly quite consistent across the two different measures of contributions.

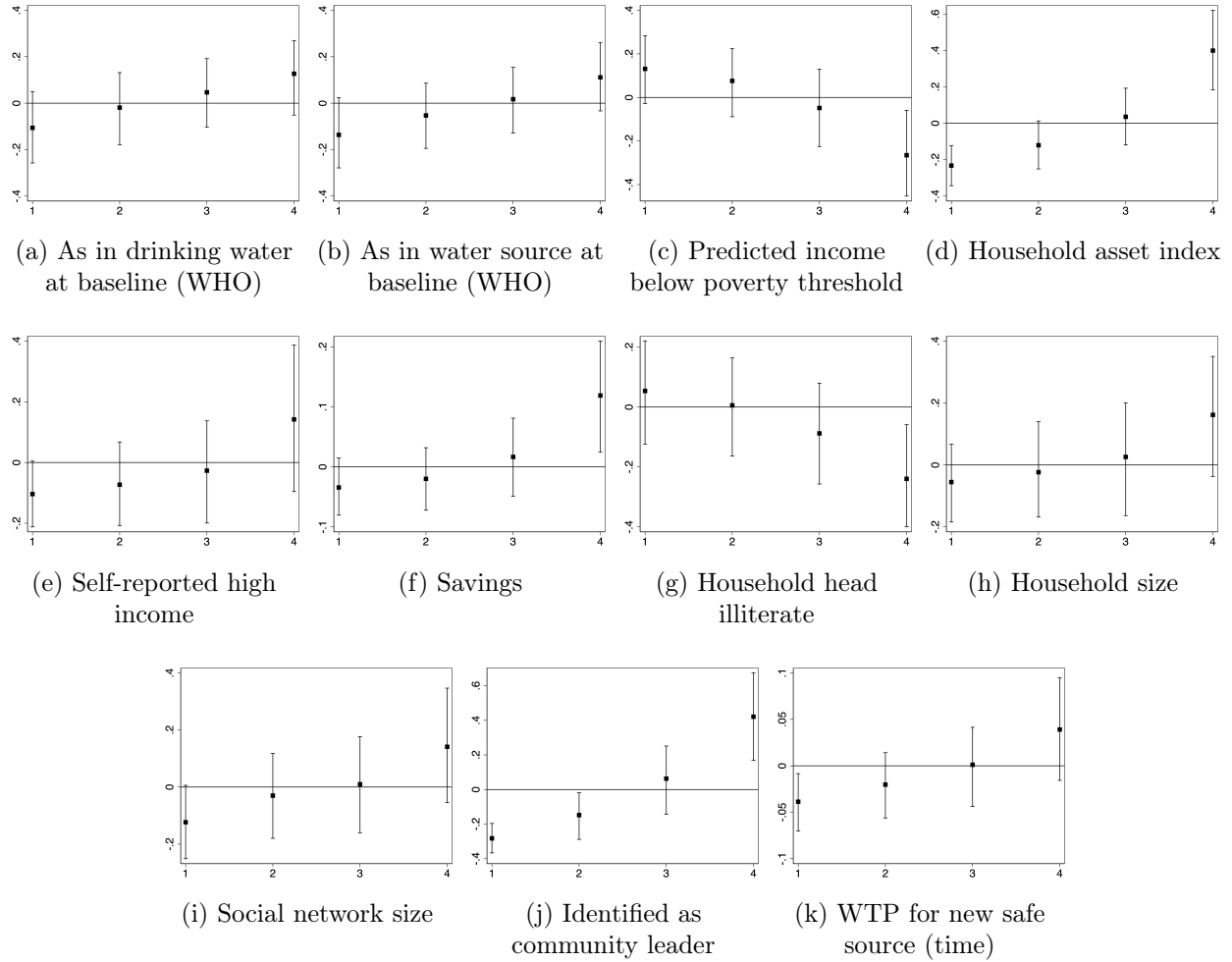
Appendix Figures L13 and L12 plot the characteristics that are correlated with the largest difference in predicted contribution rates under the cash and labour treatment arms. Again, we plot only variables for which the quartile with the highest predicted difference in contribution rates differs from the quartile with the lowest predicted difference in contribution rates with a p value of less than 0.1, and we repeat the exercise for both self-reported contributions and contributions based on project records. Measures of income—and variables correlated with income—are more strongly with predicted contributions under the cash contribution requirement than under the labour contribution requirement. Arsenic contamination rates are more strongly correlated with predicted contributions under the cash contribution requirement than under the labour contribution requirement.

