Center for Environmental and Resource Economic Policy
College of Agriculture and Life Sciences
https://cenrep.ncsu.edu

Scalable Targeting of Social Protection:

When Do Algorithms Out-Perform Surveys and Community Knowledge?

Emily Aiken, Anik Ashraf, Joshua E. Blumenstock, Raymond P. Guiteras and Ahmed Mushfiq Mobarak

Center for Environmental and Resource Economic Policy
Working Paper Series: No. 25-001
November 2025

Suggested citation: Aiken, Emily, Anik Ashraf, Joshua E. Blumenstock, Raymond P. Guiteras and Ahmed Mushfiq Mobarak, "Scalable Targeting of Social Protection: When Do Algorithms Out-Perform Surveys and Community Knowledge?" CEnREP Working Paper No. 25-001, November 2025, https://go.ncsu.edu/cenrep.wp.25.001.

Scalable Targeting of Social Protection: When Do Algorithms Out-Perform Surveys and Community Knowledge?*

Emily Aiken Anik Ashraf Joshua E. Blumenstock
Raymond P. Guiteras Ahmed Mushfiq Mobarak

November 2025

Abstract

Advances in digital data and algorithms are enabling new approaches to poverty targeting at scale. Using rich data from Bangladesh and Togo, we compare an algorithmic approach based on machine learning and mobile phone data to status quo targeting with proxy means tests and community-based targeting. While proxy means tests are most accurate, algorithmic targeting is more cost effective for programs where the budget is small relative to the number of households screened. Combining our estimates with global program data, we estimate that phone-based targeting would be the welfare-maximizing approach for up to 30% of countries' social assistance programs.

JEL Codes: C55, I32, I38

*Aiken: eaiken@ucsd.edu; Ashraf: anik.ashraf@econ.lmu.de; Blumenstock: jblumenstock@berkeley.edu; Guiteras: rpguiter@ncsu.edu; Mobarak: ahmed.mobarak@yale.edu. IRB approval was obtained from the U.C. Berkeley Committee for the Protection of Human Subjects (Protocol #2023-02-16103) and the Institutional Review Board of the Institute of Health Economics at the U. of Dhaka. We are grateful to Soumi Chandra and Leo Selker for excellent research assistance. We received helpful comments from seminar participants at CUNY – Hunter College, Columbia U., DevSouth, PacDev, the Triangle Applied Microeconomics Conference, Yale SOM, GRIPS (Tokyo), the U. Chicago AI in Social Science conference, Paris School of Economics, CIRED, and U. Washington. We thank our funders (GiveDirectly, the Global Innovation Fund, and the National Science Foundation under CAREER Grant IIS-1942702),implementing partners (GiveDirectly and a2i), BIGD for data support, and BRAC for sharing their community-based targeting protocol.

1 Introduction

Hundreds of billions of dollars are spent on social protection programs and humanitarian aid each year (ILO, 2021), yet accurately identifying poor households remains a central challenge (Hanna and Olken, 2018). When targeting fails, scarce resources are misallocated: poor households are excluded from benefits while non-poor households are included. In many low- and middle-income countries, such errors are widespread: Coady et al. (2004) find that a quarter of poverty-targeted programs in low-income countries are regressive, directing more benefits to rich households than to poorer ones.

Most programs continue to rely on traditional, in-person data collection methods — such as survey-based eligibility verification and community selection of beneficiaries — to determine eligibility. Recently, however, the spread of digital data and advances in machine learning have enabled new ways to identify beneficiaries remotely, potentially reducing the cost and time required for targeting (Aiken et al., 2022; Mukerjee et al., 2023; Lopez, 2020; Smythe and Blumenstock, 2022; GiveDirectly, 2022). Approaches that use mobile phone metadata or satellite imagery are particularly attractive because digital data can be collected at a fraction of the cost of in-person visits and scale more easily when large populations must be screened.

This paper compares three paradigms for identifying beneficiaries of a cash transfer program in southern Bangladesh, both in terms of accuracy and implementation cost, to assess when it is preferable to target beneficiaries using digital data and machine learning rather than traditional methods. To our knowledge, this is the first direct comparison of proxy means testing (PMT), community-based targeting (CBT) and phone-based targeting. To make this comparison, we collected four complementary datasets: a census of 106,000 households across 201 villages; detailed consumption surveys for 5,000 households; community wealth rankings from CBT exercises in 180 neighborhoods; and mobile phone records from all four major telecom operators covering all consenting households. In partnership with GiveDirectly, we used these approaches to deliver cash transfers of 30,000 Bangladeshi Taka (roughly \$300 USD, or \$955 at PPP) to 22,000 households. We also documented the costs of administering each method, enabling us to evaluate not only targeting accuracy but also cost-effectiveness within a welfare framework. To understand the generalizability of our results, we replicate the analysis with nationally representative survey and phone data from Togo, and relate our results to cross-country evidence on social protection budgets from the World Bank's ASPIRE database.

Our first main finding — focusing on accuracy alone — shows that algorithmic targeting with mobile phone data more accurately identifies consumption-poor households than community-based targeting, but both methods are substantially less accurate than proxy means testing. Different methods also prioritize different types of beneficiaries. For instance, CBT is more likely to select widows and widowers, echoing results from other settings (e.g., Sumarto et al., 2025), while phone-based targeting and PMTs prioritize households that devote a larger share of expenditures to food. Other survey-based targeting approaches — including the Poverty Probability Index (PPI, Kshirsagar et al., 2017) and a decentralized peer rankings approach where respondents privately rank peers' poverty status — are comparable in accuracy to phone-based targeting.[1]

Our second main finding concerns the tradeoff between accuracy and cost: we show that the welfare-maximizing targeting method depends on program scale. While PMT is the most accurate method, it is also more expensive to implement than algorithmic targeting, especially for national-scale programs that must screen large populations. We adapt the welfare framework of Hanna and Olken (2018) to account for targeting costs, and compare the simulated welfare effects of CBT, PMT and phone-based approaches in both Bangladesh and Togo. The results show that when the total program budget is small relative to the number of households screened, phone-based targeting is most cost-effective. Intuitively, when survey-based screening would consume a large share of program resources, the cost savings from algorithmic methods outweigh the loss in accuracy.

This paper contributes to three strands of research. First, a large literature evaluates the accuracy of alternative approaches to targeting social protection. We find that PMTs are generally more accurate than CBTs in identifying the consumption-poor in Bangladesh, which is consistent with prior results elsewhere (Alatas et al., 2012; Basurto et al., 2020; Premand and Schnitzer, 2021; Schnitzer and Stoeffler, 2022; Trachtman et al., 2022; Sumarto et al., 2025). Related studies have evaluated other approaches to identifying poor households, including geographic targeting (Baker and Grosh, 1994), "scorecard" approaches like the Poverty Probability Index (Kshirsagar et al., 2017), decentralized community-based targeting based on peer rankings (Alatas et al., 2016; Beaman et al., 2021; Trachtman et al., 2022), and random targeting via lotteries (Bance and Schnitzer, 2021). We extend this work by providing direct, head-to-head accuracy comparisons across all these approaches and by incorporating phone-based targeting into the comparison set. This addition is important because the rapid expansion of mobile phones in otherwise data-poor regions, combined with advances in computing and machine learning, makes phone records a promising tool for cost-effective targeting of large-scale social protection programs.

[1] Our main analysis assesses targeting accuracy by benchmarking against consumption poverty, as measured through intensive consumption expenditures surveys. Since we only collect consumption, PMT, and CBT data at a single point in time, our analysis does not consider the extent to which different methods remain accurate over time (Hillebrecht et al., 2023; Brown et al., 2018; Aiken et al., 2025).

Second, our paper contributes to the emerging literature on the use of novel digital data sources for targeting (Aiken et al., 2022, 2023b; Smythe and Blumenstock, 2022). Two prior studies in Togo and Afghanistan develop the foundational methods underlying the phone-based targeting approach we deploy in Bangladesh. This paper goes further by setting up the first direct comparison between a highly-centralized, top-down algorithmic method and the increasingly popular decentralized, bottom-up community-based approach (Alatas et al., 2012; Sumarto et al., 2025). We also extend prior work by moving beyond accuracy comparisons to identify the circumstances under which phone-based targeting is most efficient.[2]

Finally, this paper contributes to a nascent literature on the cost-effective administration of social protection and humanitarian aid programs in low-income countries. While development programs are often evaluated using some metric of cost-effectiveness (e.g., Murray et al., 2000; Dutrey, 2007; Devereux et al., 2017), there is little systematic evidence on the tradeoff between cost and accuracy across alternative targeting approaches. Two prior studies by Houssou and Zeller (2011) and Hanna and Olken (2018) measure cost-effectiveness of targeting relative to universal distribution. Our distinctive contribution is to extend this framework by providing head-to-head comparisons of the cost-effectiveness of multiple popular targeting approaches across different program scales.

2 Data and Methods

2.1 Setting and Data Collection

The primary context for our analysis is a cash transfer program we developed in partnership with GiveDirectly and the Government of Bangladesh in 2023. The program provided cash transfers of 30,000 BDT (955 USD PPP) to 22,000 households in three sub-districts of Cox's Bazar district in southern Bangladesh — Ramu, Teknaf, and Ukhia.[3] The cash transfer program was designed to target the poorest 21% of households within the program area. Our main analysis compares the accuracy and cost-effectiveness of proxy means testing (PMT), community-based targeting (CBT), and phone-based targeting in this setting. A timeline of the project is provided in Figure S1. In supplementary analysis, we use data from a cash transfer program run by GiveDirectly and the government of Togo in 2021, which we describe later.

[2] Also related are papers that show how poverty can be estimated using non-traditional data such as satellite imagery (Jean et al., 2016; Yeh et al., 2020), internet data (Fatehkia et al., 2020), mobile phone records (Blumenstock et al., 2015; Blumenstock, 2018), and administrative records from financial services companies (Engelmann et al., 2018).

[3] GiveDirectly's objective was to support poor communities such as these, which host large numbers of Rohingya refugees.

Our analysis of targeting in southern Bangladesh relies on four main sources of data:

- A **census** of all households in 201 randomly chosen villages from the three study sub-districts in Bangladesh. This accounts for roughly two thirds of the households and villages in these sub-districts.[4] The census was conducted in February and March 2023. We collected phone numbers of all adult household members, as well as basic information about household characteristics and asset ownership necessary in order to compute the Poverty Probability Index (PPI).[5] The census collected information for approximately 106,000 households. This census was also used by GiveDirectly to register potential beneficiaries for their cash transfer program.

- A March 2023 **household survey**, which collected data on consumption, expenditures, demographics and assets. In this survey, we adopted the standardized consumption module from the 2016 Household Income and Expenditures Survey (HIES) implemented by the Bangladesh Bureau of Statistics. Following the instructions published by the Bangladesh Bureau of Statistics (Ahmed et al., 2019), we use these data to construct a measure of per capita **household consumption expenditures**.[6] The household survey was conducted with a representative random sample of 5,006 households from 180 neighborhoods in the study area. Neighborhoods were selected randomly from among the 890 neighborhoods enumerated in the census, stratified by upazila, neighborhood size (by size tercile), and the share of households in the neighborhood that were a religious or ethnic minority (no minority households; <10% minority households; ≥10% minority households). Descriptive measures and summary statistics from the household survey are provided in Figures S2 and S3 and Table S1. We also included a **peer rankings** module in the household survey, based on the mechanism of Bloch and Olckers (2022). In this module, we asked each household about eight randomly selected households in their neighborhood. They were asked to report how well they knew the household and to assess the material well-being of each household both in absolute terms and relative to the seven other households on their list (details in Appendix A.7).

- Household wealth rankings from **community-based targeting exercises** conducted in November 2023 in each of the 180 neighborhoods. Our CBT exercises assembled 12-25

[4]Based on the official 2011 census, we estimate that our census covered 65% of households and 63% of villages.

[5]The PPI for Bangladesh is available at https://www.povertyindex.org/country/bangladesh. See Kshirsagar et al. (2017) for PPI methodology and assessment in Zambia.

[6]We use per capita household consumption expenditures, rather than an adult equivalence scale, following the standard consumption aggregate calculation by the Bangladesh Bureau of Statistics (Ahmed et al., 2019). We later show that our main results hold when an adult equivalence scale is used instead.

community members from all walks of life from each neighborhood to collectively identify the 20% households with the lowest socioeconomic status, who would later receive a one-time cash transfer of 1,100 Taka ($35 USD PPP). We adopted a protocol regularly implemented by BRAC to determine beneficiaries for their own social safety net programs. This protocol is described in detail in Appendix A.2.

- Complete **mobile phone metadata** from all consenting survey respondents from March to July 2023, including records of calls, texts, and mobile data usage. These data were obtained from all four mobile network operators active in the survey region. Following the data protection procedures described in our IRB protocol, we removed all personally identifying information, including phone numbers, prior to analyzing mobile phone metadata. Details on these protocols, and special considerations regarding data privacy and ethics, are discussed in Appendix A.8.

2.2 Targeting Methods

We use these data to assess several approaches to targeting social protections in the context of southern Bangladesh. Our main results focus on three targeting methods:

1. A **phone-based targeting** approach that uses machine learning to predict household per capita consumption expenditures, based on 1,578 statistics computed from that subscribers' mobile phone records (including information about calls, texts, contact diversity, mobility, and mobile data usage). Our machine learning methods are similar to those used in past work (Aiken et al., 2022, 2023b,a) and are detailed in Appendix A.1. In short, we first obtain pseudonymized mobile phone records from all four mobile network operators active in Cox's Bazar, for all phone numbers from all consenting surveyed households. These data include metadata (including pseudonymized identifiers for the caller and recipient, date, time, and duration of calls, and GPS coordinates for cell towers used) for all incoming and outgoing calls and SMS messages placed between March 1 and July 31, 2023, as well as information on daily mobile data usage. From these data, we calculated 1,578 *features* describing mobile phone use for each pseudonymized phone number in the dataset,[7] including statistics on call and text frequency, heterogeneity in contact networks, recharge patterns, mobility traces based on cell tower usage, and more. Finally, we matched mobile phone features to the household survey (for the 94% of households that provided at least one phone number that was present in the mobile

[7] Subscriber-level statistics on mobile phone use are calculated using the open source python library cider.

phone records), and used the matched dataset to train a gradient boosting model[8] to predict log per-capita consumption using mobile phone features. Table S2 shows the phone features that turn out to be the most predictive of consumption in our Bangladesh data.

2. The **community-based targeting (CBT)** rankings from each community are used directly to identify the households with the lowest socioeconomic status in each neighborhood (Appendix A.2). Rankings are normalized within each neighborhood to be between 0 and 1 to facilitate comparisons across communities.

3. The **proxy means test (PMT)** estimates household consumption expenditures using verifiable assets and household characteristics. In our household survey, we collected information on 45 covariates that are common to many PMTs (Hanna and Olken, 2018; Brown et al., 2018), including household characteristics (for example, the number of rooms and the material of the roof), demographic information (e.g., the household size and gender of the household head), and asset ownership. We use modern machine learning methods to develop a PMT that predicts log per-capita consumption expenditures from those 45 covariates (see Appendix A.3). We expect that this represents a best-case PMT, since many real-world PMTs take a more ad hoc approach to fitting the prediction rule (McBride and Nichols, 2018; Noriega-Campero et al., 2020).[9] Figure S4 lists the variables with the largest coefficients in our Bangladesh data.

In addition to the three main approaches described above, we include some results based on less common approaches to targeting:

4. **Geographic targeting** at the union (admin-5) level, based on the population-weighted wealth estimates of CIESIN (2021), which combines sub-national administrative data and gridded earth observation data to produce the Global Deprivation Index (GDI),an

[8] Gradient boosting is a nonparametric ensemble machine learning algorithm. The ensemble consists of a number of decision trees, each of which is trained to predict household poverty from the phone data features, and includes explicit regularization. The final poverty prediction for each household is an average of the predictions from each decision tree. In addition to the gradient boosting approach, we also tested other machine learning models (including linear regression, LASSO regression, and a random forest); we focus on the gradient boosting model because it has the highest overall accuracy.

[9] Using cross-validation, we evaluated several approaches to constructing a PMT, including simple linear regression, linear regression with step-wise forward selection, LASSO regression, and a random forest algorithm. When evaluated out-of-sample, the LASSO regression was most accurate, so our main results focus on the LASSO PMT, where the L1 penalty is selected via cross-validation. Appendix A.3 presents results for other PMT variants. The PMT we use as a benchmark is the output of the optimally regularized LASSO, without applying any constraints on included variables (e.g., to eliminate variables that could be easily gamed, or to reduce overall survey length). The performance of a real-world LASSO, under such constraints, would likely be slightly inferior to our benchmark PMT.

index of relative deprivation.[10] We aggregate the Global Deprivation Index at the union level, weighting by population using remotely sensed population data from Tiecke et al. (2017).

5. Other survey-based targeting approaches similar to the PMT, including Bangladesh's **poverty probability index (PPI)** and an **asset index** constructed with principal components analysis. The PPI is a scorecard poverty method based on 10 questions, including district, household members, children under ten, the highest grade completed by anyone in the household, ownership of a bicycle, refrigerator, and fan, construction material of household walls, electricity connection, and type of toilet used. The PPI scorecard was calibrated by Innovations for Poverty Action using the nationally representative 2016-17 Household Income and Expenditures Survey. Our asset index is constructed following Filmer and Pritchett (2001), using weighted principal components analysis to obtain a vector representing the direction of maximum variation in asset ownership among the 26 assets collected in our survey. In our setting, the first principal component explain on average 18% of the total variation in asset ownership.

6. **Peer rankings**, based on taking the average of the wealth ratings elicited in the household survey for a given household by their neighbors (see Appendix A.7). This is similar to the CBT in that it seeks to elicit how neighbors perceive each others' relative standing, but the information is solicited from households individually and privately rather than through the collective and public process of the CBT. We ask households to also rate themselves, so the peer ranking module also produces a "self-targeting" outcome. Unlike the CBT, these peer rankings were not incentivized and survey subjects were not told that their rankings would affect real transfers.

Data Ethics and Privacy Appendix A.8 provides details on the protocols we followed to minimize the risk of mis-use of call detail records (CDR). To summarize: we received permission from the Bangladesh Telecom Regulatory Commission to access CDR from the four major telecom operators in the country. We also secured informed consent from survey participants before accessing their CDR. To minimize the risk of data leaks and unauthorized use, the research team provided the telecom operator staff a set of phone numbers from the subset of surveyed households who granted us consent, along with code that would allow the telecom staff to extract the 1,578 features from the CDR data. Our research team never accessed the raw CDR. The telecom staff were responsible for merging a redacted

[10]The components of the gridded GDI include the child dependency ratio, infant mortality rates, the subnational human development index, the remotely sensed ratio of built-up to non-built up area, nighttime lights intensity, and changes in nighttime lights intensity from 2012 to 2020.

version of the household surveys with the CDR features; this dataset was anonymized and securely stored on an isolated server on the premises of a2i, an entity of the Government of Bangladesh. This multi-step data handling protocol was designed to ensure that the research team, GiveDirectly, and the Bangladesh government never accessed the CDR with personal identifiable information (PII), and that the telecom operators never accessed the unencrypted household survey data.

3 Accuracy of targeting methods

Our first set of results assesses how accurately each targeting method can identify the poorest households. This analysis uses data from the household survey (N=5,006) to compare each household's true consumption expenditures (as measured in the survey) to the score or ranking assigned to that household by the various targeting methods described above in Section 2.[11]

Data from a randomly selected 75% of surveyed households are used to train targeting methods that require machine learning (i.e., phone-based targeting and PMT), and performance is evaluated on the held-out 25%. We repeat this process 100 times on different random train-test splits, and report the mean and standard deviation of each performance metric across the 100 runs.[12] To illustrate, Figure 1 shows scatterplots from one train-test split of the rankings under each method vs. per-capita consumption expenditure as measured in the household survey. In the figure, it can be seen that the PMT scores (center) are most correlated with true consumption, followed by phone-based targeting (top) and then CBT (bottom).

Accuracy metrics We use three standard metrics for assessing targeting methods. The first and most intuitive is *recall*: the probability that a truly poor household is correctly classified as poor by the targeting method.[13] This is the simplest metric, but considers only

[11] We use per capita household consumption expenditures, rather than an adult equivalence scale, following the standard consumption aggregate calculation by the Bangladesh Bureau of Statistics (Ahmed et al., 2019). Figure S5 shows that our main results hold when an adult equivalence scale is used instead.

[12] Some of the targeting methods we simulate do not produce poverty rankings for all households. For instance, in the phone-based targeting approach, 6% of households cannot be assigned a ranking (2% of households in the survey did not provide a phone number or did not consent to matching survey data to mobile phone records; 4% of households provided a number that could not be matched to the data provided by the mobile telcos). 0.4% of households were not ranked in the CBT exercises and 2% of households had no peer rankings because they were not known to the community. In such cases, households that are unranked are targeted last in our targeting simulations — that is, we assume that any unranked household will be last in line to receive a transfer.

[13] Recall, also known as *sensitivity*, is equal to one minus the exclusion (type II) error rate. Since the program in Bangladesh provided transfers to 21% of households, as determined by the program's total budget,

Figure 1: Predictions vs. survey-based household PCE, by method

Notes: These scatterplots show the score or ranking from the targeting method (y-axis) against the true per-capita expenditures, as measured in the household survey (x-axis). Each dot represents a household from the survey. Results are produced from a single train-test split.

binary errors, not the magnitude of error, and depends on the fraction of households targeted by a particular program. The second metric is the *Spearman rank correlation* between the rank assigned to a household by a particular method and the household's true rank in the distribution of consumption per capita. This metric penalizes large errors in the relative ranking of households, and does not depend on whether households fall above or below a specific eligibility threshold. The third metric is the Area under the ROC (receiver operating characteristic) curve, or *AUC*, which summarizes targeting accuracy not just at a single classification threshold (in our case, the 21% quota), but rather for all possible classification thresholds (i.e., quotas that range from 0% to 100%).[14] A perfect classifier achieves an AUC value of 1, whereas a random classifier (that targets randomly chosen households to fill the quota) produces an AUC of 0.5.[15]

Targeting accuracy Figure 2 summarizes the three measures of targeting accuracy for the main targeting methods; Table 1 provides details on the performance of these and other targeting methods described in Appendix A. Phone-based targeting (Spearman r= 0.23) is more accurate at identifying the consumption-poor than CBT ($r = 0.15$). However, both approaches are substantially less accurate than PMT ($r = 0.65$). The differences between the three methods are statistically significant ($p < 0.001$, using a Wilcoxon signed-rank test). The other survey-based targeting methods perform better than phone-based targeting and CBT, but worse than the PMT (PPI $r = 0.41$; asset index $r = 0.46$).