Figure L10: CLAN, recorded contributions, pooled



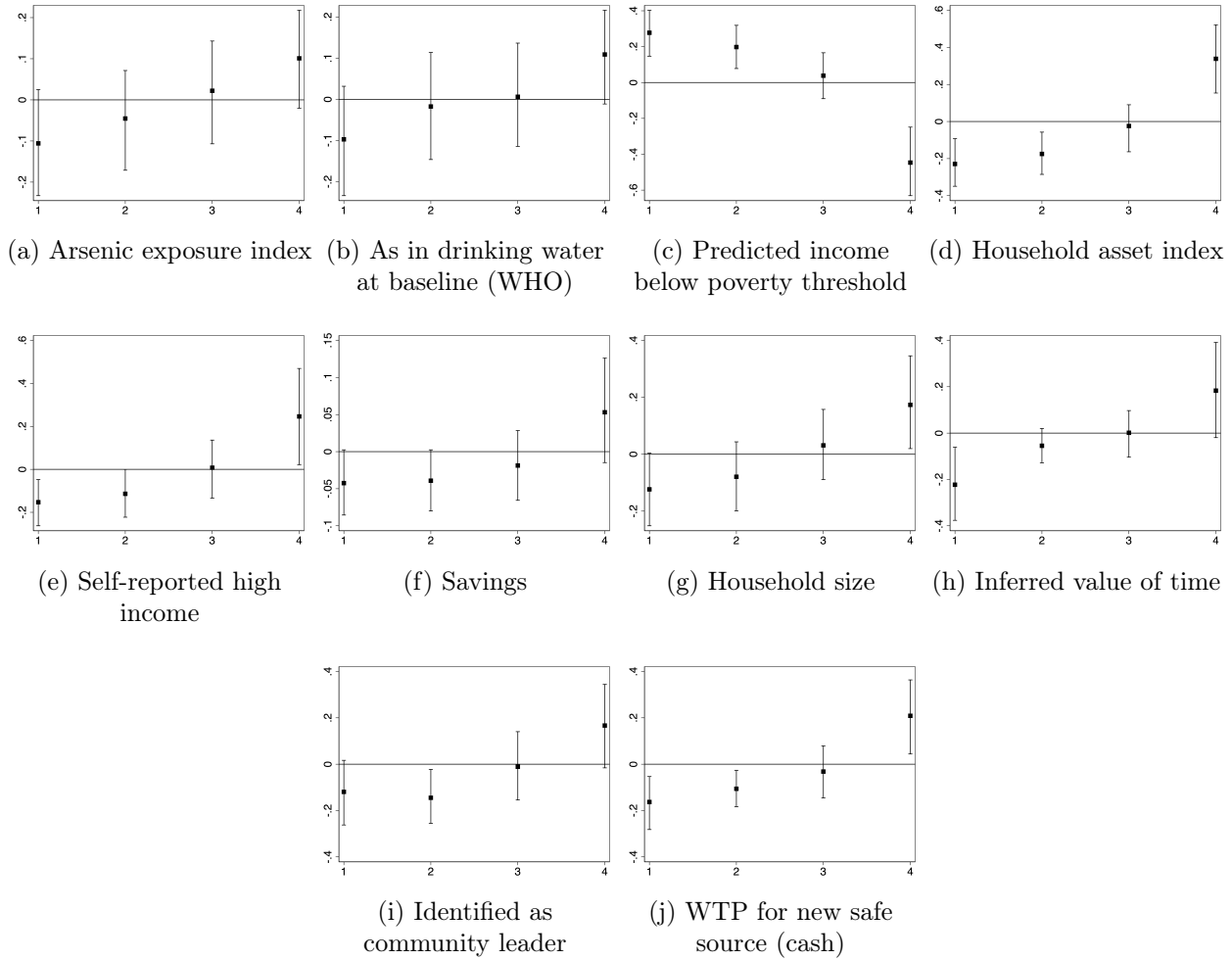
Notes Graphs plot the normalized value of the listed characteristic by quartile defined by heterogeneity score. x axis shows quartiles defined by predicted heterogeneity. y axis shows normalized value of listed characteristics. Only characteristics for which the top and bottom quartiles differ significantly ($p < 0.1$) are shown.

Figure L11: CLAN, self-reported contributions, pooled



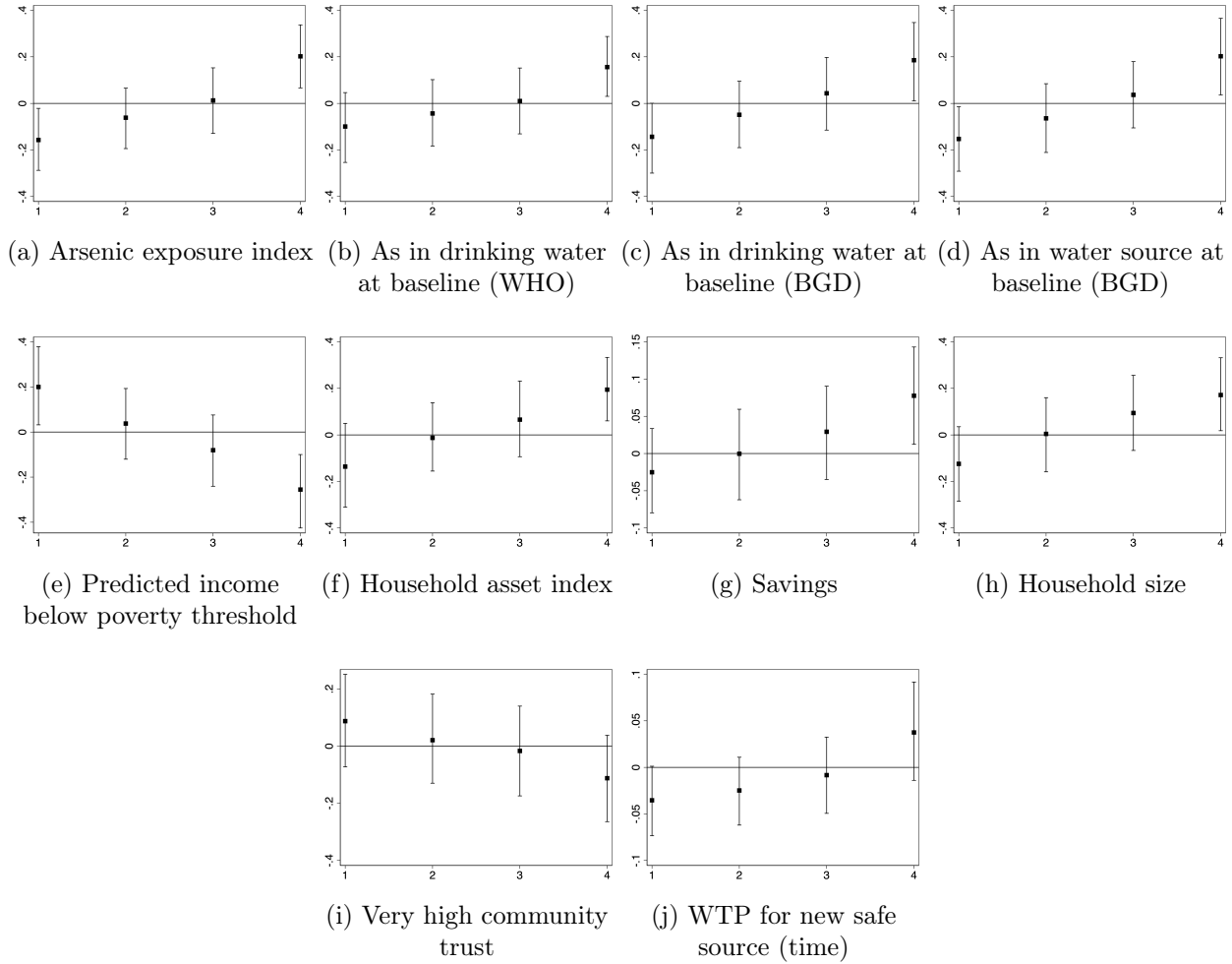
Notes Graphs plot the normalized value of the listed characteristic by quartile defined by heterogeneity score. x axis shows quartiles defined by predicted heterogeneity. y axis shows normalized value of listed characteristics. Only characteristics for which the top and bottom quartiles differ significantly ($p < 0.1$) are shown.

Figure L12: CLAN, recorded contributions, cash vs labour



Notes Graphs plot the normalized value of the listed characteristic by quartile defined by heterogeneity score. x axis shows quartiles defined by predicted heterogeneity. y axis shows normalized value of listed characteristics. Only characteristics for which the top and bottom quartiles differ significantly ($p < 0.1$) are shown.

Figure L13: CLAN, self-reported contributions, cash vs labour



Notes Graphs plot the normalized value of the listed characteristic by quartile defined by heterogeneity score. x axis shows quartiles defined by predicted heterogeneity. y axis shows normalized value of listed characteristics. Only characteristics for which the top and bottom quartiles differ significantly ($p < 0.1$) are shown.

M Screening

This section evaluates evidence for screening effects of community contribution requirements. Appendix Table M1 shows 2SLS estimates of the treatment effects on the treated, using assignment to treatment to predict fraction of wells successfully installed under each treatment arm. Compared to the reduced form comparisons across treatment arms, the additional assumption required for a causal interpretation of the estimated effects is that contributions treatments only affect impact through the share of offered wells installed.