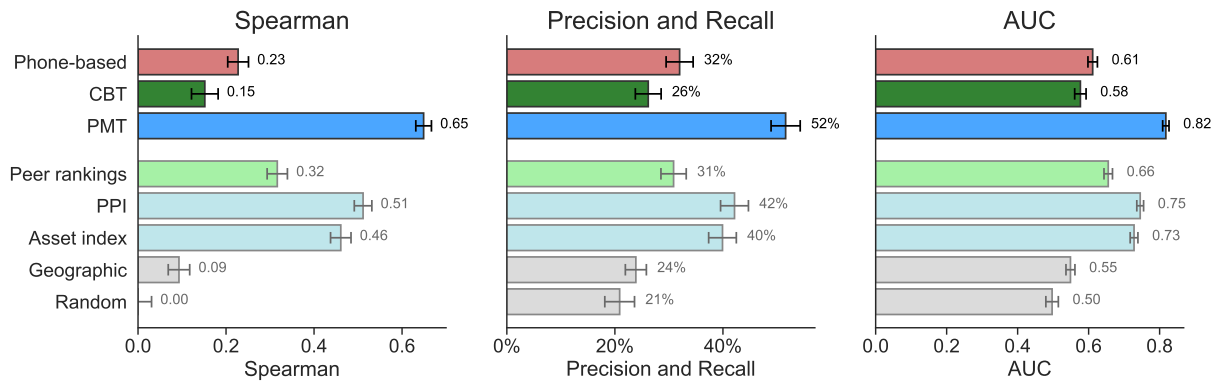
The decentralized peer ranking approach outperforms the CBT and is comparable in accuracy to phone-based targeting ($r = 0.32$). This suggests that either some knowledge households possess about the relative consumption poverty of their peers is being lost in the CBT exercise, or that the community places some direct weight on characteristics other than consumption poverty. We will explore this below. Perhaps surprisingly, people's self-assessment of their own socioeconomic status ($r = 0.40$) is more accurate than either the

recall and *precision* — the share of households classified as poor that are truly poor (one minus the type I error rate) — are equal in our setting.

[14] Specifically, the ROC curve shows how the true positive rate (recall) varies as a function of the false positive rate, for each possible classification threshold between 0 and 1. When the threshold for being classified as poor is low (e.g., if benefits are provided to any household that has more than a 5% chance of being poor), most households are targeted (resulting in many true positives, but also many false positives); by contrast, when the threshold is high, few households will be targeted (fewer true positives and fewer false positives).

[15] These accuracy metrics evaluate each method's ability to identify low-consumption households. For the PMT, PPI, and phone-based targeting, where machine learning models are trained to predict consumption, this is a natural evaluation. However, the CBT and peer-ranking exercise were not designed to elicit opinions about consumption per se. Instead, the CBT protocol followed standard practice and asked community members to identify community members with the lowest "socio-economic status", and the peer rankings asked households to identify the households that "have the least." As we discuss after presenting results on targeting accuracy, there is suggestive evidence that community-based methods prioritized non-consumption measures of vulnerability.

Figure 2: Comparing accuracies of targeting methods



Notes: These figures compare the accuracy of our targeting methods, based on Spearman correlation with consumption (left), precision and recall for identifying the 21% consumption-poorest households (middle), and area under the ROC curve (right). The length of the bar depicts the mean over 100 test-train splits, with the brackets showing two standard deviations above and below the mean.

Table 1: Accuracy metrics for all targeting methods

Targeting Method	**Spearman**	**Precision**	**AUC**
Panel A: Main targeting options			
Phone-based targeting	0.23 (0.02)	32% (3%)	0.61 (0.01)
CBT	0.15 (0.03)	26% (2%)	0.58 (0.02)
PMT (LASSO)	0.65 (0.02)	52% (3%)	0.82 (0.01)
Random	0.00 (0.03)	21% (3%)	0.50 (0.02)
Panel B: PMT variants			
PMT (OLS)	0.65 (0.02)	51% (3%)	0.82 (0.01)
PMT (Stepwise)	0.64 (0.02)	51% (3%)	0.81 (0.01)
PMT (Random Forest)	0.62 (0.02)	48% (3%)	0.80 (0.01)
Panel C: Other Survey-based targeting options			
PPI	0.51 (0.02)	42% (3%)	0.75 (0.01)
Asset index	0.46 (0.02)	40% (3%)	0.73 (0.01)
Panel D: Geographic targeting options			
Unions	0.09 (0.02)	24% (2%)	0.55 (0.01)
Villages	0.09 (0.03)	24% (2%)	0.54 (0.01)
Neighborhoods	0.08 (0.03)	24% (3%)	0.54 (0.01)
Panel E: Decentralized CBT			
All ratings	0.32 (0.02)	31% (2%)	0.66 (0.01)
Neighbor ratings only	0.23 (0.02)	28% (3%)	0.61 (0.01)
High confidence neighbor ratings only	0.32 (0.02)	31% (2%)	0.66 (0.01)
Own rating only	0.40 (0.02)	30% (1%)	0.67 (0.01)
All rankings	0.15 (0.03)	25% (3%)	0.57 (0.02)
Neighbor rankings only	0.03 (0.03)	22% (2%)	0.52 (0.02)
High confidence neighbor rankings only	0.09 (0.03)	25% (3%)	0.55 (0.02)

Notes: Each cell in this table presents the mean and, in parentheses, standard deviation of a targeting accuracy metric for one targeting method across 100 test-train splits. Metrics (Spearman, Precision / Recall, and AUC) are in columns. Each row corresponds to a targeting method described in Appendix A.

CBT ($r = 0.15$) or the decentralized ranking ($r = 0.32$). Of course, self-ratings are most susceptible to strategic behavior and gaming by beneficiaries, and are likely not a viable policy option.

Binary classification errors do not capture potential differences in the magnitude of errors. Two classification methods could have similar overall error rates for a given threshold, but a method with small mistakes (which tends to exclude households just below the threshold and include households just above the threshold) will be preferred to a method with larger mistakes (which tends to exclude households far below the threshold and include households far above the threshold). In Figure 3, we assess the magnitude of errors by showing the distribution of consumption per capita for households included and excluded by each targeting approach. Figure 3 suggests that the PMT tends to include poorer households than phone-based targeting and CBT, and that phone-based targeting includes poorer households than CBT. Similarly, the households excluded by PMT are on average richer than the households excluded by phone-based targeting, which are in turn richer than the average household excluded by CBT.

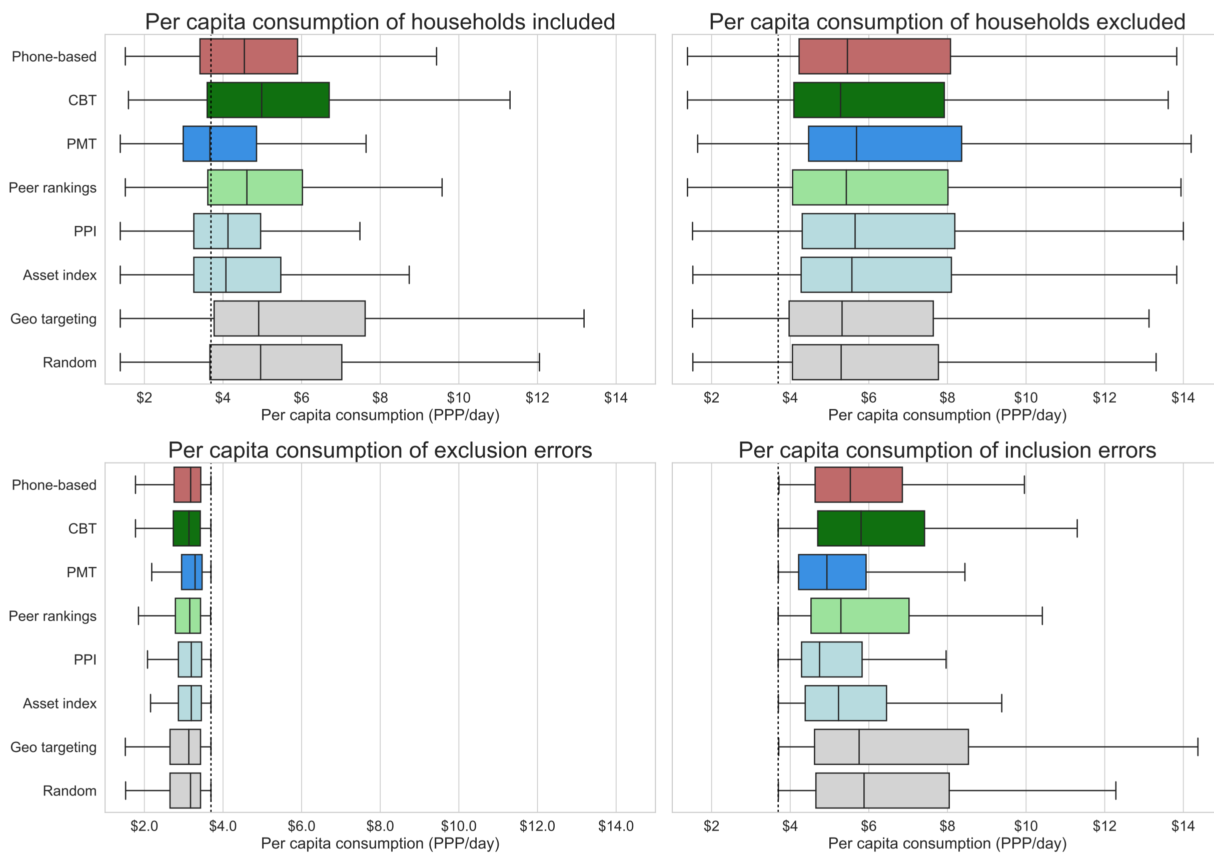
Who is targeted by each method? The different targeting methods use different information to rank households. Figure S4 highlights the variables selected by the PMT. These include demographic characteristics (large households with lots of children, disabled household head), information on asset ownership (those lacking vehicles, fridges, large plots of residential and agricultural land, and large houses with cement roofs), as well as the household's geographic location.[16] By contrast, the features of mobile phone use most correlated with per-capita consumption, shown in Table S2, are quite different. These include mobile phone recharge behavior (which captures transactions where the subscriber adds phone credit to their SIM card),[17] how frequently they use mobile data (which might be a proxy for owning a smartphone), features of their network (such as the number of unique phone numbers the user connects to for incoming or outgoing calls), and aspects of their mobility as inferred from the location of cell towers with which the phone connects.

To better understand what types of households are selected by each targeting method, Table 2 shows the results of regressing a binary indicator of household selection on a number of different characteristics at the household and community-level. In the first column, we

[16]While 5 of the top 20 PMT variables are geographic indicators (unions), purely geographic targeting performs poorly, as seen in, e.g., Figure 2. This suggests that the LASSO-based PMT can capture useful geographic information, but that relying entirely on geography will miss significant variation within geographic units.

[17]In Bangladesh, the vast majority of subscribers are on prepaid contracts. For these phones, the subscriber has to first add value to their account via recharge, and can then use the available balance on their account to make calls, send text messages, and so forth.

Figure 3: PCE distributions by inclusion / exclusion and targeting method



Notes: These figures depict the distribution of PCE per capita, by targeting method, for households included (top left), excluded (top right), wrongly excluded (bottom left), and wrongly included (bottom right). The boxes denote the 25th, 50th and 75th percentiles, while the whiskers capture the maximum and minimum. The vertical dashed line indicates the targeting cutoff (21^{st} percentile). A better-performing targeting method will tend to: include poorer households, shifting the distribution in the top left panel to the left; exclude less-poor households, shifting the distribution in the top right to the right; incorrectly exclude households closer to the cutoff (from below) rather than the poorest households, so the distribution in the bottom left panel will be compressed against the cutoff line from the left; incorrectly include households closer to the cutoff (from above) rather than the least-poor households, so the distribution in the bottom right panel will be compressed against the cutoff line from the right.

observe that phone-based targeting tends to select households that own a phone — this is mechanical, since we assign households without phones the lowest rank in phone-based targeting. However, conditional on owning a phone, households with fewer phone transactions (phone calls and text messages) are more likely to be selected. Phone-based targeting also targets households with fewer social connections (as measured in the survey); this suggests that the phone data may help reveal the extent to which households are socially isolated.[18]

Both community-based targeting and decentralized peer rankings are more likely to select widows and widowers for transfers than the PMT or phone-based targeting. This result is consistent with recent results from Indonesia by Sumarto et al. (2025), who observed that community members were making use of local, private information about the idiosyncratic challenges faced by specific households. Alternatively, the community may prioritize household characteristics other than consumption poverty in deciding how to target aid. Indeed, while phone-based targeting and the PMT prioritize households that spend a large share of their budget on food (a proxy for the household's subsistence risk (Bryan et al., 2014)), the CBT and peer rankings do not. We see no evidence in Table 2 to support concerns that the CBT systematically excludes ethnic minorities or socially isolated households (Cameron and Shah, 2014; Alatas et al., 2019).

While results thus far focus on the extent to which each method correlates with household consumption, prominent social scientists have debated whether societal welfare metrics should shift from our standard income and consumption measures to broader conceptions of "subjective well-being" (Kahneman and Krueger, 2006; Stiglitz et al., 2009). Results in Table 2 Panel A indicate that all methods select households that subjectively perceive themselves to be poor, where subjective perceptions are measured with the question, "compared to other households in your village, is your household a family that has the most, a family that has a lot, a family that has neither a lot nor a little, a family that has little, or a family that has the least?" For the CBT, this is consistent with Alatas et al. (2012), who find that community targeting is more likely to target households with lower self-reported well-being. To explore these subjective perceptions further, Appendix Figure S6 shows how accurately each targeting method predicts the subjective wealth assessment.[19] Peer rankings are most correlated with subjective perceptions, followed by the PMT, followed by phone-based targeting and CBT

[18]In the peer rankings module of the survey, each household was asked, for eight randomly selected households in their neighborhood, how well they know the household on a scale of 1-4. A household's "connectedness" is defined as the average knowledge ranking others assign to that household. Connectedness at the neighborhood level (Panel B) is defined as the average knowledge ranking for all households in the neighborhood.

[19]Appendix Figure S6 uses the same poverty score produced for each household that was used to produce Figure 2, and compare that value to the subjective wealth assessment instead of the true consumption value. The PMT and phone-based targeting models are not re-trained to predict subjective wealth.

(which are comparable). Thus, while the CBT is slightly more correlated with subjective wealth (Figure S6) than consumption expenditures (Figure 2), this does not explain the comparatively weak performance of CBT in our setting.

Heterogeneity: Do some methods perform better on specific types of households or neighborhoods? While our results thus far indicate that PMT targeting is substantially more accurate than the other options, and that phone-based performs better than community-based targeting, the aggregate results may mask important heterogeneity. For instance, CBTs might work better in more homogenous neighborhoods, or phone-based targeting might work best with active phone users. However, we find little evidence that the relative performance of different targeting methods varies systematically by neighborhoods or household type. In the top panel of Figure S7, we observe that the PMT generally performs better than phone-based targeting, which performs better than CBT, across all different types of communities — including when disaggregating by community size, by share of non-Muslim or non-Bengali minority households, etc. The bottom panel of Figure S7 tells a similar story with respect to heterogeneity by household characteristics (household size, household head gender/employment/minority status, connectedness, and amount of phone use (measured as the total number of calls and texts placed over the study period). For all types of households and neighborhoods, PMT performs best, and phone-based generally outperforms CBT.

Figure S7 also allows us to examine the absolute (as opposed to relative) performance of each targeting method across neighborhood and household type. Community-based targeting works better in more urban neighborhoods, and where average poverty levels are high. Interestingly, there is little variation in CBT performance by the minority share, size, and neighborhood connectedness. Both phone-based targeting and CBT are a bit more accurate within the set of non-minority households.

One exclusion of particular concern for phone-based targeting is that of households without phones, which account for 4% of surveyed households. However, in our setting, households without phones are not systematically poorer than households with phones: the average per capita consumption for households with phones is \$6.15 PPP (standard deviation of 3.43), and the average per capita consumption for households without phones is \$6.49 PPP (standard deviation of 3.99).

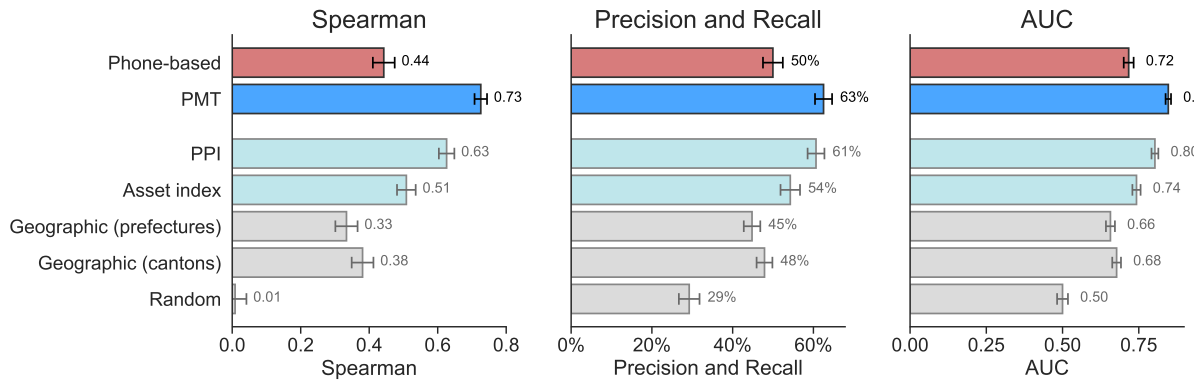
Targeting Within Neighborhoods The analysis presented thus far compares targeting methods in terms of how accurately each method identifies the poorest households from the overall study sample, which matches the goals of the GiveDirectly program. However, some programs may seek to identify the poorest households within each community, with a quota

Table 2: Correlates of inclusion and exclusion for each targeting method

	Phone-based	**CBT**	**PMT**	**Peer rankings**
Panel A: Household characteristics				
HH head female	0.008 (0.041)	0.036 (0.041)	0.067 (0.037)	0.044 (0.038)
HH head age	0.007 (0.013)	0.011 (0.013)	-0.054 (0.012)***	-0.012 (0.012)
HH head employed	0.013 (0.032)	0.004 (0.032)	0.028 (0.029)	-0.008 (0.030)
HH head minority	-0.072 (0.071)	-0.009 (0.071)	-0.008 (0.064)	-0.084 (0.066)
HH head widow/widower	-0.034 (0.056)	0.118 (0.056)*	-0.031 (0.050)	0.113 (0.052)*
HH size	-0.003 (0.013)	-0.023 (0.013)	0.155 (0.012)***	-0.032 (0.012)*
Connectedness (in)	-0.033 (0.015)*	0.002 (0.015)	-0.009 (0.013)	-0.015 (0.014)
Connectedness (out)	-0.021 (0.009)*	-0.006 (0.009)	0.020 (0.008)*	0.010 (0.008)
Own phone	0.320 (0.061)***	-0.065 (0.061)	0.011 (0.055)	-0.126 (0.057)*
Phone transactions	-0.080 (0.012)***	0.008 (0.012)	-0.033 (0.010)**	0.019 (0.011)
Food consumption share	0.031 (0.013)*	0.009 (0.013)	0.036 (0.012)**	0.001 (0.012)
Subjective wealth	-0.038 (0.014)**	-0.057 (0.014)***	-0.094 (0.012)***	-0.156 (0.013)***
Panel B: Neighborhood characteristics				
# of Households	0.065 (0.015)***	-0.004 (0.015)	0.026 (0.013)*	0.037 (0.014)**
Land area (square km)	-0.035 (0.014)*	0.006 (0.014)	-0.029 (0.013)*	-0.024 (0.013)
Density	0.010 (0.016)	-0.006 (0.016)	-0.029 (0.015)*	-0.040 (0.015)**
Urban	0.013 (0.097)	0.008 (0.097)	0.040 (0.087)	-0.017 (0.090)
% Minority	-0.005 (0.025)	-0.002 (0.025)	-0.039 (0.022)	0.016 (0.023)
Connectedness	0.059 (0.019)**	0.008 (0.019)	0.029 (0.017)	0.006 (0.017)
Average consumption	-0.001 (0.014)	0.011 (0.014)	-0.074 (0.013)***	0.025 (0.013)
Inequality (Gini)	0.008 (0.014)	-0.002 (0.014)	0.040 (0.013)**	-0.027 (0.013)*
Panel C: Controls				
Per capita consumption	-0.048 (0.016)**	-0.016 (0.016)	-0.031 (0.014)*	-0.028 (0.015)
Constant	-0.092 (0.067)	0.257 (0.067)***	0.157 (0.060)**	0.331 (0.062)***
N	1,252	1,252	1,252	1,252

Notes: Each column represents a separate regression, where the dependent variable is an indicator for whether a household was targeted by the method listed in the column header. The independent variables shown are a representative set of household (Panel A) and neighborhood characteristics (Panel B), while controlling for consumption expenditures (Panel C). Each regression includes all independent variables (from all panels). All explanatory variables are standardized. Household connectedness (in) represents the average knowledge that other households had of the household in question during the peer ranking exercise; household connectedness (out) represents the average knowledge that the household in question had of other households in their community. Neighborhood connectedness represents the average self-reported knowledge that households have of other households in their community, elicited during the peer rankings exercise in our household survey. Subjective wealth is measured with the question "compared to other households in your village, is your household a family that has the most, a family that has a lot, a family that has neither a lot nor a little, a family that has little, or a family that has the least?" in our household survey. Regressions are run using data from a single train-test split. Standard errors are in parentheses.

Figure 4: Targeting accuracy in Togo study



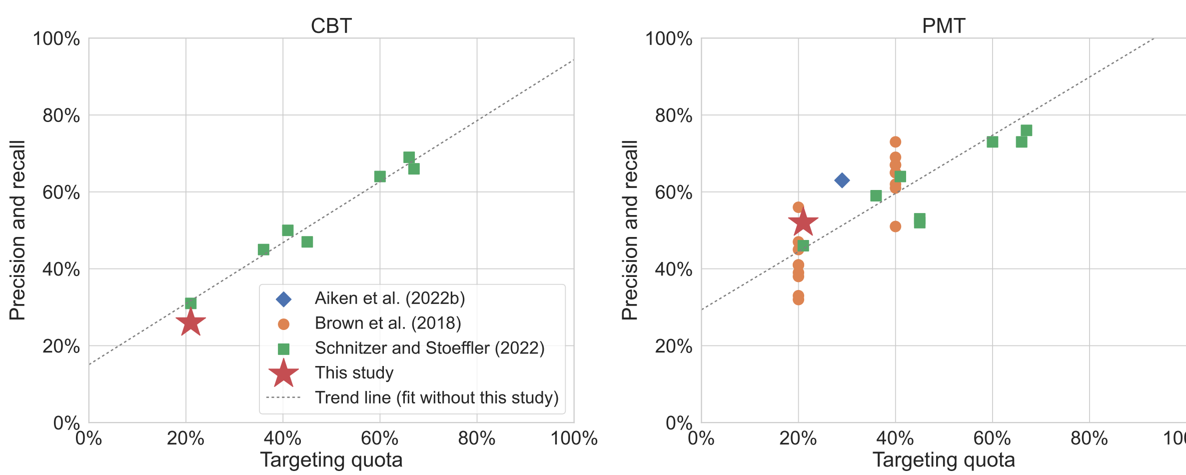
Notes: This figure shows the accuracy of different targeting methods for the country of Togo, reproducing results in Aiken et al. (2022). As in our analysis in Bangladesh in Figure 2, accuracy is calculated over 100 random train-test splits, and error bars show two standard deviations above and below the mean for each metric. This bootstrapping procedure explains very slight differences to the results presented in Aiken et al. (2022), where 1,000 train-test splits were used.

assigned at the community level. Importantly, in the CBT we implemented, communities were asked to rank households from poorest to richest, and were told that the poorest 20% of households within each community would receive a transfer. It is therefore possible that – while the CBT is weaker than phone-based targeting overall – it is better at identifying the poorest share of households within each community. To assess this possibility, we repeat the targeting evaluation with the objective of identifying the poorest 21% of households *within each neighborhood*. In Figure S8, we show that while the absolute accuracy of each targeting method declines with this evaluation approach (this is unsurprising, since geographic variation between communities is no longer used for targeting), the quality of targeting approaches relative to one another is unchanged: PMT is still the most accurate, followed by phone-based targeting, followed by CBT.