For effects on arsenic exposure, the treatment effects on the treated for the cash contribution arm are positive in sign for the index and for three of the four index components, implying an increase in arsenic exposure rather than the intended reduction. However, all these estimates have considerable uncertainty, and confidence intervals do not exclude reductions in exposure. We cannot rule out the null of equal treatment effects on the treated under any pair of treatment arms for any of the arsenic exposure measures. Noisy measures of arsenic exposure compound the effects of small samples and the small number of compliers in the cash treatment arm, making it hard to draw definitive conclusions.

Effects on the more precisely-estimated measures of well use show a pattern of results that is consistent with some positive screening. Well use rates are about 20% higher in communities that successfully install wells in the cash treatment arm compared to communities that successfully installed wells under either of the other two treatment arms. However, we cannot rule out the null of equal treatment effects on the treated for any pair of treatment arms for either measure of well use, implying that the pattern could arise due to chance.

Table M1: Treatment effects on the treated

	Control Mean	Estimated impact			C = W	L = W	C = L	N
	(1)	Cash (2)	Labour (3)	Waiver (4)	(5)	(6)	(7)	(8)
a) Change in exposure to arsenic contamination								
Arsenic exposure index	-0.10 (0.03)	0.07 (0.21)	-0.11* (0.07)	-0.12* (0.06)	0.336	0.923	0.351	13646
Household drinking water (WHO)	-0.10 (0.02)	0.08 (0.15)	-0.05 (0.05)	-0.02 (0.03)	0.438	0.551	0.336	13588
Household drinking water (BGD)	-0.02 (0.02)	0.12 (0.11)	-0.01 (0.03)	-0.03 (0.03)	0.119	0.564	0.172	13588
Water source (WHO)	-0.07 (0.02)	-0.10 (0.17)	-0.11** (0.05)	-0.07** (0.04)	0.873	0.454	0.948	13588
Water source (BGD)	-0.01 (0.02)	0.02 (0.13)	-0.02 (0.04)	-0.08** (0.04)	0.371	0.148	0.670	13588
b) Use of project wells								
Share of households	0.00 ()	0.34*** (0.09)	0.27*** (0.03)	0.28*** (0.04)	0.535	0.862	0.466	7122
Share of water	0.00 ()	0.13*** (0.04)	0.11*** (0.01)	0.11*** (0.02)	0.627	0.952	0.603	7122

Notes Table reports 2SLS estimates of treatment effects on the treated. The sample is all randomly-sampled households in study communities. Clustered standard errors are in parentheses. For measures of arsenic contamination, regressions include baseline and endline data and community fixed effects. For measures of use, regression includes endline data only. Weights are applied so that each community counts equally in summary statistics. Significance levels obtained from analytical standard errors. p values from pairwise hypothesis tests in columns 5-7. Asterisks in columns 2-4 reflect significance of impacts with respect to the control group: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

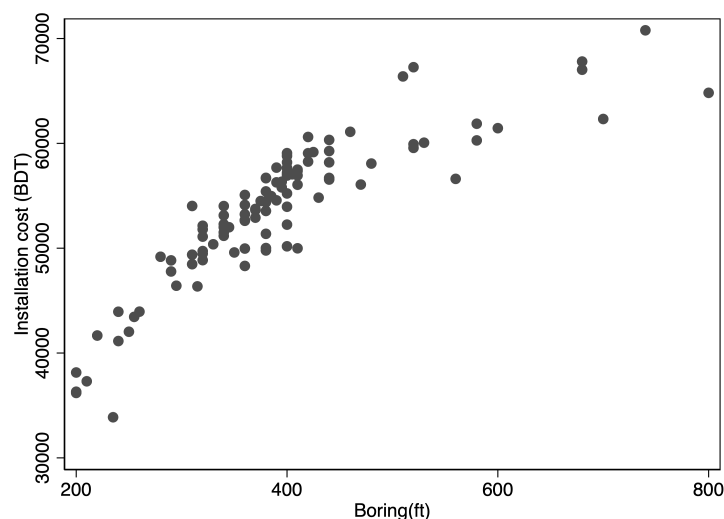
N Cost data

We assemble data from several sources to give a comprehensive picture of the costs of implementation. This appendix describes how we aggregate the cost data to the community level.

Well installation costs We obtain the exact payments to the contract for each installed well from project accounts. The price paid to the contractor for a successfully installed safe well is a function of the depth to which contractors had to drill to install the well and the depth of the installed well, as well as the parts required for installation, following a predetermined price schedule. Figure N1 illustrates.