Generalizability The performance of phone-based targeting in Bangladesh is broadly consistent with what prior work has found evaluating a similar set of targeting approaches in Togo. In Figure 4, we replicate the results of Figure 2, instead using data from Togo (Aiken et al., 2022). In both settings, we find that the PMT is more accurate than phone-based targeting. However, the gap between phone-based targeting and PMT is wider in Bangladesh (63% difference in precision and recall and 34% difference in AUC) than in Togo (26% difference in precision and recall and 18% difference in AUC). The previous work in Togo did not include CBT as a possible targeting approach.[20]

While the targeting performance we observe — across *all* methods — is low in absolute terms (AUC = 0.52-0.82; precision and recall of 23%-52%), this performance is in fact consistent with what has been reported in prior work. Figure 5 plots the precision and recall, as a function of the share of households targeted, for the CBTs and PMTs documented in three papers: (1) Schnitzer and Stoeffler (2022), which evaluates the targeting accuracy of seven CBT-based and eight PMT-based social protection programs run in parts of Burkina Faso, Cameroon, Mali, Niger, and Senegal with targeting quotas ranging from 21% to 67%; (2) Brown et al. (2018), which simulates PMT-based country-level targeting in eight African countries with 20% and 40% targeting quotas; and (3) Aiken et al. (2022), which calculates targeting accuracy nationwide in Togo for a PMT with a 29% targeting quota. In Figure 5, we see that targeting accuracy increases approximately linearly with the share of households targeted, despite large variations in data, program implementation, and study contexts. As indicated by the two stars, the accuracy of our CBT (left panel) and PMT (right panel) are close to the predicted values (dotted lines) for our targeting quota of 21%.

Figure 5: Comparison of targeting accuracy with other studies



Notes: These figures show comparisons of our results on targeting accuracy (red stars) with past studies that also use a quota approach to targeting evaluation (green squares for Schnitzer and Stoeffler (2022), blue diamonds for Brown et al. (2018), and orange dots for Aiken et al. (2022). The targeting error rate is shown as a function of the targeting quota.

One possible reason that targeting in our Bangladesh setting is challenging is the narrow

[20]Our finding that PMT is more accurate than CBT is consistent with most other papers that have compared the two methods (Schnitzer and Stoeffler, 2022; Premand and Schnitzer, 2021; Alatas et al., 2012), but the differences we observe are larger: we find that switching from CBT to PMT doubles precision and recall (from 26% to 52%) and increases AUC by 41% (from 0.58 to 0.82).

geographic scope of the program, which focused on 180 neighborhoods in Cox's Bazar district. In this relatively small region, there is less variation in poverty than at a national scale. To more directly illustrate how targeting performance is lower in more homogeneous populations, Figure S9 simulates targeting on subsets of households in the survey. When a small number of uniformly poor households are included (left side of plots), performance is quite poor; as the sample grows, targeting performance increases.

Combining targeting methods It is possible that different approaches to targeting could be complementary, in the sense that the combination of two targeting methods could improve overall targeting accuracy. Figure S10 shows the results of a simple strategy for combining targeting approaches. Our algorithm for augmenting method A with targeting method B is to replace the very last (least poor) household targeted for a transfer under method A with the poorest household targeted under method B who was excluded under method A. Such replacements can be repeated until all method-A-targeted households are replaced with method-B-targeted households. This yields a continuum of A-B combined targeting, where the "mixing parameter" (share of A-targeted households replaced with B-targeted households) varies from 0% to 100%. Figure S10 shows that combining rankings using this method does not improve overall targeting accuracy. Neither the phone + PMT nor the CBT + PMT approaches improve precision and recall relative to solely using the PMT rankings. We find that adding decentralized peer rankings can improve pure phone-based targeting by a few percentage points, but that the combination of phone + CBT does not improve over pure phone-based targeting.

An alternative, intuitive approach to combined targeting simply gives the targeting algorithm access to variables from multiple data sources. For example, for the combined phone-based targeting + PMT method, we give a machine learning algorithm access to all phone features and all PMT variables in the ML model. Similarly, to combine PMT and CBT, we include CBT rankings and all PMT variables. Figure S11 shows that algorithms that have access to data from multiple data sources generally do not improve accuracy relative to using a single data source.

4 Cost-effectiveness results

Thus far, our results indicate that proxy means testing is more accurate than phone-based targeting, and that phone-based targeting is more accurate the community-based targeting. However, the costs associated with these different targeting methods also vary substantially. For instance, phone-based targeting is typically much cheaper than proxy means testing

— especially for large scale programs — because phone-based targeting does not require in-person primary data collection for screening and therefore has essentially zero marginal cost. This creates a tradeoff between cost and accuracy. In this section, we use a simple framework based on Hanna and Olken (2018) to characterize the conditions under which the different targeting methods would be more cost-effective.

We assume that the implementer has a total budget B and chooses between targeting methods to direct as much money as possible to the neediest individuals. Method m incurs targeting costs C_m, leaving $T_m = B - C_m$ for transfers. We assume that the implementer wishes to target I "included" households out of a total population of $I + E$ (included and excluded) households, and is constrained to make equal payments b_m to each recipient, with $b_m = (B - C_m)/I$.[21] We assume a household constant relative risk aversion (CRRA) utility function, so household i's utility is given by

$$U_i = \frac{c_i^{1-\sigma}}{1-\sigma},$$

where consumption c_i is equal to the household's pre-program consumption level y_i plus the transfer b if the household receives it, so $c_i = y_i + 1\{i \in I\}\, b$. We assume that the implementer has an objective function that maximizes the sum of household utilities

$$\begin{aligned} V &= \frac{1}{1-\sigma} \sum_{i=1}^{N} c_i^{1-\sigma} \\ &= \frac{1}{1-\sigma} \sum_{i \in I} (y_i + b)^{1-\sigma} + \frac{1}{1-\sigma} \sum_{i \in E} y_i^{1-\sigma}. \end{aligned}$$

Due to diminishing marginal utility, the implementer prefers to allocate transfers to households with lower pre-program y_i, but faces a tradeoff if identifying and targeting such households is more costly and therefore reduces b.

After fixing a hypothetical program's budget and the number of people screened, we calculate the screening costs associated with different targeting approaches, and then calculate the total budget remaining that can be provided as benefit transfers. Fixing the targeting threshold at 21% — as in GiveDirectly's program in Bangladesh — we then allocate the transfers to the targeted households and calculate V under each possible targeting regime. Following Hanna and Olken (2018), we use $\sigma = 3$ to calculate V. We first calculate V_0, the value of the implementer's objective function in the absence of the program, and $V_{1\mathrm{B}}$, the "first-best" or "oracle" value, i.e., if the implementer could costlessly obtain the exact ranking

[21] We take the total budget B and the number of people targeted I as given, although in principle either or both could depend on the accuracy of the targeting method. We constrain the implementer to equal transfers for simplicity, which reflects the design of most real-world social protection programs.

of all households and target perfectly. The gain in this first-best scenario, then, is $V_{1B} - V_0$. We then calculate V_m, the value of the objective function for each method m (i.e., PMT, phone-based targeting, CBT), and report G_m, the gain relative to the first-best:

$$\text{Relative gain from method } m = G_m = \frac{V_m - V_0}{V_{1B} - V_0}. \tag{1}$$

In comparing programs, we will refer to methods that produce a higher G_m as more "cost-effective" to capture the intuition that they produce a higher sum of household utilities for a fixed budget B.

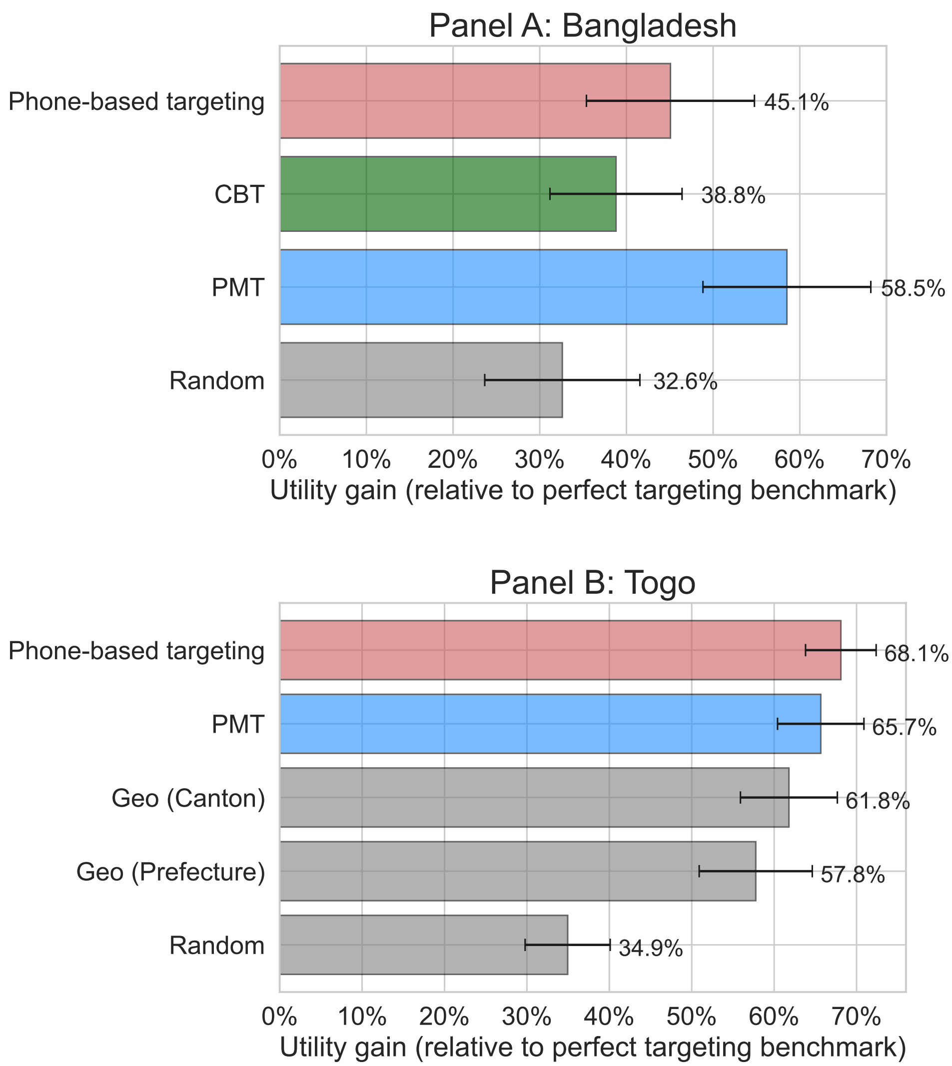
Data on Costs. Screening costs are a key input for computing G_m. To analyze cost-effectiveness, we therefore supplement our survey data and mobile phone records with detailed information on the costs of administering each targeting approach. We estimate the cost of the PMT using costs from our census, which lasted approximately 15 minutes (similar to the time required for a typical PMT scorecard) and cost approximately \$1.25 per household, in addition to fixed costs of \$6,300 for enumerator training and equipment. This marginal cost per household for a PMT is lower than typical: past work that reviewed the published PMT costs in the research literature found the median reported PMT cost to be \$4.00 (Aiken et al., 2023b). Phone-based targeting incurs a fixed cost for researcher time in implementing the machine learning method, but the marginal cost per household is approximately zero. Our CBT exercises had a variable cost of \$2.33 per household screened, plus a fixed cost for training and equipment of \$19,300. Because CBT is both more expensive *and* less accurate than phone-based targeting in our setting, we focus primarily on comparisons between PMT and phone-based targeting.[22] Both the PMT and phone-based targeting approaches require a detailed household consumption survey for model calibration. In Bangladesh, this cost \$46,600 for 5,000 households — though both the PMT and phone-based targeting required this survey, so this element does not meaningfully affect our cost-effectiveness comparisons.[23]

PMT is most cost-effective in Bangladesh We begin by computing G_m generated by the GiveDirectly program in Bangladesh (described in Section 2), which had a budget of roughly \$5 million and screened around 100,000 households. Based on our own surveys, we estimate the total screening costs for a PMT in this setting at \$177,900, leaving around

[22]Our CBT cost per household is similar to the median value of \$2.20 reported in a recent review (Aiken et al., 2023b).

[23]The consumption survey cost of roughly \$10 per household in our setting is much lower than other costs reported in the literature: Kilic et al. (2017) report costs ranging from around \$50-500 per household surveyed (\$200,000 to over \$4 million total) for nationally representative consumption surveys enumerated as part of the Living Standards Measurement Surveys program.

Figure 6: Gains by targeting method



Notes: These figures show G_m, the gains in the implementer's objective function, for different targeting methods m. As in Equation (1), gains are relative to those obtained under costless perfect information. Calculations are based on the design and parameters of the cash transfer programs implemented in Bangladesh (top panel) and Togo (bottom panel) (see Section 4). The Bangladesh program had a $5 million budget for 100,000 households screened, targeting 21% of households, with targeting accuracy shown in Table 1. The Togo program had a $5 million budget for 207,000 households screened, targeting 29% of households, with targeting accuracy shown in Figure 4.

$4.82 million for cash transfers. The screening costs for phone-based targeting were $46,600, leaving around $4.95 million for cash transfers. The GiveDirectly program in Bangladesh targeted 21% of households (21,000 households), implying a choice between a PMT program that provides $229 cash transfers to the 21,000 households predicted to be poorest by the PMT, or a phone-based program that provides $236 cash transfers to the 21,000 households predicted to be poorest by phone-based targeting.

Figure 6 (top panel) shows the gains in utility G_m from the GiveDirectly transfers in southern Bangladesh under the assumption of CRRA utility, relative to the gains generated by costless perfect targeting (V_{1B}). Despite the higher costs of the PMT, the PMT's higher targeting accuracy results in larger gains than phone-based targeting (58.5% of the gains of costless perfect targeting vs. 45.1%). The CBT approach, which is both more costly and less accurate than phone-based targeting, produces smaller gains (38.8%).

Phone-based targeting is most cost-effective in Togo To better understand the generalizability of our findings, we repeat this calculation for the GiveDirectly-Novissi (GD-Novissi) program in Togo described in Aiken et al. (2022). GD-Novissi also had a budget of roughly $5 million, screened roughly 207,000 households, and targeted the poorest 29% among the households screened. For our analysis of cost-effectiveness of GD-Novissi, we use nationally representative survey data from Togo collected in 2018, matched to mobile phone records from the same year.

Several key differences between the two research settings in Bangladesh and Togo are worth noting: First, nearly twice as many households were screened in Togo, though the program had roughly the same total budget as the Bangladesh program. Second, screening costs per household were higher in Togo ($4.00) than in Bangladesh ($1.25). Third, community-based targeting data was collected in Togo, so we can only compare PMT and phone-based targeting there. Finally, as we saw when comparing Figure 2 and Figure 4, phone-based targeting in Togo was substantially more accurate (relative to the PMT) than in Bangladesh. This was likely due to the fact that (a) households in Togo were drawn from a nationally representative sample, whereas the households in Bangladesh were drawn from a relatively small and homogeneous region; and (b) the survey in Togo was restricted to households that provided a phone number that could be matched to mobile phone metadata, so non-phone-owning or unmatched households are not included in the analysis. In contrast, households without phones are included in the Bangladesh analysis and assumed to be targeted last under the phone-based targeting approach.

The bottom panel of Figure 6 shows the gains from different targeting approaches for the GD-Novissi program in Togo. The results are qualitatively different, as the gains from phone-based targeting (68.1% of the gains from costless perfect targeting) exceed the gains produced by PMT (65.7%). This results from the higher PMT screening costs in in Togo, which we estimate to be \$311,650, leaving around \$4.7 million for cash transfers — or \$78 per beneficiary. By contrast, screening costs for phone-based targeting in Togo would be only $46,600$, leaving \$4.95 million for cash transfers — or \$82 per beneficiary.

These contrasting results from Bangladesh and Togo illustrate how the relative cost-effectiveness of phone-based targeting and proxy means testing depends on the scale of the program (both the total program budget and the number of households that need to be screened), as well as the relative accuracy of the targeting methods being compared and the cost of screening each household. Our next set of results analyzes these tradeoff more systematically by varying the scale and scope of several hypothetical transfer programs.

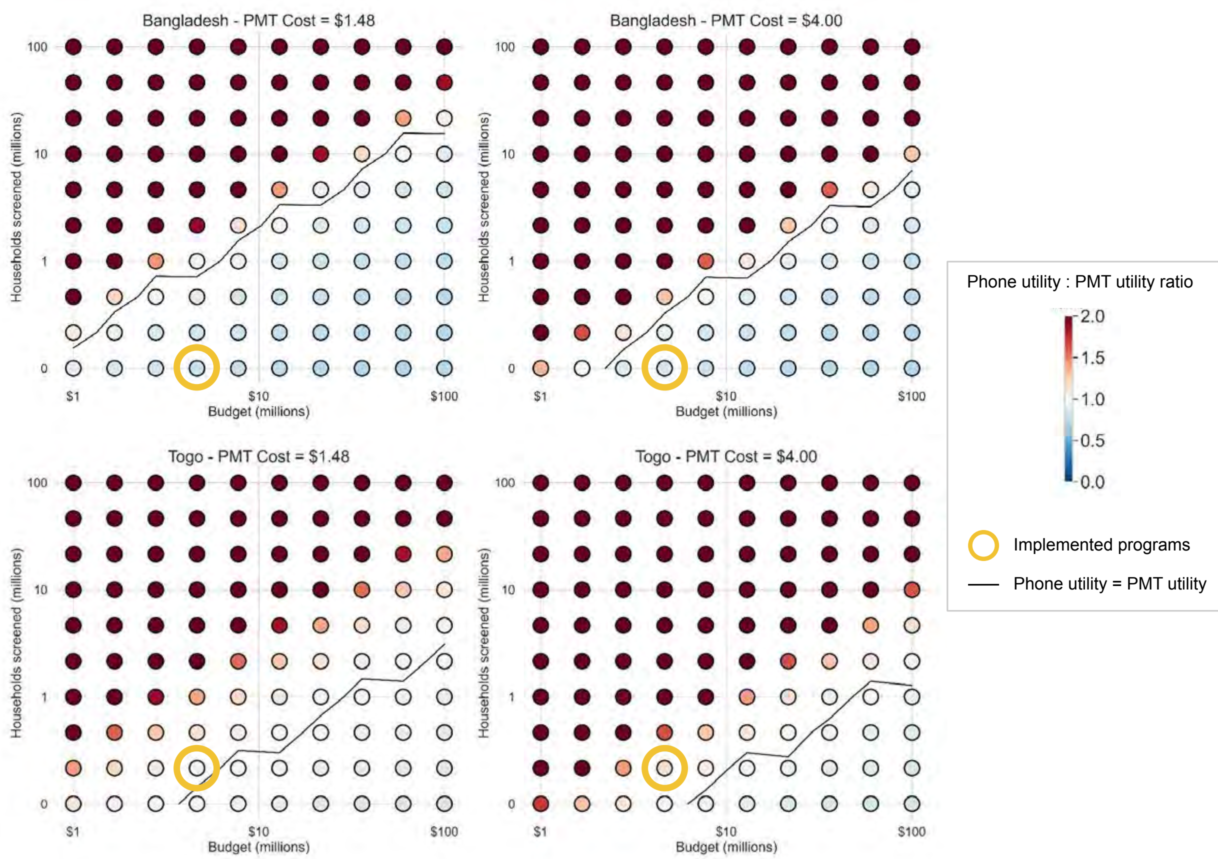
The most cost-effective approach depends on program scale The GiveDirectly programs we analyze were relatively small — both in terms of the total budget and the number of households screened for eligibility — relative to national-scale social protection programs typically run by governments. For example, government cash transfer programs in Bangladesh typically have budgets of \$10-300 million, and a mandate to screen all 41 million households in the country for eligibility.[24]

Figure 7 provides a visual comparison of the performance of phone-based targeting and PMTs for a range of hypothetical programs, varying both the total program budget (horizontally) and the number of households screened (vertically). Each dot corresponds to a hypothetical program, with the shading determined by the ratio of the total household utility of phone-based vs. PMT targeting, V^{phone}/V^{PMT}. When transfer programs have a large budget relative to the size of the population screened, like GiveDirectly's programs in Bangladesh and Togo, then PMT-based targeting results in larger gains (blue dots). However, for programs with small budgets relative to the number of households screened, which characterizes many real-world government-run social protection programs in Bangladesh and elsewhere, phone-based targeting is preferred (red dots). This is mainly because the marginal cost of screening additional beneficiaries using mobile phone meta-data is negligible.

The top-left panel of Figure 7 roughly corresponds to the cost structure of the GiveDirectly program in Bangladesh. The circled point illustrates that for the actual program that was implemented in Bangladesh, the PMT would have yielded higher V than phone-based targeting. This is partly because the variable per-household screening cost for the PMT in Bangladesh (\$1.25) was unusually low, reflecting uniquely low costs of data collection in

[24] These figures are taken from the budgets of large cash assistance programs in the fiscal year 2019-2020, reported in World Bank (2021).

Figure 7: Cost-effectiveness of PMT vs. phone-based targeting as parameters vary



Notes: Figures depict ratios of gains between phone-based targeting and proxy means testing (PMT) as a function of a hypothetical social protection program's budget (x-axis) and households screened (y-axis). Red shades (darker) represent program scales at which phone-based targeting is preferred, blue shades (lighter) represent program scales at which PMT is preferred, and the line identifies the "decision threshold". Left: PMT variable cost of $1.25 per household screened, based on the costs of our surveys in Bangladesh. Right: PMT variable cost of $4.00 per household screened, based on the median of values reported in the literature. Above: Using data from Bangladesh. Below: Using data from Togo.

Bangladesh. The right two panels of Figure 7 show how the relative performance of PMT changes if the cost of screening households in Bangladesh were in line with the median per-household screening cost reported in the literature of \$4.00 (Aiken et al., 2023b). For more typical PMT screening costs, the scope and scale of programs where phone-based targeting is preferred to PMT expand.[25] More broadly, Figure 7 highlights that a key factor in determining which targeting method performs best is the ratio of the program budget to the number of households screened. Phone-based targeting looks relatively more attractive for national-scale programs with small budgets, or more precisely, for programs where the budget is small relative to the number of households that must be screened.

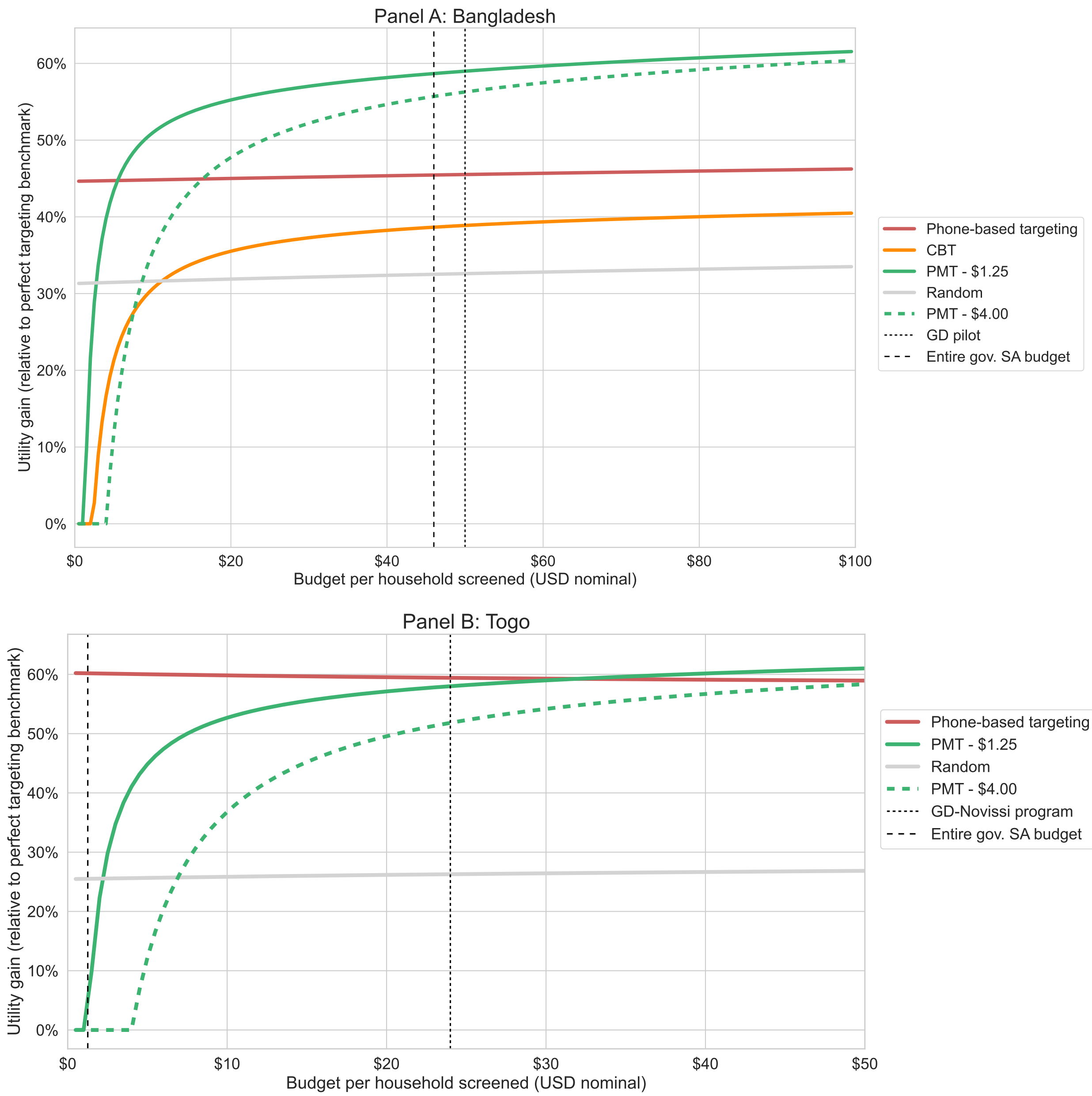
Figure 8 more directly illustrates how the cost-effectiveness of different methods depends on this key parameter, the *program budget per household screened*.[26] In Bangladesh (Panel A), phone-based targeting (red line) is preferred to the PMT for programs with budgets of less than \$4.00 per household screened if the PMT costs \$1.25 (solid green line); however, if the PMT costs \$4.00 (dashed green line), phone-based targeting is preferred for budgets of up to \$15 per household screened. In Togo (Panel B), phone-based targeting is preferred for a wider range of program budgets: when the PMT variable cost of \$1.25 is used, phone-based targeting is preferred for programs with budgets under \$31 per household screened; with a PMT cost of \$4.00, phone-based targeting is preferred for budgets under \$51 per household screened.

To anchor these comparisons to real-world social protection program scenarios, Figure 9 plots budgets as a function of the number of households screened for a number of countries (across the GDP per capita spectrum) using data from the World Bank's ASPIRE database. The figure shows where existing programs fall relative to the two thresholds of \$51 per household (using costs and accuracy from Togo) and \$15 per household (using costs and accuracy from Bangladesh). Sixty-six of 95 countries have budgets over \$51/hh, and so PMT would be preferable under both thresholds; 10 of 95 countries have budgets that are sufficiently low (less than \$15/hh) to prefer phone-based targeting under both thresholds; and 19 of 95 are intermediate cases, with PMT preferred using cost and accuracy estimates from Bangladesh but phone-based targeting preferred using cost and accuracy estimates from Togo. Appendix E provides more details on each program represented in the figure.

[25] Figure S12 further illustrates how the performance of PMTs and phone-based targeting vary with other important aspects of program design, including the fraction of beneficiaries targeted, the coefficient of relative risk aversion, and the variable cost of the PMT.

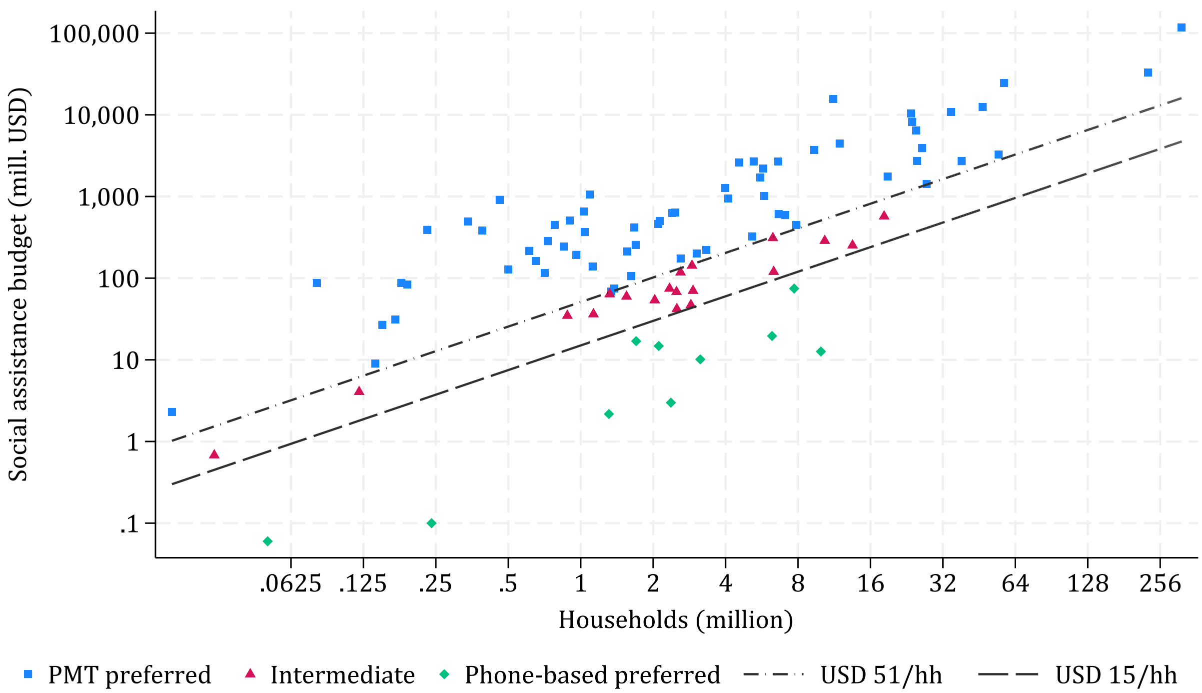
[26] In constructing Figure 8, we ignore fixed costs. For medium- to large-scale programs, these will be a tiny fraction of total costs; for example, fixed costs make up 4% of screening costs for a PMT-targeted program screening 1 million households, but only 0.4% of screening costs for a PMT-targeted program screening 10 million households.

Figure 8: Gains by method as a function of budget per household screened



Notes: These figures show gains in the implementer's objective function as a function of the program's budget per household screened. As in Equation (1), gains are relative to what could be achieved with costless perfect information. The top panel uses our data from Bangladesh. This analysis requires fixed costs to be dropped from calculations, implicitly assuming that fixed costs are negligible when the number of households screened are sufficiently large. The solid green line uses the PMT variable cost of $1.25 per household screened, based on the costs of our surveys in Bangladesh, and the dashed green line the value of $4.00 per household screened, based on the median of values reported in the literature. The bottom panel uses data from Togo. In each figure, the dashed vertical lines show the budget of the aid program by GiveDirectly and the entire government cash assistance budget.

Figure 9: Social assistance: total budgets vs households screened



Notes: This figure plots countries' budgets for cash-based social assistance transfers (y-axis) versus the number of households to be screened (x-axis). Each of the 95 points represents a country. The upper dashed diagonal line denotes a budget of $51 per household screened, which is the cutoff between cost-effectiveness of PMT (above) and phone-based targeting (below) based on screening costs and accuracy from the Togo study. The lower dashed diagonal line denotes the PMT vs. phone-based cutoff of $15 per household screened, using parameters from the Bangladesh study. For programs above the upper line (blue squares, 66 observations), PMT is preferred to phone-based targeting under both scenarios. For programs below the lower line (green diamonds, 10 observations), phone-based targeting is preferred to PMT in both scenarios. For programs in between the two (red triangles, 19 observations), PMT is preferred to phone-based targeting using parameters from Bangladesh, but phone-based targeting is preferred to PMT using parameters from Togo. Data are from the World Bank's Aspire database, World Bank Open Data and the Global Data Lab, and are described in greater detail in the main text.

Sensitivity of cost-effectiveness results The analysis in this section has emphasized how the choice of the preferred targeting method (i.e., the one that maximizes G_m) depends on the size of the program relative to the number of households screened. Here, we briefly illustrate how this determination is influenced by other aspects of the program design and assumptions about the shape of the utility function.

Perhaps most notably (and intuitively), the results in Figure 8 are sensitive to the accuracy of the underlying targeting technology. In comparing Panels A and B of Figure 8, we saw that in Togo, where phone-based targeting is more accurate, phone-based targeting was preferred for a larger range of programs (i.e., Figure 8). Figure S13 shows how the gains from phone-based targeting increase as the targeting algorithm becomes more accurate (see Appendix B for details on these simulations). In both Bangladesh (Panel A) and Togo (Panel B), when the Spearman correlation between phone-based poverty estimates and consumption is around 0.20 (as in Bangladesh), programs with budgets under \$15 per household screened should use phone-based targeting. As the correlation increases to 0.40 (as in Togo), phone-based targeting performs better for programs up to \$40 per household screened. The exact points at which phone-based targeting would be preferred over a PMT are shown in Table S3.

Figure S12 illustrates how three other factors likewise influence the performance of phone-based targeting relative to the PMT. In the left panels, we observe how screening costs affect the relative gains (G_m) of each targeting method: as the variable cost of the PMT increases, the gains from PMT decrease. In Togo, phone-based targeting is more cost-effective for almost all PMT costs (bottom-left panel); in Bangladesh (top-left panel), phone-based targeting is only preferred when PMT costs exceed roughly \$14 per household (for context, the highest PMT cost reported in the survey by Aiken et al. (2023b) is \$9.50).

The middle panels illustrate the sensitivity of our results to the coefficient of relative risk aversion σ: when σ is small (i.e., the cost of targeting errors are relatively small), the differences between methods are smaller, while at very high values of σ gaps increase. Intuitively, if household utility is close to linear in consumption (σ small), then there is little aggregate loss from mistakenly allocating a transfer to a relatively well-off household rather than a worse-off household. If the curvature of the household utility function is large (σ high), the aggregate loss from this misallocation is large. In Togo (bottom-middle panel), we observe that when σ is very high, and consumption of the most poor matters more, the PMT becomes preferred: despite its high cost relative to phone-based targeting, the additional accuracy becomes more valuable, because as σ increases the cost of errors becomes greater.

Finally, the right panels of Figure S12 show that the relative performance of different methods is not particularly sensitive to the fraction of the population that is eligible for transfers. For all three of our main targeting methods, higher targeting thresholds (that is, providing benefits to a larger share of households, but a lower level of benefit to each targeted household) lead to better performance relative to the first-best. However, changing the targeting threshold does not meaningfully change the ordering of which method is preferred: in Bangladesh, PMT uniformly outperforms phone-based targeting at all targeting thresholds, and phone-based targeting always outperforms CBT. In Togo, PMT outperforms phone-based targeting up to a threshold of approximately 25%, and the two perform similarly at higher thresholds.

5 Discussion and Conclusion

Our paper produces two key findings. First, in southern Bangladesh, targeting poor households using machine learning and mobile phone data (AUC = 0.61) is more accurate than community-based targeting (AUC = 0.58), but less accurate than proxy-means testing via household surveys (AUC = 0.82). Second, we provide the first direct comparison of targeting approaches based on cost-effectiveness, building on the welfare framework introduced by Hanna and Olken (2018). We show that it would be more cost-effective to use traditional approaches like proxy-means testing to target social protection programs with large budgets relative to the number of households screened, but algorithmic targeting approaches are worth considering for programs with thinly stretched budgets (below $10-50 per household screened). Data from real-world government-run social protection programs suggest that most program budgets are sufficiently large that proxy-means testing is the most efficient targeting approach. However, 10-30% of countries in the World Bank ASPIRE database have small enough social protection budgets relative to the size of their population such that phone-based targeting would be preferred.

Robustness and Limitations The insights we present may be sensitive to the details of real-world social protection programs. First, our empirical analysis draws heavily on two specific programs implemented in Bangladesh and Togo, with household surveys and administrative cost data collected there. We have added context by using data from other programs (e.g., the World Bank's ASPIRE database), and through simulations of results under counterfactual parameters (e.g., to show how conclusions would differ if phone-based or community-based targeting were substantially more accurate, as in Figure S13). However, to make specific recommendations in individual countries, or to generalize across a wider range of environments, more work is needed to better calibrate the costs and accuracy of targeting approaches, especially novel digital approaches like phone-based targeting.

Second, our analysis is focused on a one-time program in which beneficiaries received transfers soon after the data used to determine eligibility was collected. In practice, PMT and CBT targeting sweeps are typically conducted only every few years (Barca and Hebbar, 2020). This lowers per-year screening costs proportionally, making PMT and CBT look more attractive relative to phone-based targeting, as shown in Figure S12. However, targeting accuracy also typically deteriorates as data become out-of-date (Hillebrecht et al., 2023; Brown et al., 2018). For instance, Aiken et al. (2025) estimate that PMT targeting accuracy decreases, on average, by 9 percentage points for each year that the PMT data are out of date. Aiken et al. (2022) show that the accuracy of phone-based targeting also degrades over time, as the underlying relationship between phone use and poverty changes.

Third, our measurement of costs focuses only on the financial costs of screening households. The private costs to households who participate in screening activities, such as responding to household surveys or participating in community meetings, are not accounted for in our cost analysis. Imputing the value of people's time using survey data on hourly wages increases the cost of targeting via PMTs by approximately 9%. For CBTs, targeting costs would increase by 94% for households that participate in CBT meetings, although non-participating households would not incur such costs.[27] Accounting for these non-monetary costs would make phone-based targeting look relatively more attractive.

A final limitation of our analysis is that we have abstracted away from the concern that households might strategically alter their behavior to game the targeting regime after the targeting mechanism has been implemented (Goodhart, 1975; Lucas, 1976). For instance, households might look for ways to manipulate the information that is collected in the PMT (Camacho and Conover, 2011; Banerjee et al., 2020) or change the way they use their mobile phones (Björkegren et al., 2020) in order to become eligible for benefits. Alternatively, it is possible that community members might try to influence the decisions made by CBT selection committees (Han and Gao, 2019; Alatas et al., 2019). Ex ante, it is not obvious which of the targeting processes would be most impacted by these factors. In principle, the "black box" nature of machine learning algorithms may make them more difficult for beneficiaries to game. At the same time, beneficiaries may prefer – or regulations may require – decision rules that are transparent (Walmsley, 2021).[28]

[27]We estimate respondents' opportunity costs of time by multiplying the average duration of PMT surveys and community meetings by the average hourly-income of an earning household-member in our sample. The latter is simply the average household-income of the households we surveyed divided by the average number of earners per household in 2022 reported by Bangladesh Bureau of Statistics (2023). Expressing this time cost as a proportion of the per-household variable cost of the PMT survey or community meeting provides estimates of how the corresponding targeting costs increase when accounting for respondents' time. The relatively large increase for CBT is partly because, although an enumerator's costs for a single CBT meeting are divided among many households, each meeting participant still spends as much time in the meeting as the enumerator does.

[28]For instance, the European Union's GDPR requires that beneficiaries receive "meaningful information

Beneficiary Satisfaction One final consideration that we do not account for in our cost analysis is the possibility that beneficiaries have preferences over targeting methods. Prior experimental comparisons of CBTs to PMTs have found mixed results: Alatas et al. (2012) find that community members randomly assigned to a CBT program in Indonesia were more satisfied with the targeting process than those assigned to a PMT; however, Premand and Schnitzer (2021) find that community members assigned to a PMT in Niger were more likely to perceive the process as legitimate than those assigned to a CBT. We collected data on people's satisfaction with the process and their perceptions of fairness for both community- and phone-based targeting, after transfers were delivered. We present those comparisons here, but there are several caveats to interpreting those results: (a) The value of the transfers allocated by phone-based targeting and CBT were very different ($300 vs. $9); (b) they were delivered at different times (9 months vs. 2 months before the satisfaction survey); (c) we provided almost no information about the process of phone-based targeting, whereas a community meeting accompanied the CBT; and (d) unlike Alatas et al. (2012) and Premand and Schnitzer (2021), the targeting method was not randomly assigned; CBT transfers were made in a random subset of villages that had already received phone-based transfers.

We conducted the satisfaction survey in October 2024 with 1,100 households in villages that received both phone-based and CBT transfers. While most respondents remembered the programs, their understanding of how eligibility was determined was quite poor (see Appendix C and Appendix Table S4 for details). As shown in Appendix Table S5, satisfaction scores for CBT were significantly higher than for phone-based targeting: respondents were 27 percentage points more likely to report being "satisfied" or "very satisfied" (as opposed to "somewhat" or "not at all" satisfied) when asked about the CBT process than when asked about the phone-based process. They were also 35 percentage points more likely to perceive the CBT process as fair compared to their perception of the phone-based targeting process.[29] Those who actually received CBT transfers were especially fond of the CBT, and those who received phone-based transfers preferred the phone-based approach.

Concluding remarks Our analysis compares three different paradigms for poverty targeting, and shows how the preferred approach is likely to depend on program scale. In particular, the cost-effectiveness of the program depends critically on both program budget and the size of the population being screened. Our results suggest that although phone-based targeting is less accurate than proxy-means testing, it can be a more cost-effective way to determine

about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" (Selbst and Powles, 2017).

[29] Respondents were asked to give a yes or no response to the question, "In your opinion, was the selection process to receive cash aid in the program fair?"

eligibility for social protection programs with small budgets relative to their scale. We hope that this and subsequent analysis can further elucidate if and when algorithmic techniques would be a useful addition to the toolbox of targeting approaches available to policymakers.

References

Ahmed, A. and Bakhtiar, M. M. (2023). *Proposed indicators for selecting needy participants for the Vulnerable Women's Benefit (VWB) Program in urban Bangladesh*. Intl Food Policy Res Inst.

Ahmed, F., Genoni, M. E., Roy, D., and Latif, A. (2019). Official methodology used for poverty estimation based on the Bangladesh Household Income and Expenditure Survey 2016/17. *The Bangladesh Development Studies*, 42(2/3):289–319.

Aiken, E., Bellue, S., Blumenstock, J., Karlan, D., and Udry, C. R. (2023a). Estimating impact with surveys versus digital traces: Evidence from randomized cash transfers in Togo. Technical report, National Bureau of Economic Research.

Aiken, E., Bellue, S., Karlan, D., Udry, C., and Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603(7903):864–870.

Aiken, E., Ohlenburg, T., and Blumenstock, J. (2025). Moving targets: When does a poverty prediction model need to be updated? In *Proceedings of the 6th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, pages 117–117.

Aiken, E. L., Bedoya, G., Blumenstock, J. E., and Coville, A. (2023b). Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan. *Journal of Development Economics*, 161:103016.

Alatas, V., Banerjee, A., Chandrasekhar, A. G., Hanna, R., and Olken, B. A. (2016). Network structure and the aggregation of information: Theory and evidence from Indonesia. *American Economic Review*, 106(7):1663–1704.

Alatas, V., Banerjee, A., Hanna, R., Olken, B. A., Purnamasari, R., and Wai-Poi, M. (2019). Does Elite Capture Matter? Local Elites and Targeted Welfare Programs in Indonesia. *AEA Papers and Proceedings*, 109:334–339.

Alatas, V., Banerjee, A., Hanna, R., Olken, B. A., and Tobias, J. (2012). Targeting the poor: Evidence from a field experiment in Indonesia. *American Economic Review*, 102(4):1206–1240.

Baker, J. L. and Grosh, M. E. (1994). Poverty reduction through geographic targeting: How well does it work? *World development*, 22(7):983–995.

Bance, P. and Schnitzer, P. (2021). Can the luck of the draw help social safety nets?

Banerjee, A., Hanna, R., Olken, B. A., and Sumarto, S. (2020). The (lack of) distortionary effects of proxy-means tests: Results from a nationwide experiment in Indonesia. *Journal of Public Economics Plus*, 1:100001.

Bangladesh Bureau of Statistics (2023). *Final Report on Household Income and Expenditure Survey (HIES) 2022*. Bangladesh Bureau of Statistics (BBS), Dhaka, Bangladesh. Accessed: 2025-05-29.