Under the cash contribution arm, 6000 BDT was paid in cash by the community, so the cost to the NGO was 6000 BDT lower than the total amount transferred to the contractor. Under the labour contribution arm, the contractor rebated 3000 BDT to the NGO after installation, so the cost to the NGO was 3000 BDT lower than the price would have been under the contribution waiver.

Figure N1: Price paid to contractor for well installation



Notes Each point is a successfully-installed well. Data on costs and depth from project records

Refunds Field staff also recorded cases when they refunded cash contributions to communities if wells could not be installed (2 wells) and when they compensated households for time expended providing labour contributions for wells that could not finally be installed (7 wells).

Field staff time use Field staff kept extensive records of their activities. We classify these into ten categories, shown in order of decreasing frequency in Table N1. Staff reported the date of each activity and the time spent in minutes on each activity. In the one community which declined to organize a meeting, field staff did not complete an activity report. We assume that staff spent two days attempting to organize a meeting in this community.

Table N1: Field staff activities

Activity	% of recorded activities
Monitoring installation	25%
Meeting preparation	13%
Caretaker selection	13%
Post-installation monitoring	13%
Pre-installation coordination	13%
General meetings	11%
Small group meetings	11%

Notes Table reports percentage of recorded activities coded in each category, unweighted by time spent on the activity. Field staff reported on average 4 activities per work day.

We convert the time data to a share of each work day. We estimate the total time taken for each activity by adding 30 minutes of travel time to the recorded time the activity took. We sum the total amount of minutes recorded on each day. We then assign each activity a share of the total time recorded in activities for that day. The unit of this measure is the “team-day”: if the team completed only one activity on a day, we consider that the team spent one team-day on that activity in that community. An implicit assumption is that either the team worked separately for the whole day or together for the whole day. If the team worked, e.g., separately for the morning and together for the afternoon, our approach will overweight the time spent separately. Not all time is spent in the field, and field staff did not keep detailed records of which projects they worked on in the office. Our approach makes what we think is a fairly reasonable assumption that time in the office was spent in proportion to time in the field in each community. For all days in which staff worked in the field, we are thus able to assign a share of their time to one of the activities listed above and to a specific community. We can also separate out time spent on post-installation monitoring, which we considered a research activity rather than an implementation activity. Our partner NGO would not normally undertake as extensive post-installation monitoring as we did under this project. Appendix Figures N2a and N2b summarize staff time costs per community and per successfully installed well, respectively, by activity.

Other NGO project costs We use comprehensive project accounts data to estimate all other costs. We first exclude costs that we account to subprojects that we ran in parallel to this study (Cocciolo, 2019; Cocciolo, Habib and Tompsett, 2019; Ghisolfi, 2020). We then classify costs as belong to seven categories: overheads (including office costs), wages, survey costs (including enumerator wages and tablets for data entry), water quality testing, well installation costs, post-well installation costs, and local travel. Consistent with NGO Forum’s policy, we apply a 7% indirect cost rate to all expenses.

We sum up all additional costs associated with well installation attempts. The total amount recorded on the budget associated with installation is 29% larger than the total amount paid to the contractor, because additional costs are accounted to this line item, such as travel costs or bank fees. We count this additional cost (29% times the costs paid directly

to the contractor) in this category. We also include the field engineer's salary under this category, because the field engineer attended both successful and unsuccessful installations.

We also sum up all additional costs associated with tubewell construction that apply to installed wells: the costs of training caretakers and providing them with tools and the costs of laboratory tests conducted to confirm water quality after installation.

We sum up all other costs directly associated with implementation that scale on a per-community basis, including printing of maps and brochures and the costs of holding meetings.

The staff costs, overheads, and local travel recorded on the budget cover periods when the project was implemented and periods when we were collecting baseline or endline data. For each project year, we count the number of days on which we collected data and the number of days for which our field staff reported any field activity. We calculate the total staff costs, travel, and overheads associated with implementation proportional to the share of days for which we recorded outputs in a year that field staff reported implementation activities. (i.e., $\text{number of team-days on which team recorded implementation activities} / \text{number of days on which any activity was recorded}$).

Assembling project cost data at the community level We estimate total implementation costs per community, as follows:

- The costs paid to the contractor for wells installed in the community are assigned to that community.
- Implementation costs that scale per community are divided equally between communities.
- Additional installation costs that scale with attempted wells are assigned proportionally to the number of attempted installations in each community.
- Additional installation costs that scale with installed wells are assigned proportionally to the number of successfully installed wells in each community.
- We then assign staff costs, overheads, and local travel costs proportionally to the share of time spent in each community recorded in the activity reports.

Figure N2a: Field staff time per community

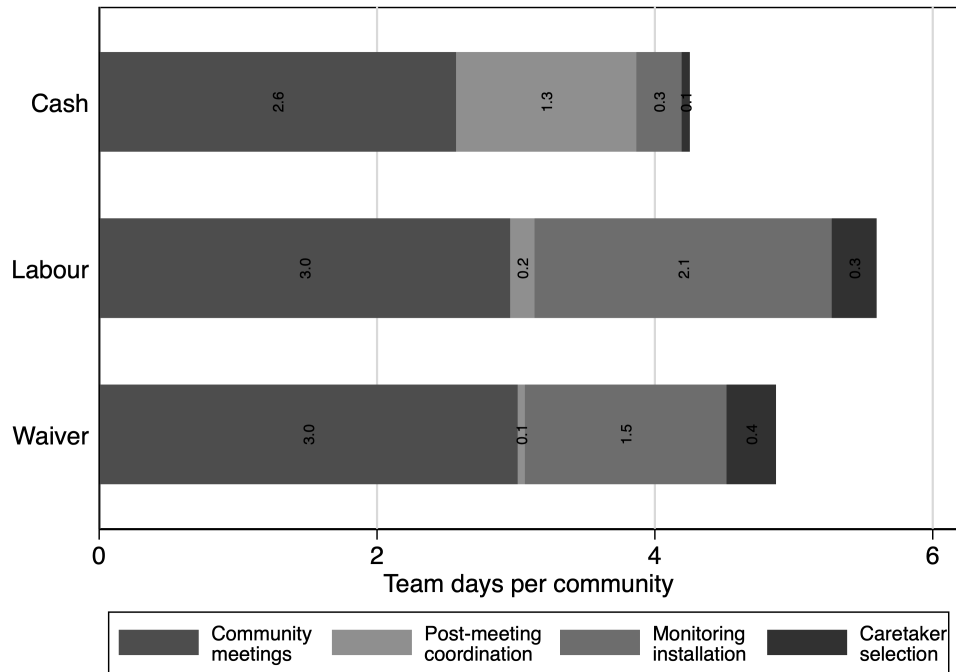
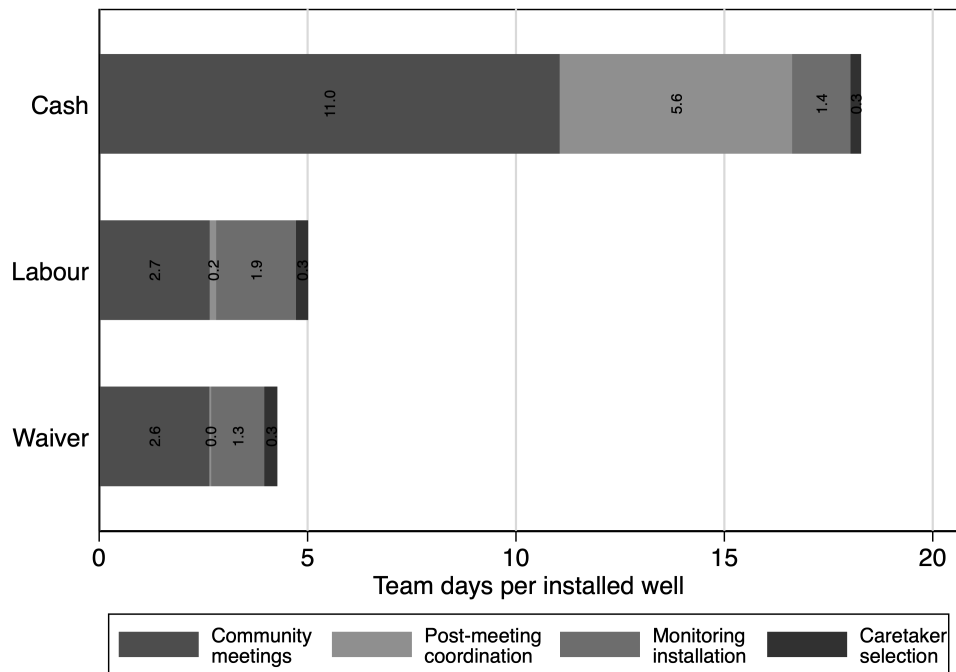