Barca, V. and Hebbar, M. (2020). On-demand and up to date? Dynamic inclusion and data updating for social assistance. Healthy DEvel, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), https://health.bmz.de/studies/on-demand-and-up-to-date/.

Basurto, M. P., Dupas, P., and Robinson, J. (2020). Decentralization and efficiency of subsidy targeting: Evidence from chiefs in rural Malawi. *Journal of Public Economics*, 185:104047.

BBS (2020). Poverty Maps of Bangladesh 2016: Key Findings. Bangladesh Bureau of Statistics. Archived at https://bit.ly/BBS-PovertyMap-2016-Archive.

Beaman, L., Keleher, N., Magruder, J., and Trachtman, C. (2021). Urban networks and targeting: Evidence from Liberia. In *AEA Papers and Proceedings*, volume 111, pages 572–576.

Björkegren, D., Blumenstock, J. E., and Knight, S. (2020). Manipulation-Proof Machine Learning. *arXiv:2004.03865 [cs, econ]*. arXiv: 2004.03865.

Bloch, F. and Olckers, M. (2021). Friend-based ranking in practice. In *AEA Papers and Proceedings*, volume 111, pages 567–571.

Bloch, F. and Olckers, M. (2022). Friend-Based Ranking. *American Economic Journal: Microeconomics*, 14(2):176–214.

Blumenstock, J., Cadamuro, G., and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076.

Blumenstock, J. E. (2018). Estimating economic characteristics with phone data. In *AEA papers and proceedings*, volume 108, pages 72–76.

Brown, C., Ravallion, M., and Van de Walle, D. (2018). A poor means test? Econometric targeting in Africa. *Journal of Development Economics*, 134:109–124.

Bryan, G., Chowdhury, S., and Mobarak, A. M. (2014). Underinvestment in a profitable technology: The case of seasonal migration in bangladesh. *Econometrica*, 82(5):1671–1748.

Camacho, A. and Conover, E. (2011). Manipulation of Social Program Eligibility. *American Economic Journal: Economic Policy*, 3(2):41–65.

Cameron, L. and Shah, M. (2014). Can Mistargeting Destroy Social Capital and Stimulate Crime? Evidence from a Cash Transfer Program in Indonesia. *Economic Development and Cultural Change*, 62(2):381–415. Publisher: The University of Chicago Press.

Chi, G., Fang, H., Chatterjee, S., and Blumenstock, J. E. (2022). Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3):e2113658119.

CIESIN (2021). Global gridded relative deprivation index (GRDI), version 1.

Coady, D., Grosh, M., and Hoddinott, J. (2004). Targeting outcomes redux. *The World Bank Research Observer*, 19(1):61–85.

Devereux, S., Masset, E., Sabates-Wheeler, R., Samson, M., Rivas, A.-M., and Te Lintelo, D. (2017). The targeting effectiveness of social transfers. *Journal of Development Effectiveness*, 9(2):162–211.

Dutrey, A. P. (2007). Successful targeting? Reporting efficiency and costs in targeted poverty alleviation programmes.

Engelmann, G., Smith, G., and Goulding, J. (2018). The unbanked and poverty: predicting area-level socio-economic vulnerability from m-money transactions. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1357–1366. IEEE.

Fatehkia, M., Tingzon, I., Orden, A., Sy, S., Sekara, V., Garcia-Herranz, M., and Weber, I. (2020). Mapping socioeconomic indicators using social media advertising data. *EPJ Data Science*, 9(1):22.

Filmer, D. and Pritchett, L. H. (2001). Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of India. *Demography*, 38(1):115–132.

GiveDirectly (2022). Canva partnership tackling extreme poverty in Malawi.

Goodhart, C. (1975). *Monetary Relationships: A View from Threadneedle Street.* University of Warwick.

Han, H. and Gao, Q. (2019). Community-based welfare targeting and political elite capture: Evidence from rural China. *World Development*, 115:145–159.

Hanna, R. and Olken, B. A. (2018). Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. *Journal of Economic Perspectives*, 32(4):201–226.

Hillebrecht, M., Klonner, S., and Pacere, N. A. (2023). The dynamics of poverty targeting. *Journal of Development Economics*, 161:103033.

Houssou, N. and Zeller, M. (2011). To target or not to target? the costs, benefits, and impacts of indicator-based targeting. *Food Policy*, 36(5):627–637.

ILO (2021). World social protection report 2020–22: Social protection at the crossroads–in pursuit of a better future.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Jiang, X., Lim, L.-H., Yao, Y., and Ye, Y. (2011). Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244.

Kahneman, D. and Krueger, A. B. (2006). Developments in the measurement of subjective well-being. *Journal of Economic perspectives*, 20(1):3–24.

Kilic, T., Serajuddin, U., Uematsu, H., and Yoshida, N. (2017). Costing household surveys for monitoring progress toward ending extreme poverty and boosting shared prosperity. *World Bank Policy Research Working Paper*, (7951).

Kshirsagar, V., Wieczorek, J., Ramanathan, S., and Wells, R. (2017). Household poverty classification in data-scarce environments: A machine learning approach.

Lopez, J. (2020). Experimenting with poverty: The SISBEN and data analytics projects in Colombia.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1(Supplement C):19–46.

McBride, L. and Nichols, A. (2018). Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review*, 32(3):531–550.

Mukerjee, A. N., Bermeo, L. X., Okamura, Y., Muhindo, J. V., and Bance, P. G. A. (2023). Digital-first Approach to Emergency Cash Transfers: Step-kin in the Democratic Republic of Congo. *World Bank Working Paper Series*, (181798).

Murray, C. J., Evans, D. B., Acharya, A., and Baltussen, R. M. (2000). Development of WHO guidelines on generalized cost-effectiveness analysis. *Health economics*, 9(3):235–251.

Noriega-Campero, A., Garcia-Bulle, B., Cantu, L. F., Bakker, M. A., Tejerina, L., and Pentland, A. (2020). Algorithmic targeting of social policies: fairness, accuracy, and distributed governance. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 241–251.

OECD (2013). *OECD framework for statistics on the distribution of household income, consumption and wealth.* Organisation for Economic Co-operation and Development Publishing.

Premand, P. and Schnitzer, P. (2021). Efficiency, legitimacy, and impacts of targeting methods: Evidence from an experiment in Niger. *The World Bank Economic Review*, 35(4):892–920.

Schnitzer, P. and Stoeffler, Q. (2022). Targeting for social safety nets: Evidence from nine programs in the Sahel. *Available at SSRN 4017172.*

Selbst, A. D. and Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242.

Smythe, I. S. and Blumenstock, J. E. (2022). Geographic microtargeting of social assistance with high-resolution poverty maps. *Proceedings of the National Academy of Sciences*, 119(32):e2120025119.

Stiglitz, J. E., Sen, A., Fitoussi, J.-P., et al. (2009). Report by the commission on the measurement of economic performance and social progress.

Sumarto, S., Satriawan, E., Olken, B. A., Banerjee, A., Tohari, A., Alatas, V., and Hanna, R. (2025). Community targeting at scale. Working Paper 33322, National Bureau of Economic Research.

Tiecke, T. G., Liu, X., Zhang, A., Gros, A., Li, N., Yetman, G., Kilic, T., Murray, S., Blankespoor, B., Prydz, E. B., et al. (2017). Mapping the world population one building at a time. *arXiv preprint arXiv:1712.05839*.

Trachtman, C., Permana, Y., and Sahadewo, G. (2022). How much do our neighbors really know? The limits of community-based targeting. *University of California, Berkeley Working Paper*.

Walmsley, J. (2021). Artificial intelligence and the value of transparency. *AI & SOCIETY*, 36(2):585–595.

World Bank (2021). Bangladesh social protection public expenditure review.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., and Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature communications*, 11(1):2583.

Appendix - For Online Publication Only

A Construction of targeting approaches

A.1 Phone-based targeting

The phone-based targeting approach is implemented in a similar manner to past work on predicting poverty from mobile phone metadata (Blumenstock et al., 2015; Blumenstock, 2018; Aiken et al., 2022, 2023b). Pseudonymized mobile phone metadata (call detail records, or CDR) were shared with our research team by all four mobile network operators active in Cox's Bazar, for all phone numbers collected in our census of 100,000 households conducted in February 2023 (the census collected all phone numbers from all adult household members). These data included the following information, for five months from March 1 to July 31, 2023:

- Records of incoming and outgoing calls, including a pseudonymized identifier for the caller and recipient, date, time, and duration of the call, and GPS coordinates of the cell tower through which the call was placed and received
- Records of outgoing SMS messages, including a pseudonymized identifier for the sender and recipient, date and time of the message, and GPS coordinate of the cell tower through which the message was sent
- Records of mobile data usage, which we aggregate into the amount of mobile data (in megabytes) used by each subscriber per day

From these data sources, we calculate 1,578 "features" describing mobile phone use for each pseudonymized phone number in the dataset. We use open source python library cider[30] to calculate these features, which include information on call and text frequency, heterogeneity in contact networks, recharge patterns, mobility traces based on cell tower usage, and more (see cider's documentation for a complete list of features).

Next, we match the mobile phone features to the census and household survey, which are used to train our machine learning models and conduct the evaluation. For households that provided only a single phone number in the census (69% of households), each household is matched to the mobile phone metadata from the phone number provided. For households with multiple phone numbers recorded in the census (25% of households), the mobile phone

[30] https://global-policy-lab.github.io/cider-documentation/

metadata from the most senior member of the household is used.[31] The remaining 6% of households either did not provide a phone number of provided a phone number that did not produce any records in the March 1 - July 31 time period. These households were not included in the training of the ML model, and their phone-based poverty rankings were considered missing in the targeting evaluation, and therefore they are targeted last by the phone-based targeting approach. To build intuition, Table S2 shows the mobile phone features most correlated with measures of poverty from the survey: for example, the mean recharge amount is the feature most correlated with per capita consumption expenditures ($\rho = 0.19$), the PPI ($\rho = 0.23$), and the asset index ($\rho = 0.23$).

The dataset of mobile phone metadata features matched to poverty "labels" from the household survey (N = 4,820) is used to train the ML model. We train and evaluate the machine learning pipeline in the same way that we train and evaluate other ML-based targeting approaches, like the PMT (see Appendix A.3): We divide the matched features - household survey dataset (N = 4,820) into a 75% training set and 25% test set. We train the model to predict per capita consumption expenditures on the training set using the gradient boosting model available in cider, with hyperparameters selected via three-fold cross validation. The main ML model is trained to predict log-transformed per capita consumption expenditure; we also test models to predict the PPI and asset index. We then produce predictions on the test set, which are used for evaluation. As with the other targeting methods, we repeat the process for 100 different random train-test splits, to produce confidence intervals in our downstream targeting evaluation. Sample weights are used in both training and evaluation.

In the phone-based targeting approach, households are targeted from poorest to richest based on their phone-based poverty prediction. Households without phones are targeted last.

A.2 Community-based targeting (CBT)

Community-based targeting (CBT) exercises were conducted in each of the 180 neighborhoods included in our study, following the protocol used by BRAC. The CBT protocol is summarized as follows. First, in neighborhoods of more than 100 households, enumerators split

[31] An alternative approach to phone-based targeting for households with multiple phones would be to aggregate together poverty predictions from all phones to obtain a predicted measure of household poverty. Figure S14 shows the overall targeting accuracy of phone-based targeting when the ML model is trained on data from all phone numbers provided, and predictions are aggregated together for households with multiple phones (for single-phone households, the single prediction for that household's phone is still used). Three approaches to aggregation are tested: taking the average predicted poverty score, the minimum score, and the maximum score, as well as the status quo approach of taking the poverty prediction for the most senior household member. While there is little difference in the accuracy of these aggregation approaches — at least partly because relatively few (25%) households provided multiple phone numbers — taking the minimum score has lower targeting accuracy (31%) than the other three options (34%).

neighborhoods into contiguous segments of 50-100 households and conducted separate CBTs in each. Enumerators worked with senior community members to identify 12-25 households to join the meeting, inviting households from all walks of life and ensuring participation from women, students, farmers, businessmen, and laborers. Each meeting began with a "social mapping" exercise in which a community map was drawn with each household identified by name and occupation. Meeting attendees then worked together to rank the socio-economic status of all households in the community by placing index cards representing each household on a string in the order of their socio-economic status.[32] To make the CBT exercises consequential, participants were informed at the start of the meeting that the 20% poorest-ranked households would receive a one-time cash transfer of 1,000 Taka ($31.88 USD PPP) following the meeting.

The normalized ranking within each village is used to identify the poorest households for our community-based targeting method. The implicit assumption of this approach is that poverty distributions across villages are comparable. Households that are not ranked in the community-based targeting approach (0.4% of households) are considered to be targeted last for benefits by the CBT approach.

A.3 Proxy means test (PMT)

The PMT implementation follows standard approaches in the literature (Hanna and Olken, 2018; Brown et al., 2018). We use the following demographic and housing-related variables as PMT predictors:

- **Household head demographic variables:** Age, gender, marital status, highest level of education, worked in past seven days, disability status
- **General household demographic variables:** Household size, number of children under 10, number of children under 18, highest education level of any household member, union of residence
- **Housing variables:** Number of rooms, has a kitchen, has a stove, has electricity, has a toilet, ownership status of house, ownership status of land, main material of roof, main material of walls
- **Asset ownership variables:** TV, fridge, fan, stove, furniture, cell phone, solar panel, bicycle, rickshaw, vehicles, crop inventory, poultry, goats, cows, unpowered agricultural

[32] Our instructions to the meeting participants were based on the CBT used by BRAC, and noted, "... we will ask you to conduct a ranking of households in your community based on their socio-economic status."

equipment, powered agricultural equipment, fishing nets, non-engine-powered boat, engine-powered boat, business assets, owned place of business, owned dwelling, owned residential land, owned agricultural land, cash on hand

Continuous variables are scaled to a 0-1 range and winsorized with a 99% limit. Categorical variables are converted into a set of mutually exclusive dummy variables; we combine any categories that make up less than 1% of observations into a generalized "other" category for each variable.

We then fit a model to predict log-transformed per capita consumption from these input variables on the training set, and produce predictions on the test set (separately for each train-test split). We experiment with four options for the machine learning model underlying the PMT:

- **Simple linear regression:** Implemented with Python's statsmodels API via weighted least squares. We fit the regression model on the training set, and produce predictions for the test set.

- **Linear regression with step-wise forward selection of predictor variables:** For this option, the training set is again divided into a 50% true training set and a 50% validation set. We implement stepwise forward selection on the training set – that is, we search across all predictor variables to find the single best predictor of consumption (based on R^2 score on the test set), we then search across all remaining predictors to add a second for a two-predictor model, and continue adding predictors until the test-set accuracy decreases with additional predictors. Once this stopping criterion is met and the predictor subset is identified, we use Python's statsmodels API (via weighted least squares) to fit a final simple linear regression using only this subset of predictors on the entire training set, and produce predictions for the test set.

- **LASSO regression:** LASSO regression uses a regularization term to automatically perform feature selection to avoid overfitting to the training set. We implement the LASSO with scikit-learn's Lasso model, and tune the regularization parameter using three fold cross validation on the training set.

- **Random forest:** We use scikit-learn's RandomForestRegressor model, and tune hyperparameters via three fold cross validation on the training set. The ensemble size is chosen from [50, 100] and the maximum tree depth is chosen from [2, 4, 8].

Overall, we generally observe similar predictive performance of these different PMT variants: the LASSO is best with average $R^2 = 0.38$ (standard deviation 0.02), followed by OLS also at $R^2 = 0.38$ (standard deviation 0.03), then stepwise forward selection at $R^2 = 0.37$ (standard deviation 0.03), and finally the random forest at $R^2 = 0.33$ (standard deviation 0.02). In our main results we therefore show only the LASSO results, but in our supplementary results we show all four PMT variants. In general, these R^2 values are on the low end in comparison to reported R^2 values for PMTs elsewhere: for example, Brown et al. (2018) report R^2 values ranging from 0.32 in Ethiopia to 0.64 in Burkina Fasso and Hanna and Olken (2018) report R^2 values of 0.53 in Indonesia and 0.66 in Peru. One explanation for the low PMT R^2 in our context is the subnational and highly geographically concentrated nature of our survey — these other PMTs were trained and evaluated at a nationwide scale.

Figure S15 Panel A shows the PMT (using a LASSO regression) distribution for one example train-test split.

A.4 Geographic targeting

Bangladesh's most recent official poverty map is only available at the upazila (sub-district) level. (BBS, 2020) With only three upazilas in our household survey, geographic targeting at the upazila level is not a relevant targeting approach in our setting. We therefore use high-resolution poverty maps based on nontraditional data sources to simulate geographic targeting.

Our satellite-based poverty estimates come from the gridded Global Deprivation Index (GDI) released by NASA/Columbia's SEDAC center last year (CIESIN, 2021). The GDI uses subnational administrative datasets and gridded earth observation datasets to produce an "index of relative deprivation" in approximately a 1km global grid. The index consists of six components: (1) child dependency ratio from gridded population of the world datasets, (2) infant mortality rates from the global subnational infant mortality rates dataset, (3) the subnational human development index from the Global Data Lab, (4) the ratio of built-up to non-built-up area using data from Facebook's High Resolution Settlement Layer and OpenStreetMap, (5) nighttime lights intensity from VIIRS, and (6) changes in nighttime light intensity from 2012 to 2020. The average of these six components makes up the GDI.[33]

We aggregate the GDI to three different geographic levels, for three variants of geographic targeting. For each level of aggregation, we talk the weighted average of GDI tiles contained (or partially contained) within the boundary, with weights determined by the population contained within the tile. The population density layer is also based on remote sensing and released by Meta (Tiecke et al., 2017). The three levels of aggregation are as follows, ordered

[33] We prefer the GDI to the Relative Wealth Index (RWI) released by Meta (Chi et al., 2022) that has been used in previous work on remote sensing-based geographic targeting (Aiken et al., 2022; Smythe and Blumenstock, 2022) because RWI data are missing for much of the eastern portion of Cox's Bazar.

from lowest to highest resolution:

- **Unions:** We use publicly available union shapefiles[34] to aggregate the GDI to the union (admin-4) level. These shapefiles do not contain urban wards, the admin-4 unit in urban areas. To obtain extents for the eight wards in our census dataset, we use the same process used to identify village and neighborhood extents, described in detail below. There are 23 admin-4 units in total for households in our household survey: 10 in Ramu, 5 in Ukhia, and 9 in Teknaf, ranging from 0.05-137 square km (median of 21 square km). 97% of admin-4 units overlap with at least one GDI tile, with the median containing 28 tiles. For the remaining 3% of admin-4 units, the poverty level assigned is that of the closest GDI tile.

- **Villages:** To our knowledge, there are no publicly available village shapefiles for Bangladesh. To calculate the boundary of each village, we take the convex hull of all GPS coordinates recorded for households in that village in the census. Any household that is not closer than 2km to at least 20 other households in the same village is considered an outlier, and not included in the process of calculating the convex hull. We then take the weighted average of all GDI tiles overlapping the convex hull of the village. There are 105 villages in total in our household survey: 37 in Ramu, 25 in Ukhia, and 43 in Teknaf, ranging from 0.01-27 square km (median of 0.70 square km). 96% of villages contain at least one GDI tile, with the median containing four tiles. For the remaining 4% of villages, the poverty level assigned is that of the closest GDI tile.

- **Neighborhoods:** We repeat the same process to identify the convex hull of each neighborhood based on GPS coordinates recorded in our census. Again, any household that is not closer than 2km to at least 20 other households in the same neighborhood is considered an outlier, and not included in the process. There are 180 neighborhoods in total in our household survey: 60 in Ramu, 60 in Teknaf, and 60 in Ukhia, ranging from less than 0.01 square km to 3 square km. 94% of neighborhoods overlap with at least one GDI tile, with the median containing two tiles. For the remaining 6% of neighborhoods, the poverty level assigned is that of the closest GDI tile.

Figure S16 shows the poverty maps produced through this technique, at the union, village, and neighborhood level.

[34] `https://data.humdata.org/dataset/cod-ab-bgd`

A.5 Poverty probability index (PPI)

We implement the Bangladesh PPI released by Innovations for Poverty Action, which was calibrated using the 2016-17 Household Income and Expenditures Survey (which is nationally representative). The PPI consists of a scorecard of ten questions: district (Cox's Bazar for all our households), housing members, children under ten, the highest grade completed by anyone in the household, ownership of a bicycle, refrigerator, and fan, construction material of household walls, electricity connection, and type of toilet used. In our data, all questions except for electricity and the number of children under 10 were collected in the census (the remaining two were collected in the household survey). The final score represents the probability that the consumption of the household in question falls below the national poverty line. The mean PPI among our surveyed households is 54.18, with a standard deviation of 12.97. Figure S15 Panel B shows the distribution of the PPI in our household survey.[35]

A.6 Asset index

The asset index is constructed following Filmer and Pritchett (2001). We use principal components analysis (PCA, implemented with Python's wpca package) to obtain a vector representing the direction of maximum variation in asset ownership among each of the 26 assets collected in the survey (where each asset variable is a binary indicator for ownership of the asset). The PCA is fit using only the training set; we then project the data for each test set household onto this vector. Across 100 train-test splits, the first principal component explains on average 18.14% of the total variation in asset ownership (standard deviation of 0.22%). Figure S15 Panel C shows an example asset index distribution from one of the train-test splits.

A.7 Peer rankings

In our household survey, we included a peer rankings module. In this module, each household was asked to rank eight randomly selected households from their neighborhood, as well as themselves. Randomization was done to ensure that every household ranked eight other households, and each household would be ranked eight times. For each household j ranked by i, we asked i how well they knew household j (on a scale of 1-4),[36] and we asked i to rank

[35]The PPI is similar to other categorical or scorecard-based targeting approaches. A paticularly relevant one in the Bangaldesh setting is IFPRI's categorical targeting approach (Ahmed and Bakhtiar, 2023); however we do not include this approach in our analysis because it was designed for urban areas only.

[36]The options were, "1. Among my closest relatives; 2. Know very well; 3. Know a little bit; 4. Do not know"

the absolute welfare of j on a five-category scale.[37] Finally, we asked i to provide a relative ranking, from worst-off to best-off, of the eight households in i's list.

In the peer ranking module, if a household reported not knowing one of the households it was supposed to rank, they were not required to rank that household. As such, most households are not ranked exactly eight times — the median household is ranked four times by neighbors (plus once by themselves). 97% of households have at least one neighbor ranking, and 93% of households have at least one high-confidence neighbor ranking. Figure S17 shows the distribution of the number of times each household was ranked.

To determine the final peer ranking of each household j, we aggregate rankings by taking the average ranking of all households i that ranked j. We test six different approaches: one for each of the different types of ranking (absolute vs. relative), and one for each of three different variants based on the strength of the i-j connection: one that uses all rankings; one that only uses neighbor rankings (i.e., by dropping self-ranking); and one that uses self-rankings and rankings of neighbors only if i reported knowing j "very well" or better. We also test an absolute poverty ranking that uses only the self-ranking. Figure S18 compares the accuracy of each of these approaches to using the peer rankings data to target cash transfers.

Absolute welfare estimates. To obtain the community-based absolute welfare rating for each household, we simply take the average of the welfare ratings of all other households that rated it. Again, we produce three variants of this estimate: One for all ratings (including self-ratings), one for only neighbor ratings, and one for only high-confidence neighbor ratings (plus the self rating). We also look at using the self rating alone.

Relative welfare estimates. To obtain the community-based relative welfare ranking for each household, we use the HodgeRank algorithm, originally introduced by Jiang et al. (2011), and recently used for community-based targeting analysis by Bloch and Olckers (2021). Hodgerank aggregates pairwise comparisons between items (in our case, households), where each pairwise comparison represents an assessed "distance" between the two items (in our case, the difference in wealth between the two households). To produce these assessed distances, for each ranked household, we take the distance between rankings for each pair of households, normalized by the total length of the ranking. Following Bloch and Olckers (2021), if any pairwise comparison appears more than once in our dataset (16% of pairwise comparisons), we use the average (normalized) difference in ranking as input to the algorithm.

[37] The prompt was, "Is the household headed by [NAME OF HOUSEHOLD HEAD] a family that has the most, a family that has a lot, a family that has neither a lot nor a little, a family that has little, or a family that has the least?"

The Hodgerank algorithm has the benefit of a "goodness of fit" measure describing the degree of local inconsistency in the underlying rankings relative to the aggregate ranking. In our analysis, local inconsistency ranges from 0.31 when all rankings are used, to 0.28 when only neighbor rankings are used, to 0.23 when only high-confidence neighbor rankings and self-rankings are used. The inconsistency values reported by Bloch and Olckers (2021) using data from Alatas et al. (2016) in Indonesia tend to be lower: the median inconsistency across neighborhoods is 0.15.

For both the welfare rankings we assume that any household without a ranking is considered richer than all ranked households for the purposes of targeting — that is, they would be missed in targeting based on community rankings. Figure S19 shows the distributions of the six targeting rankings.

A.8 Further Details on Data Privacy

In developing our phone-based targeting algorithm, we adopted a comprehensive set of security protocols to safeguard data confidentiality. First, we obtained informed consent from surveyed households to use their survey responses and phone usage data to determine their eligibility for a cash transfer program. Second, the analysis of raw Call Detail Records (CDRs) associated with their phone numbers was conducted exclusively by telecom operators on their own premises by personnel who already had authorized access to such data. Neither the research team nor any affiliates of the project — including personnel from GiveDirectly or the Government of Bangladesh (GoB) — were granted access to the raw CDRs. Instead, the research team provided the telecom operators with a list of relevant phone numbers alongside a set of algorithmic instructions designed to extract 1,578 aggregated features from the corresponding CDRs. These features comprised summary statistics of household phone usage behaviors and thus, did not contain any data pertaining to individual calls or text messages. Finally, the telecom operators merged this feature dataset with a separate dataset containing encrypted variables derived from household surveys and fully anonymized the resulting dataset. The final anonymized dataset was securely stored on an isolated server on the premises of Aspire to Innovate (A2i, an entity of the Government of Bangladesh), where all subsequent data analyses were conducted. We submitted the data sharing protocol as part of our IRB application to the University of California Berkeley and received approval under CPHS Protocol 2023-02-16103. We also obtained explicit permission from Bangladesh Telecom Regulatory Commission to use the aforementioned phone data for our project and an additional IRB approval from the University of Dhaka.

This data sharing and handling protocol effectively minimized the potential for exposure of Personally Identifiable Information (PII) data. On one hand, it ensured that our project did not expose phone data with PII to the research team or GiveDirectly. On the other hand, the telecom operators did not have access to un-encrypted survey data. The government did not have access to either phone or survey data with PII, as they received access only to the anonymized dataset. Thus, given the robust anonymization and secure data handling practices, individual households could not be identified by the research team, GiveDirectly, telecom operators, or any government entity once the full set of survey and phone data were available together.

Once completed, the targeting algorithm identified beneficiary households from the full set of households listed in the census. A2i transmitted the list of hashed phone numbers of these selected households to the telecom operators, without disclosing any additional information. The operators then performed the de-hashing process to retrieve the original phone numbers, which were subsequently provided to GiveDirectly for the implementation of their cash transfer program.

B Simulating Counterfactual Targeting

B.1 Simulating Improved Community-Based Targeting

In our main analysis (Section 4), we find that phone-based targeting substantially out-performs community-based targeting (CBT) in Bangladesh. This raises the question: how accurate would the CBT need to be in order to out-perform phone-based targeting? To answer this question, we simulate an improved CBT by taking a weighted average of a household's CBT rank and its true consumption rank, weighting the consumption rank progressively higher to move CBT rankings closer to the correct rankings. Figure S13 Panel A reproduces Figure 8 including these simulations of the improved CBT, for four different accuracy levels. Once the CBT's accuracy substantially exceeds that of phone-based targeting (Spearman's $\rho = 0.50$, compared to 0.23 for phone-based targeting and 0.65 for PMT), the CBT is the best approach for budgets in the range of \$10-30 per household screened, using the median PMT variable cost from the literature (\$4.00).

B.2 Simulating Improved Phone-Based Targeting

Our main comparison between PMT and phone-based targeting is likewise impacted by the relative accuracy of the two methods. For example, in Togo — where phone-based targeting accuracy is higher ($\rho = 0.40$) than in Bangladesh ($\rho = 0.23$) — there is a broader scope of programs for which phone-based targeting achieves a higher utility impact than PMT (Figure 8). To more systematically show the relationship between the accuracy of phone-based targeting and the choice between phone-based targeting and PMT, we simulate improved phone-based targeting in the same way we simulate improved CBT: we take a weighted average of a household's phone-based targeting rank and its true consumption rank, weighting the consumption rank progressively higher to move phone-based targeting rankings closer to the correct rankings. Figure S13 Panels B (Bangladesh) and C (Togo) reproduce the results from Figure 8 including these simulations of phone-based targeting with higher accuracy. In both Bangladesh and Togo, when the Spearman correlation between phone-based poverty estimates and consumption is around 0.20 (as in Bangladesh), programs with budgets under \$15 per household screened should use phone-based targeting. As the correlation increases to 0.40 (as in Togo), phone-based targeting performs better for programs up to \$40 per household screened. Table S3 further illustrates the impacts of improving the accuracy of phone-based targeting, showing the budget at which aid programs should switch from phone-based targeting to PMT targeting, as a function of the accuracy of the phone-based approach.

C Beneficiary Satisfaction

We conducted a short survey in October 2024 to assess how households perceived the two cash transfer programs that had occurred in their neighborhoods. The survey included 1,100 randomly selected households from the 180 neighborhoods in which the CBT program was conducted, which were a random subset of the 200 villages in which phone-based transfers were delivered by GiveDirectly. We used a stratified random sampling approach, through which we selected four households from each of the 180 neighborhoods, one household that had received only a CBT transfer, one that had received only a phone-based transfer, and two households that had received neither transfer. We prioritized households from the baseline survey, but if there were no such households that fit the stratification criteria in a particular neighborhood, we supplemented the sample with households from the broader census.

In the survey, each respondent was first asked a set of questions about the phone-based cash transfer program, and then subsequently asked the same set of questions about the CBT program. For each program, we asked respondents (i) whether they remembered the program, and (ii) if they were a beneficiary of the program. We then asked them (iii) to describe, in their own words, how they thought eligibility was determined. Finally, we asked two questions about (iv) their perceptions of the program:

1. How satisfied were you with how the approach determined who was eligible to receive cash aid? [Not at all satisfied, Somewhat satisfied, Satisfied, Very satisfied]

2. In your opinion, was the selection process to receive cash aid in the program fair? [Yes, No]

C.1 Recall and comprehension

Of the 1,100 survey respondents, 75% remembered the phone-based "program run by GiveDirectly where cash was delivered via mobile money in January-February", and 88% remembered the CBT "program run by our survey firm where cash was delivered via mobile money and physical cash in June-July." 20% of respondents reported being a direct beneficiary of the phone-based program, and 64% reported knowing a beneficiary. A similar share of respondents reported being CBT beneficiaries (20%) or reported knowing a CBT beneficiary (69%).

While recall of the programs was high, comprehension was low, particularly for the phone-based program. Table S4 shows a sample of responses to the open-ended question they were asked about how they thought eligibility was determined for each of the two programs.

C.2 Satisfaction and perceptions of fairness

Respondent satisfaction and perceptions of fairness were substantially higher for the CBT process than for the phone-based targeting process. In particular, 58% of respondents report being "satisfied" or "very satisfied" with the CBT, while 32% were satisfied or very satisfied with the phone-based targeting process. (These estimates, as well as the others reported in this section, use sample weights to make responses representative of the beneficiary population). Likewise, 79% perceived the CBT to be fair, while 45% perceived phone-based targeting to be fair.

To better tease apart the factors that correlate with satisfaction and fairness, we regress the two measures of satisfaction on the targeting approach, with the specification:

$$\text{Satisfaction}_{i,m} = \beta\text{CBT}_{i,m} + \mu_i + \epsilon_{i,m} \tag{2}$$

In these regressions, $\text{Satisfaction}_{i,m}$ is a binary variable indicating the satisfaction (1 if satisfied or very satisfied; 0 otherwise) of respondent i for targeting method m (phone-based vs. CBT). $\text{CBT}_{i,m}$ is a binary variable that takes the value 1 for questions about the CBT and 0 for the corresponding question about phone-based targeting. We include a respondent fixed effect μ_i to isolate differences within a given respondent in satisfaction across the two different targeting methods.[38] The main coefficient of interest, β, tracks respondents' propensity to indicate greater satisfaction with the CBT relative to the same question about phone-based targeting.

In some specifications, we also include interaction terms between household characteristics X_i and the CBT dummy. These interaction terms help us understand whether certain types of respondents are systematically more likely to prefer the CBT to phone-based targeting:

$$\text{Satisfaction}_{i,m} = \beta\text{CBT}_{i,m} + \gamma(\text{CBT}_{i,m} * X_i) + \mu_i + \epsilon_{i,m} \tag{3}$$

Results in Table S5 indicate a general preference for CBT. In the first specification (columns 1 and 3), corresponding to equation (2), we observe that on average, households are 26.5 percentage points more likely to report being satisfied with the CBT process than with the phone-based process, and are 35.2 percentage points more likely to consider the CBT process fair, relative to the phone-based process. In both cases, where the specifications include respondent fixed effects but no other control variables, $p < 0.01$.

In columns 2 and 4, corresponding to equation (3), we see that the relative evaluation of CBT vs phone-based targeting varies greatly depending on respondent characteristics, such

[38] Results without respondent fixed effects are similar but less precise.

as whether this household actually received a transfer allocated by CBT or phone-based targeting. Perhaps unsurprisingly, beneficiaries' perceptions of a targeting method and program tend to be more positive if they received benefits from it (rows 2 and 3). We also observe that people who participated in the CBT process (row 4), and who are aware of the CBT program (row 5), are more satisfied with the CBT process, and more likely to perceive the CBT process as fair.

D Supplementary Figures and Tables

Figure S1: Project timeline

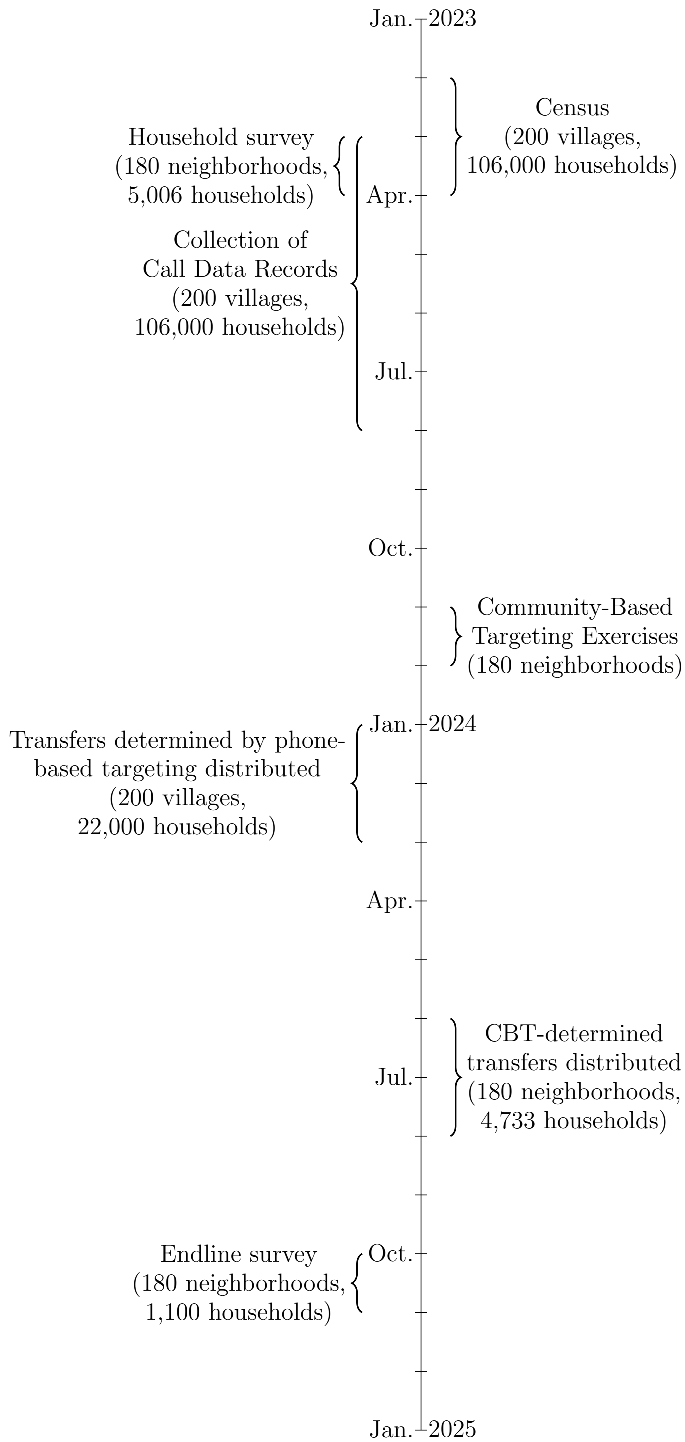
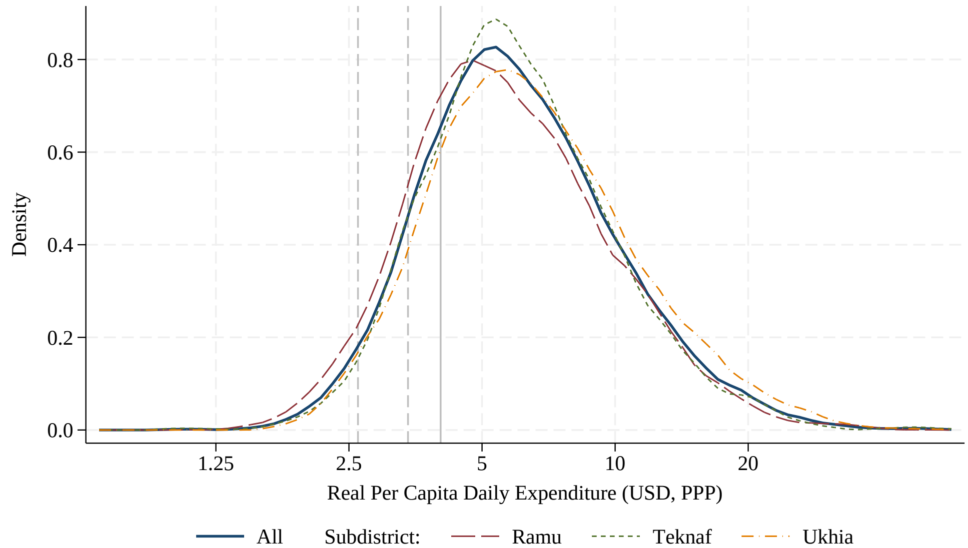
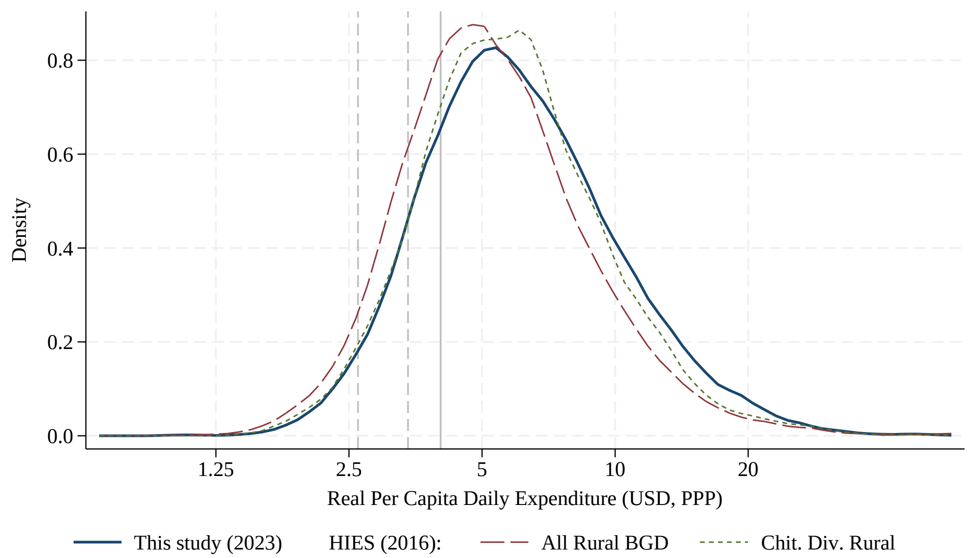


Figure S2: Density of household real per-capita daily consumption

(a) By subdistrict (upazila)

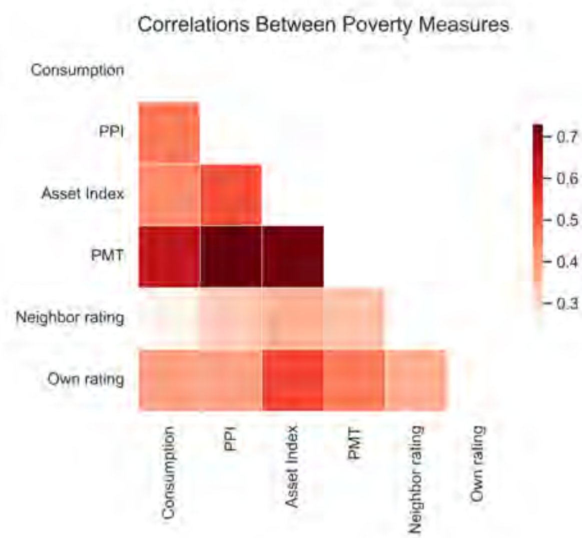


(b) Compared to 2016 HIES



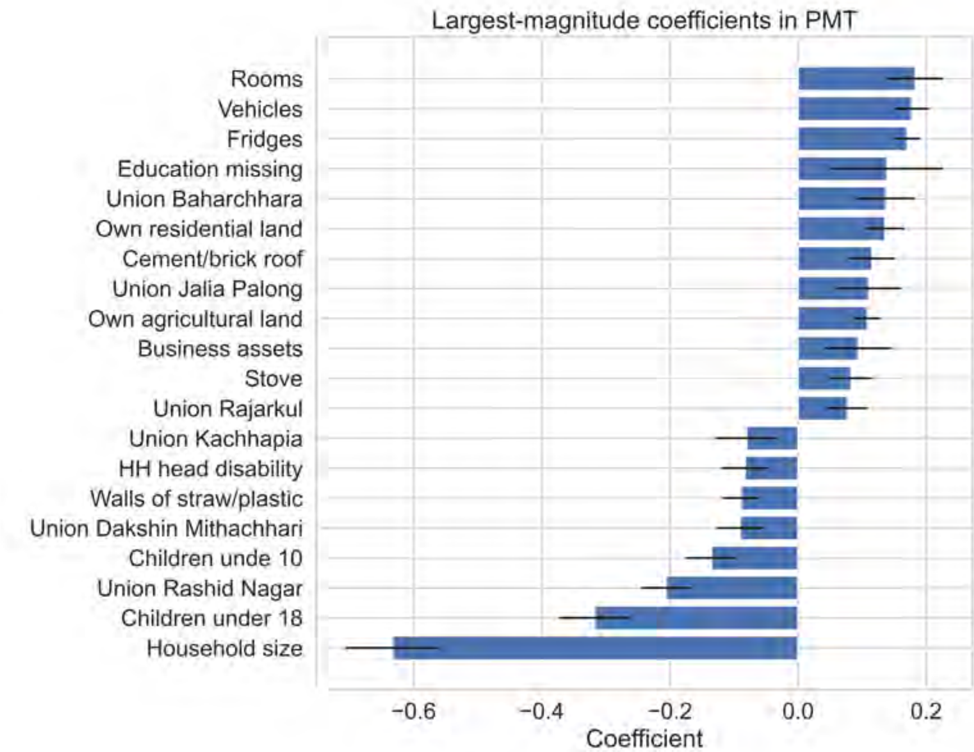
Notes: These figures plot the density of real household per-capita daily consumption in 2023 USD (PPP) from the household survey (solid line). The solid vertical line represents the 21st percentile (PPP USD 4.03). The dashed vertical lines indicate the lower and upper poverty lines for rural Bangladesh (PPP USD 2.62 and 3.40, respectively). In the top panel, we additionally plot the density by sub-district. In the bottom panel, we plot the same variable from the 2016 Bangladesh Household Income and Expenditure Survey (HIES) for two sub-groups, all rural households in Bangladesh (long dash) and rural households in Chittagong division (short dash). Our study was conducted in three sub-districts (upazilas) of Cox's Bazar district (zila) in Chittagong. Observations from the HIES are weighted using the 2016 HIES household inverse probability weights. 2016 nominal consumption in BDT is converted to 2023 BDT using the Bangladesh CPI, and then to USD at purchasing power parity at the mean 2023 PPP exchange rate for personal consumption of 30.7 BDT/USD.