Figure N2b: Field staff time per successfully installed well

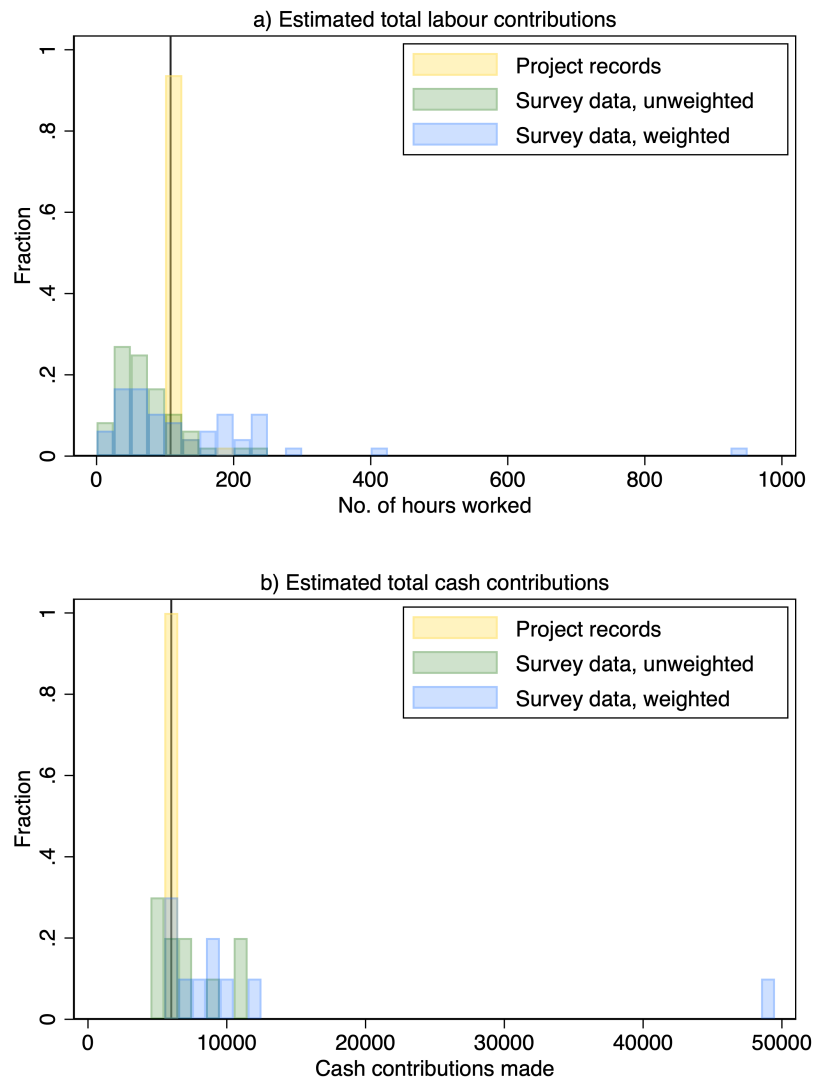


Community costs We use project records and self-reported data to estimate community costs in three categories.

1. We record cash contributions at the community level from project records and from survey data. Our project records state that a total of 6000 BDT were paid for each tubewell. When we use survey data, we weight household data by the inverse of the probability of inclusion in the sample. Sampling introduces error into our estimate of the total contributions, which in some cases can be extreme, i.e., when a household from the random sample reports paying the entire contribution.
2. We record the total amount of labour contributions using project records and survey data. The project records state that a minimum of 108 hours of labour were provided by the community for each successfully installed well (6 shifts $\times$ 6 hours $\times$ 3 workers). In a few cases, project records state that communities continued to provide labour if an installation process stretched into a fourth or fifth day. Similarly to the cash contribution data, the estimates of total contributions from survey data are affected by sampling error. To translate these data into monetary terms for the cost-effectiveness comparison, we compare several different approaches to valuation: i) valuing at the market wage; ii) valuing at the reservation wage as reported by the household at baseline, assigning contributors the median reservation wage among other contributors in their community if they were not surveyed at baseline; iii) valuing at the household's inferred valuation of time; and iv) asking households that reported providing labour contributions to report how much they would have been paid for working the number of hours they reported contributing. Appendix Figure N4 shows the distribution of valuations for the labour contributions provided for each constructed tubewell. In general, valuations that use the inferred value of time are smaller than valuations that use the reservation wage, the unskilled daily wage, or the self-reported value of the labour contributions provided.
3. We estimate the amount of time communities spend on non-contribution costs using self-reported survey data. Households listed the number of household members who assisted the small group and general meetings. We assume 2 hours of time per household member who attended the general meeting and 1 hour of time per household member who attended the small group meeting. A limitation of our data is that households did not separately report which household members attended the small group meetings. Thus we assume that all household members who attended the general meeting also attended the small group meetings, if the household reports that household members attended both meetings. In most cases, only one household member attended the general meeting. We also sum the amount of hours that households who reported helping to coordinate the project reported spending.

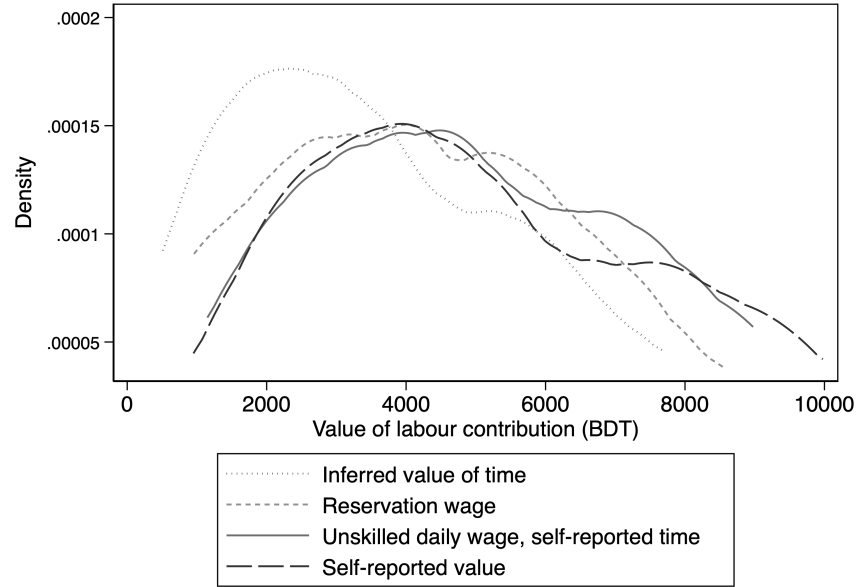
Cost-effectiveness We sum all costs and divide by the number of wells constructed in each treatment arm to estimate cost per successfully-installed well. Appendix Table N2 summarizes.

Figure N3: Total contributions per successfully installed well: project records vs self-reports



Notes Sample includes successfully installed wells. Weights account for sampling probabilities. Vertical black line shows required contribution.

Figure N4: Valuation of labour contribution



Notes Sample includes successfully installed wells. Extreme right hand side outliers not shown.

Table N2: Relative cost-effectiveness

	Cost-effectiveness relative to waiver	
	Cash	Labour
Installation costs only	0.92	0.92
NGO monetary costs only	2.43	1.03
NGO costs and community contributions	2.47	1.07
NGO costs, community contributions and community time costs	2.63	1.10

Notes Table reports cost per successfully installed well under the listed treatment arm, scaled by cost per successfully installed well under the waiver treatment arm. A value of 1 implies equal costs per successfully installed well; lower values imply higher cost-effectiveness and vice versa.

References

- Chernozhukov, Victor, Ivan Fernández-Val, and Alfred Galichon.** 2009. “Improving Point and Interval Estimates of Monotone Functions by Rearrangement.” *Biometrika*, 36(3).
- Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val.** 2025. “Fischer-Schultz Lecture: Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments, with an application to immunization in India.” *Econometrica*, 93(4).
- Cocciolo, Serena.** 2019. “Experience of Inclusive Institutions and the Value of Participation: Experimental Evidence from Bangladesh.” Working Paper.
- Cocciolo, Serena, Md. Ahasan Habib, and Anna Tompsett.** 2019. “Understanding institutional persistence: Exposure to democratic decision-making and the value of autonomy and democracy.” Pre-registered report and pre-analysis plan.
- Cocciolo, Serena, Selene Ghisolfi, Md. Ahasan Habib, SMA Rashid, and Anna Tompsett.** 2021. “Do community water sources provide safe drinking water? Evidence from a randomized experiment in rural Bangladesh.” *World Bank Economic Review*, 35(4).
- Ghisolfi, Selene.** 2020. “Contribution Requirements and Redistribution Decisions: Experimental Evidence from Bangladesh.” Working paper.
- Gibbons, Charles E., Juan Carlos Suárez Serrato, and Michael B. Urbancic.** 2019. “Broken or Fixed Effects?” *Journal of Econometric Methods*, 8(1).
- GIEWS.** 2021. “Global Information and Early Warning System. Country Brief: Bangladesh.”
- Imbens, Guido W., and Donald B. Rubin.** 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.
- Kremer, Michael, Jessica Leino, Edward Miguel, and Alix Peterson Zwane.** 2011. “Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions.” *Quarterly Journal of Economics*, 126(1).
- Lin, W.** 2013. “Agnostic notes on regression adjustments to experimental data: Reexamining Freedman’s critique.” *Annals of Applied Statistics*, 7(1).