Figure S3: Correlation between key poverty outcomes



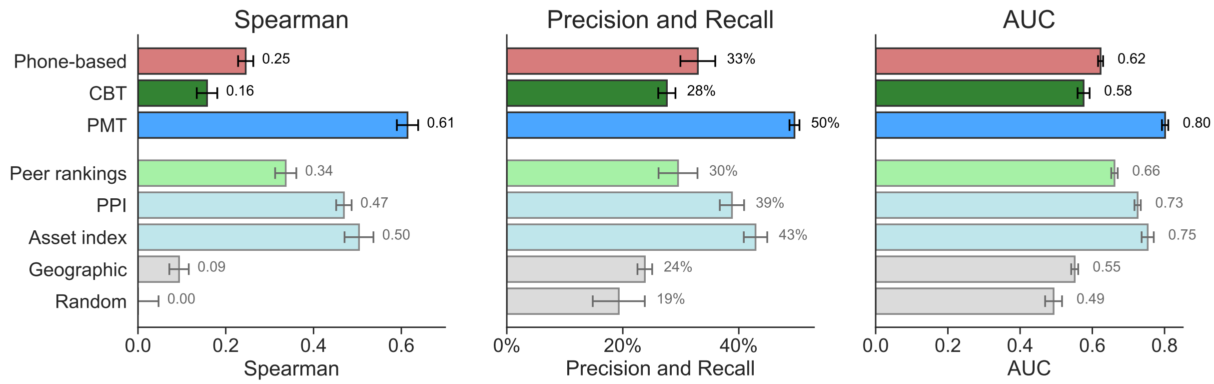
Notes: This figure displays a matrix of correlations between key poverty outcomes from the baseline household survey.

Figure S4: PMT variables with largest estimated coefficients in LASSO



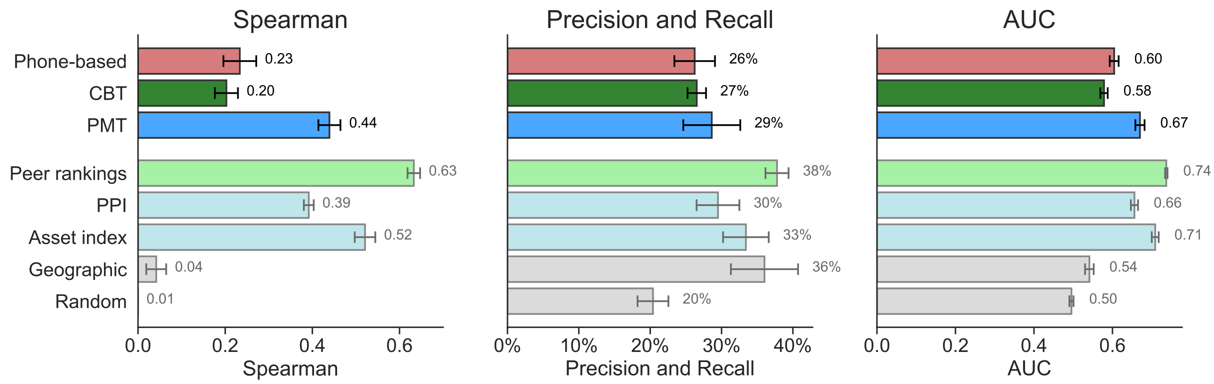
Notes: This figure displays the 20 PMT variables with the largest (in absolute value) coefficients in the LASSO estimation. A positive coefficient means that a household is predicted to have higher consumption per capita. Continuous variables are scaled to a 0-1 range. Coefficients are averaged over all 100 train-test splits, with error bars showing two standard errors above and below the mean coefficient across the 100 splits.

Figure S5: Targeting accuracy for consumption per adult equivalent



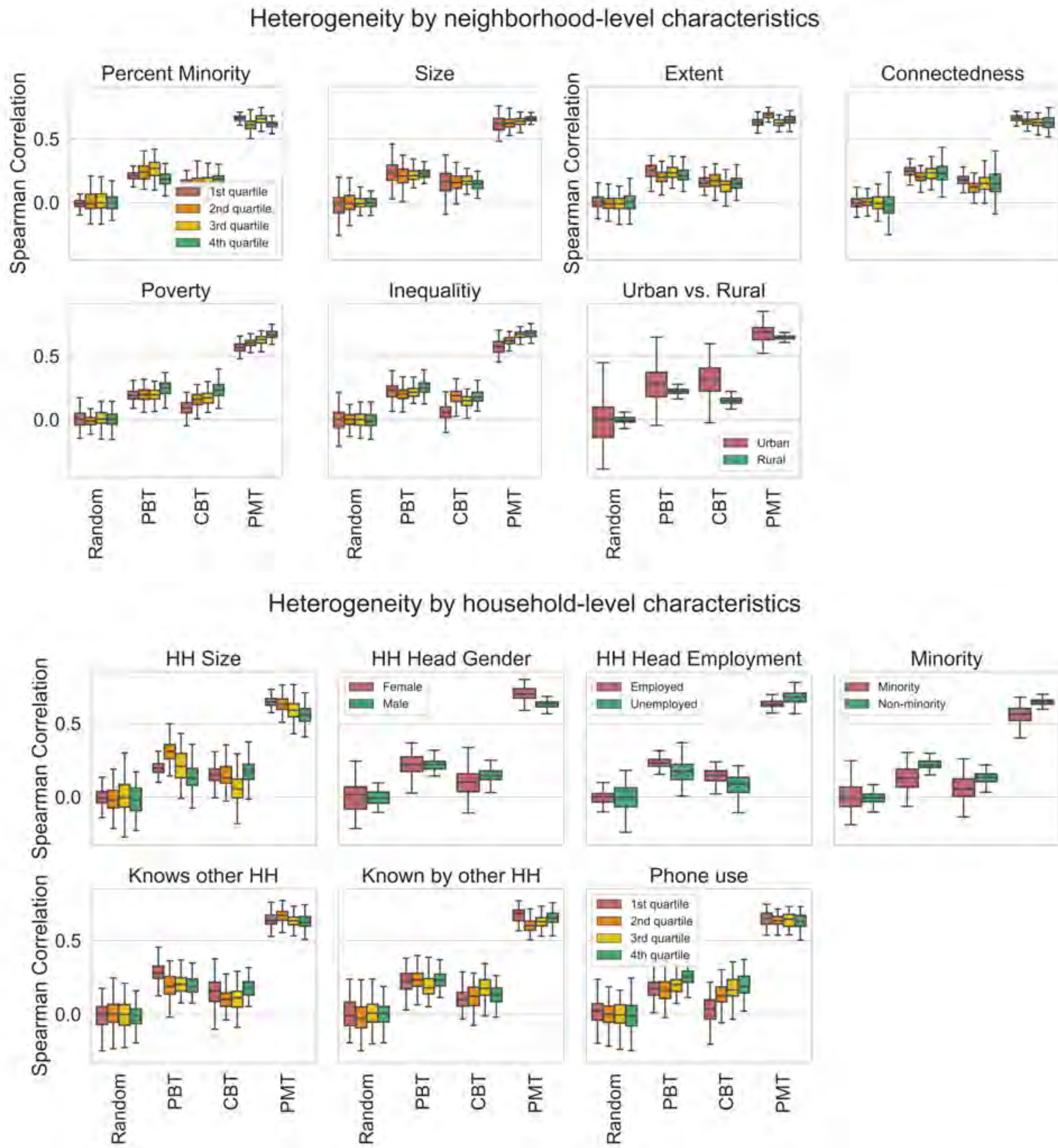
Notes: this figure reproduces our main targeting accuracy results from Figure 2 when consumption per adult equivalent, rather than per capita consumption, is used as the targeting benchmark. Adult equivalence is calculated following the OECD modified scale (OECD, 2013).

Figure S6: Targeting accuracy for identifying households that perceive themselves to be poor



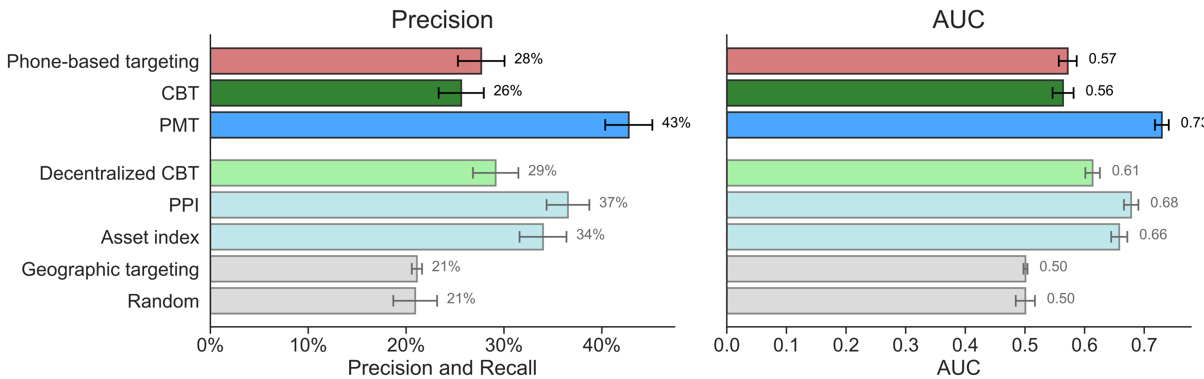
Notes: These figures show targeting accuracy when the benchmark is changed from consumption expenditures to the question, "compared to other households in your village, is your household a family that has the most, a family that has a lot, a family that has neither a lot nor a little, a family that has little, or a family that has the least?". Note that the machine learning approaches for phone-based targeting and PMT are not calibrated to predict subjective well-being: they are calibrated to predict consumption expenditures, as in Figure 2. The length of the bar depicts the mean over 100 test-train splits, with the brackets showing two standard deviations above and below the mean.

Figure S7: Heterogeneity in targeting accuracy



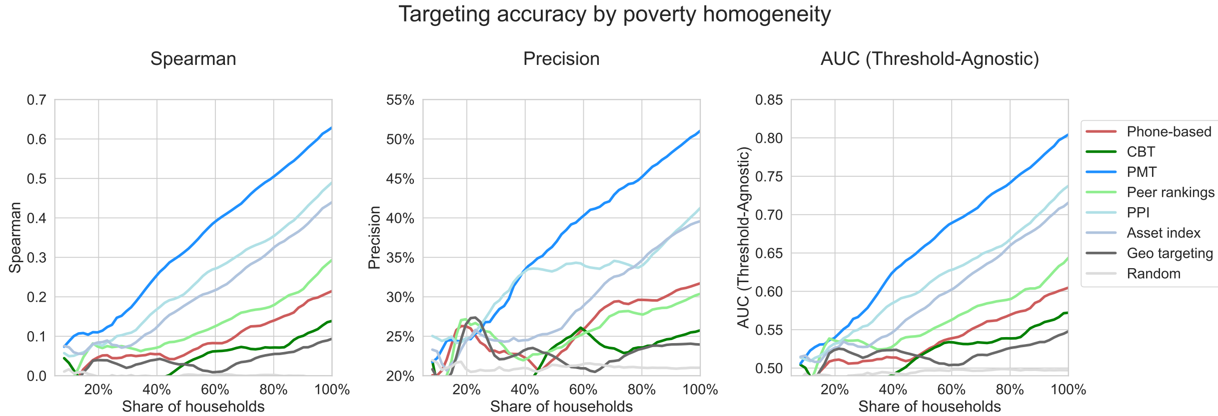
Notes: These figures show how targeting accuracy varies for different types of neighborhoods (top panel) and different types of households (bottom panel). Colors indicate neighborhood or household type. For instance, the "size" figure in the top panel indicates that the PMT outperforms all other methods (for each quartile of size, the PMT is better than other methods), and that the PMT performs slightly better on larger neighborhoods than smaller neighborhoods (the green PMT bar is higher than the red/orange/blue PMT bar). Each plot shows the distribution of Spearman correlations (over the 100 random train-test splits) for each group.

Figure S8: Targeting accuracy comparison within neighborhood



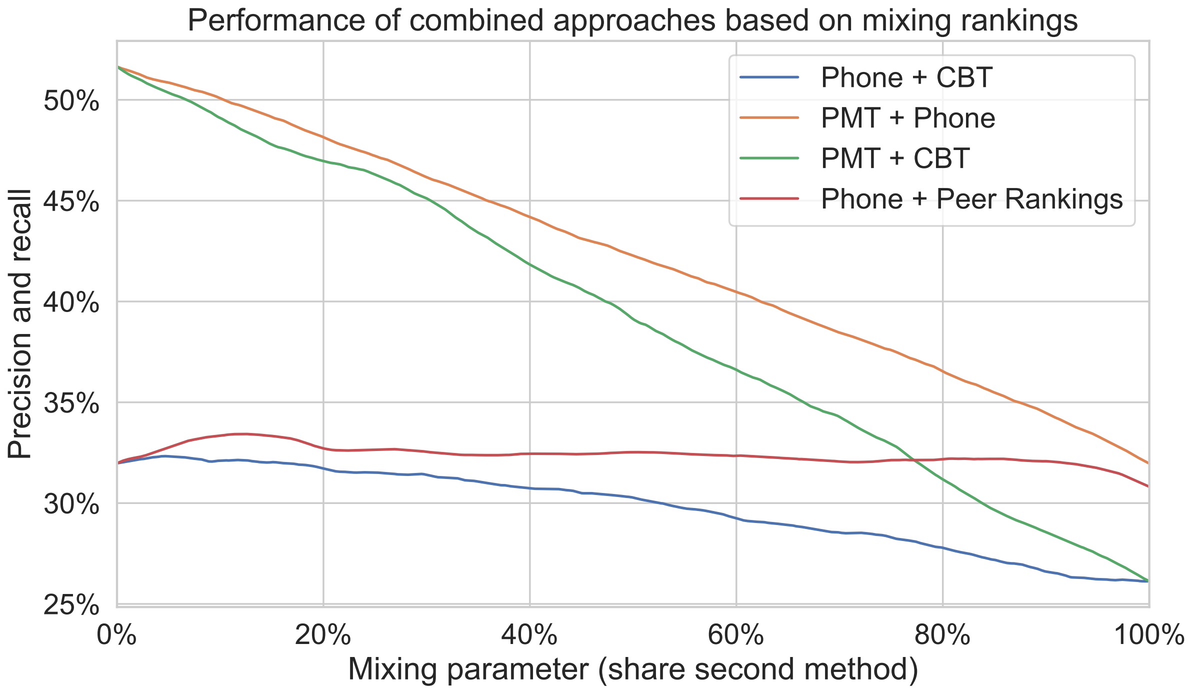
Notes: These figures compare accuracy by method for identifying the poorest households within neighborhood, rather than across the entire population, as in Figure 2. Accuracy based on precision and recall for identifying the 21% consumption-poorest households in each neighborhood (left), and area under the ROC curve (right). Accuracy is calculated over 100 random train-test splits, and error bars show two standard deviations above and below the mean for each metric.

Figure S9: Targeting accuracy by poverty homogeneity



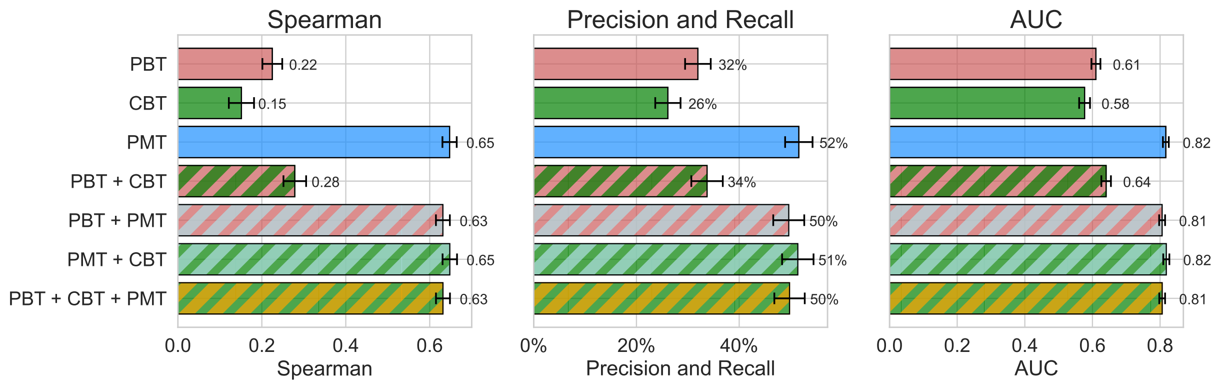
Notes: These figures compare the accuracy of each targeting method as a function of the poverty homogeneity of the population. The x-axis represents the share of households from our survey included, ranked by poverty: thus 20% indicates restricting the targeting evaluation to the 20% poorest households in our survey.

Figure S10: Accuracy of combining rankings across methods



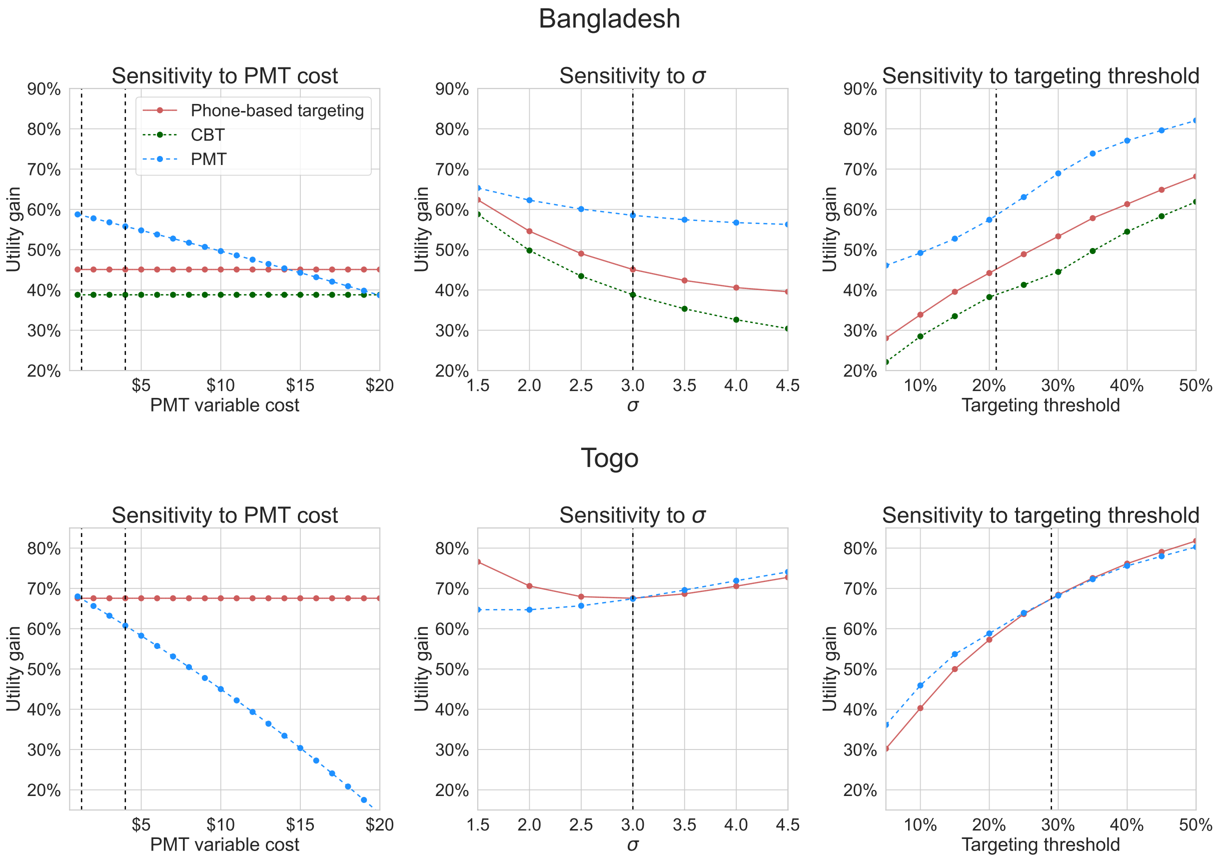
Notes: Figure shows the accuracy of targeting approaches that combine rankings from two different data sources, following the methods described in Section 3. The x-axis represents the "mixing parameter": the share of rankings that are taken from the second method in the pair (as opposed to the first). Four combined methods are tested: phone + CBT rankings (where the x-axis represents the share of rankings taken from the CBT), PMT + phone rankings (x-axis represents the share of rankings taken from phone-based targeting), PMT + CBT rankings (x-axis represents the share of rankings taken from the CBT), and Phone + Decentralized CBT (x-axis represents the share of rankings take from the peer ranking approach). Precision and recall measures are averaged over 100 bootstrap simulations.

Figure S11: Accuracy of combining data across methods



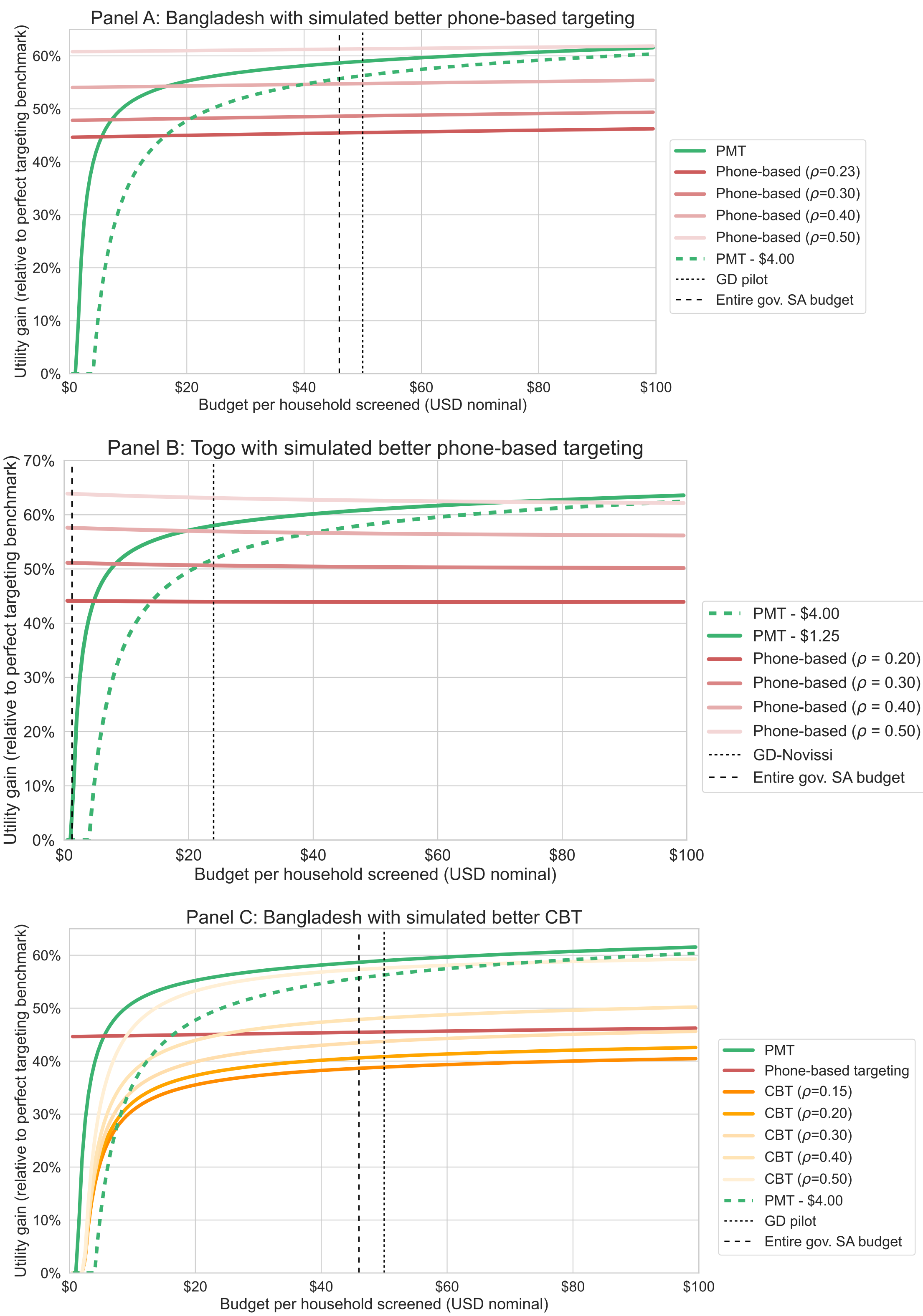
Notes: This figure displays the accuracy of ML-based approaches that combine multiple data sources into a single targeting approach, following the methods described in Section 3. "PBT", "CBT" and "PMT" indicate phone-based, community-based and proxy-means-test targeting, respectively. The horizontal length of the bar indicates the mean over 100 test-train splits, while the brackets show two standard deviations above and below the mean.

Figure S12: Sensitivity of relative performance of targeting methods



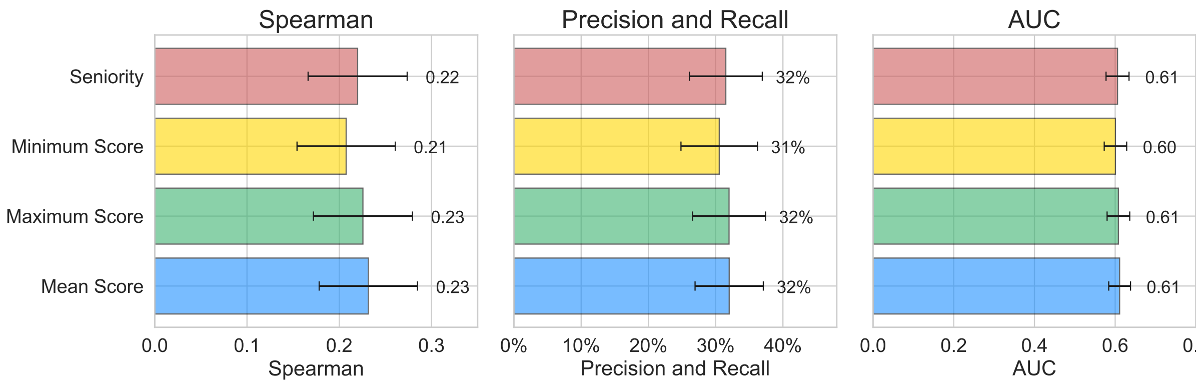
Notes: These figures show how the relative performance of each targeting method, G_m from equation (1), differs as we vary key parameters: the PMT variable cost (left in each row), σ (the coefficient of relative risk aversion for the CRRA utility function, center in each row), and the targeting threshold (share of households included, right in each row). Top: GiveDirectly program in Bangladesh (using the same data as the left panel of Figure 6). Bottom: GD-Novissi program in Togo (using the same data as the right panel of Figure 6). Dashed vertical lines indicate the values of these parameters used in the main analysis (e.g., Figures 6 and 8).

Figure S13: Sensitivity of relative performance to increases in accuracy



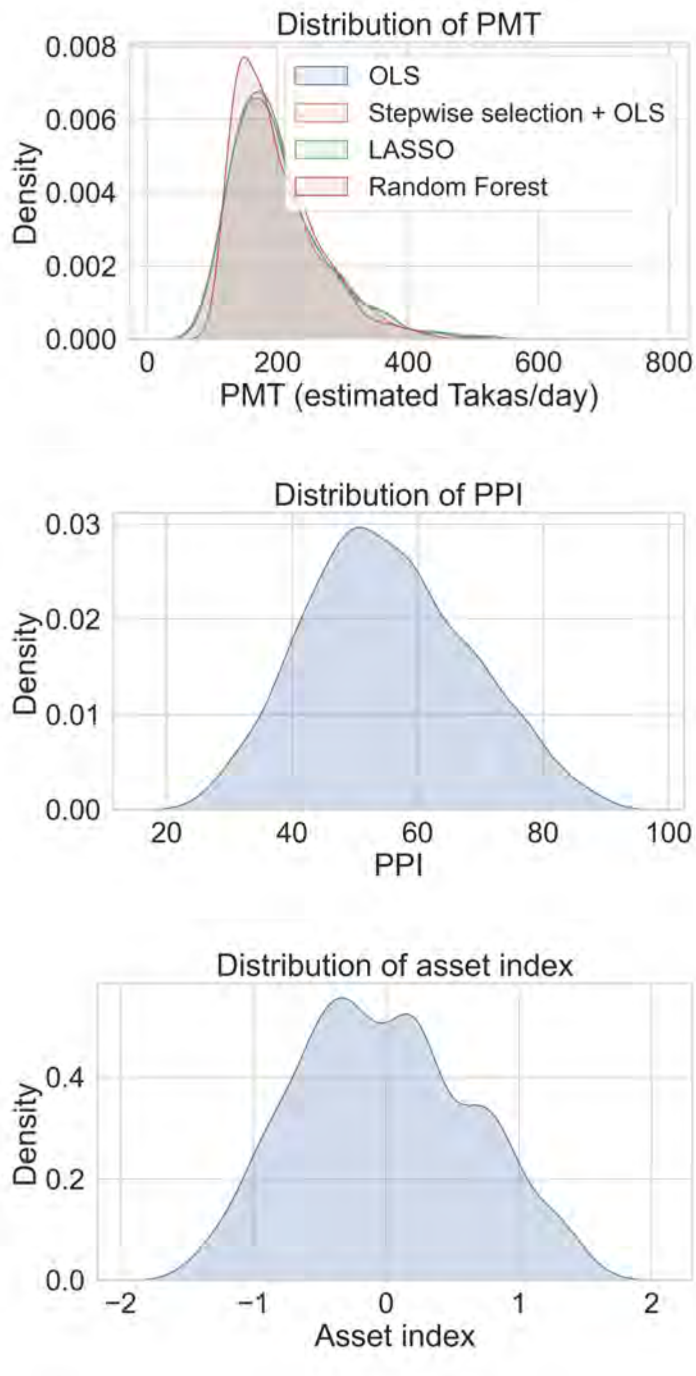
Notes: Figures replicate Figure 8 with the addition of simulated improvements in accuracy of phone-based targeting (top two panels) and community-based targeting (bottom panel). See Appendix B for details on how higher-accuracy CBT and phone-based targeting methods are simulated.

Figure S14: Households with Multiple Phones



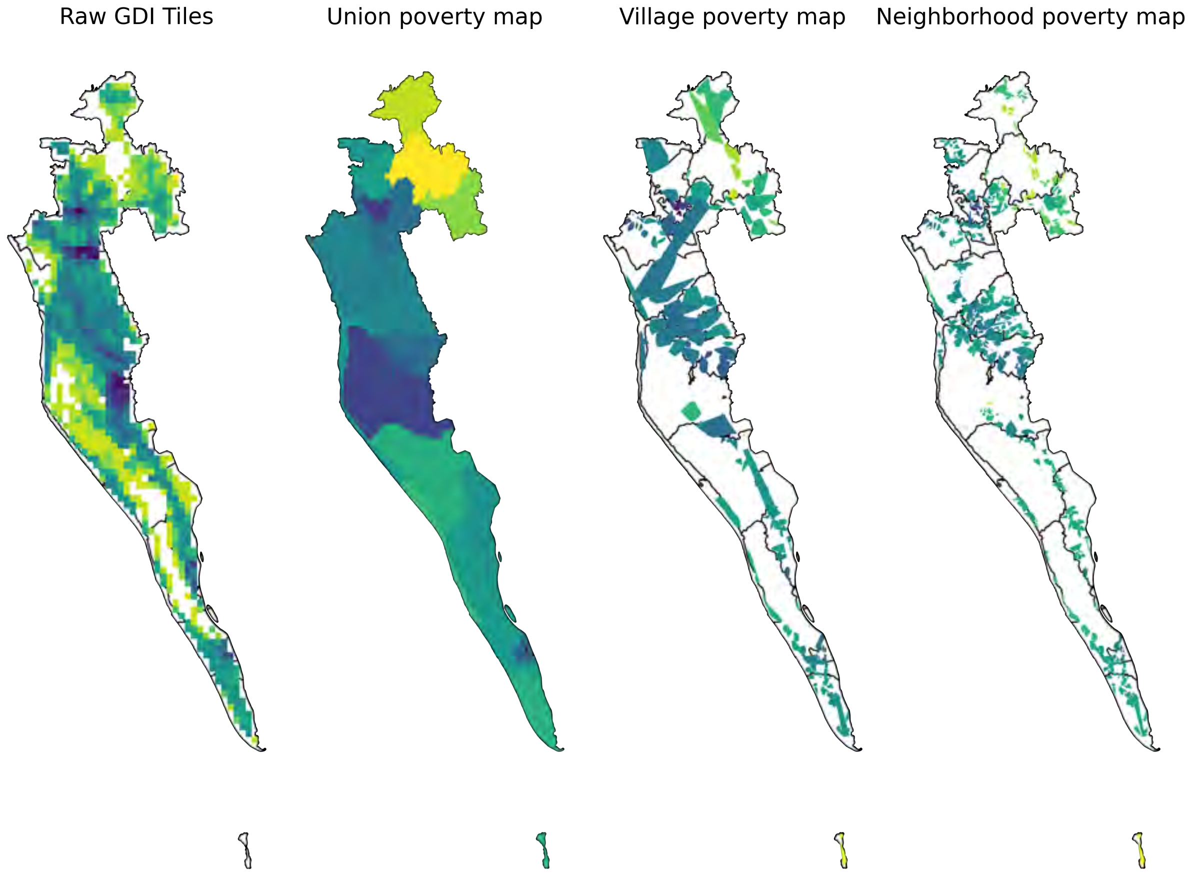
Notes: these figures present accuracy metrics for four different approaches to aggregating phone-based predictions for households with multiple phones. Targeting accuracy is calculated using the household survey dataset, as in the main targeting evaluation (Figure 2), and the approach for households providing only a single phone number (68%) or no phone numbers (3%) is unchanged. However, for households providing multiple phone numbers (29%), different approaches to aggregating poverty predictions from those phone numbers are tested: taking the prediction from the most senior member (as is implemented in the main targeting evaluations in this paper), taking the mean across predictions, taking the minimum across predictions, and taking the maximum across predictions.

Figure S15: Distribution of proxy measures



Notes: these figures present kernel density estimates showing the distribution of the PMT (left, with four versions corresponding to the four machine learning models tested), PPI (middle), and asset index (right), for one example train-test split.

Figure S16: Poverty maps

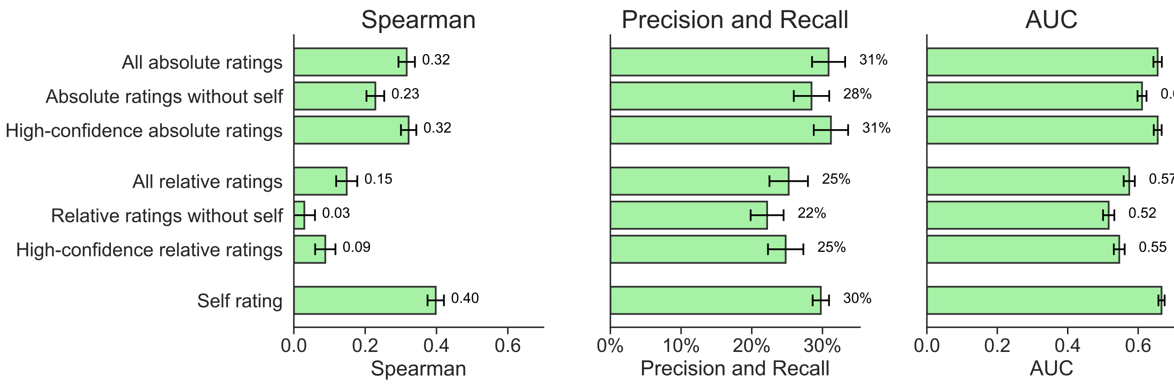


Notes: these figures display poverty maps produced by aggregating the Global Deprivation index (GDI) at the union, village, and neighborhood level, as described in Appendix A.

Figure S17: Distribution of rankings per household

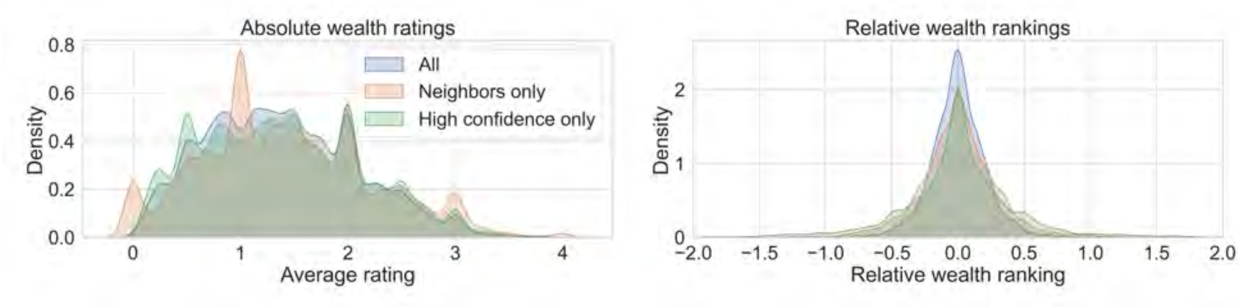
Notes: Figures display the number of rankings per household obtained in the peer rankings module of baseline survey, when keeping all rankings (left), only peer rankings (middle), and only high-confidence rankings (right).

Figure S18: Accuracy of peer ranking approaches



Notes: these figures show the accuracy of different approaches to aggregating peer ranking data for targeting. The first three sets of bars show the accuracy of approaches based on absolute ratings: each household is asked about the level of poverty of eight other households. The second three sets of bars show the accuracy of approaches based on relative rankings: each household is asked to order the eight other households in terms of poverty. The final bar shows the accuracy of self-assessments of poverty.

Figure S19: Distribution of aggregated peer rankings



Notes: these figures display distributions of aggregated peer rankings produced by averaging absolute ratings of wealth (left) and using the HodgeRank algorithm to aggregate relative rankings of wealth (right)

Table S1: Summary statistics from household survey

Variable	**Mean**
Panel A: Consumption	
Per Capita Daily Consumption (Takas)	215.27 (131.90)
Per Capita Daily Consumption (USD PPP)	6.48 (3.97)
Panel B: Additional survey-based poverty proxies	
PPI	54.75 (12.63)
Asset Index	0.00 (0.66)
PMT (Inferred Takas)	198.61 (70.54)
Panel C: Neighbor and self-assessments of poverty	
Neighbor-based poverty rating (1-5)	2.39 (0.79)
Self-assessed poverty rating (1-5)	2.22 (0.82)
Panel D: Household characteristics	
Household members	4.99 (1.97)
Number of rooms	2.67 (1.28)
Electricity access	0.82 (0.38)
Own house	1.18 (0.72)
Panel E: Household head characteristics	
Female	0.17 (0.38)
Age	41.85 (13.69)
Worked in past week	0.80 (0.40)
Has a disability	0.04 (0.20)

Notes: This table displays summary statistics from the baseline household survey. Standard deviations are shown in parentheses.

Table S2: Correlations between mobile phone features and poverty measures

	Per capita consumption		Asset Index	
	Feature	ρ	*Feature*	ρ
1	Mean recharge value	0.19	Mean recharge value	0.23
2	Max recharge value	0.16	Max recharge value	0.19
3	Min recharge value	0.14	# Call contacts (weekdays	0.18
4	# Days with mobile data use	0.13	# Call contacts (weekday, daytime)	0.18
5	# Call contacts (weekday, daytime)	0.10	# Days with mobile data use	0.17
6	# Call contacts (daytime)	0.10	# Call contacts (weekday)	0.17
7	# Call contacts (weekday)	0.10	# Call contacts	0.17
8	# of divisions visited	0.10	% of calls at night (weekday)	-0.17
9	# of subdistricts visited	0.10	% of calls at night	-0.17
10	# Call contacts (anytime)	0.10	# Weekend call contacts (daytime)	0.16
N		4,820		4,820

Notes: Mobile phone features with the strongest bivariate correlations with each poverty measure from the survey are shown, in descending order, calculated using the dataset of mobile phone features matched to household survey data (N = 4,820). A "recharge" occurs when someone adds credit (of monetary value) to the SIM card, which can be used to make calls. "Call contacts" refer to the number of unique phone numbers with which the phone made incoming and outgoing calls. "# of divisions/subdistricts" refer to the number of unique geographic jurisdictions visited by the SIM, based on observed cell tower connections. "Days with mobile data use" refers to the number of unique days that the SIM card owner is observed to use mobile data.

Table S3: Policy implications of phone-based targeting accuracy

	Low-cost PMT ($1.25)		High cost PMT ($4.00)	
Spearman	**Bangladesh**	**Togo**	**Bangladesh**	**Togo**
0.20	$4	$4	$15	$13
0.30	$6	$7	$20	$21
0.40	$17	$19	$40	$39
0.50	$98	$71	Over $100	$94
0.60	Over $100	Over $100	Over $100	Over $100

Notes: Budgets per household screened at which aid programs should switch from phone-based targeting to PMT, as a function of the accuracy of phone-based targeting accuracy (PMT accuracy is held fixed). Calculations are made using the simulated improved phone-based targeting methods from Figure 8, separately for a PMT with variable costs of $1.25 per household screened (left) and $4.00 per household screened (right).

Table S4: Example responses to questions about comprehension of eligibility criteria

Respondent	**How they thought eligibility was determined in GD program**	**How they thought eligibility was determined for CBT-based transfers**
1	I have no idea about this program. So I don't know how the eligibility criteria were determined.	Most of the rich got aid, so don't think it's appropriate.
2	I don't know how they decided because no one told me about it.	I know that the real poor families have been determined through meetings or getting together in the area.
3	Don't know how eligibility is determined under this program.	I heard they had organized a meeting but I am not familiar with the process they used for checking the eligibility for cash assistance.
4	I don't know anything about this programme. I don't know how the eligibility for the aid programme was determined.	She doesn't know the selection process, but they heard it helped 20% of poor people.
5	I don't know about this method, so I'm not sure how the selection was made.	I really liked how households were selected for cash assistance in this project because it was based on everyone's opinions and the households were surveyed accordingly.
6	I don't know how they determined it, so I have no idea about the process.	Through a lottery in a neighborhood meeting.
7	I do not know how eligibility selection program of providing aid has been set.	The poor were selected in a community gathering.
8	I don't know how eligibility for cash assistance was determined in this program; I've never heard of it.	"It was decided through the meeting with everyone's opinion because I was at the meeting so I know.
9	I am not aware of this process. But I heard that they provided money through mobile.	I think that the rich and the poor in the area, all together chose who is the richest, who is the poorest, and the poorest were given the money.
10	I don't know how the eligibility was determined but many others and I received monetary aid through the process.	She heard that assistance will be provided to 20% of people living in poverty, but no further details have been shared.

Notes: This table displays the open-ended answers provided by ten randomly selected survey responses to the question "How do you think the program determined if someone were eligible to receive cash aid? Why do you think this?"

Table S5: Beneficiary satisfaction and perceptions of fairness

	(1)	(2)	(3)	(4)
	Share satisfied or very satisfied		**Share viewing process as fair**	
CBT (relative to phone-based)	0.265***	0.094*	0.352***	0.119
	(0.027)	(0.052)	(0.038)	(0.095)
CBT * Beneficiary of CBT		0.338***		0.098
		(0.041)		(0.061)
CBT * Beneficiary of Phone		-0.444***		-0.457***
		(0.042)		(0.061)
CBT * CBT meeting participant		0.154***		0.134**
		(0.035)		(0.053)
CBT * Aware of CBT program		0.324***		0.512***
		(0.053)		(0.088)
CBT * Aware of phone program		-0.251***		-0.300***
		(0.039)		(0.061)
N	2,026	2,026	2,026	2,026

Notes: this table presents results from regressions showing correlates of beneficiary satisfaction and perceptions of fairness, for CBT-based vs. phone-based targeting approaches. The main coefficient of interest, **CBT**, indicates whether satisfaction is systematically higher for questions about the CBT relative to the same question about phone-based targeting (estimated using equation (2) (columns (1) and (3)) or equation (3) (columns (2) and (4)). All specifications include household fixed effects to isolate differences in perceptions of the two processes, for a given respondent. All regressions use sample weights to account for sample stratification.

E National Social Assistance Budgets and Scope

Table E1: Budgets and recommended targeting methods for real-world social assistance programs

Country	Year	SA budget (mill. USD)	Households (mill)	Budget per HH (USD)	Best method (BD data)	Best method (TG data)
Panel A: Social assistance in Bangladesh (based on World Bank (2021))						
Typical single program	2019	$30-311	41	$0.73-7.59	Phone-based	Phone-based
Entire SA budget	2019	$1,900	41	$46.34	PMT	Phone-based
Panel B: Social assistance elsewhere (based on World Bank ASPIRE database)						
Guinea-Bissau	2015	$0.10	0.24	$0.43	Phone-based	Phone-based
Sao Tome and Principe	2017	$0.06	0.05	$1.22	Phone-based	Phone-based
Togo	2020	$2.99	2.37	$1.26	Phone-based	Phone-based
Myanmar	2016	$12.64	9.96	$1.27	Phone-based	Phone-based
Papua New Guinea	2015	$2.17	1.31	$1.66	Phone-based	Phone-based
Madagascar	2020	$19.58	6.24	$3.14	Phone-based	Phone-based
Cameroon	2016	$10.14	3.14	$3.23	Phone-based	Phone-based
Somalia	2016	$14.78	2.11	$7.00	Phone-based	Phone-based
Tanzania	2016	$74.66	7.71	$9.68	Phone-based	Phone-based
Lao P.D.R	2021	$16.94	1.70	$9.95	Phone-based	Phone-based
Niger	2017	$46.98	2.87	$16.40	PMT	Phone-based
Zambia	2016	$41.92	2.51	$16.72	PMT	Phone-based
Congo, D.R.	2016	$252.52	13.48	$18.73	PMT	Phone-based
Uganda	2016	$119.74	6.34	$18.90	PMT	Phone-based
Samoa	2016	$0.68	0.03	$22.28	PMT	Phone-based
Rwanda	2020	$70.19	2.93	$23.93	PMT	Phone-based
Burundi	2021	$53.85	2.03	$26.48	PMT	Phone-based
Zimbabwe	2015	$67.87	2.50	$27.20	PMT	Phone-based
Kenya	2017	$287.13	10.33	$27.80	PMT	Phone-based
Ethiopia	2017	$572.40	18.24	$31.37	PMT	Phone-based
Honduras	2018	$74.61	2.34	$31.90	PMT	Phone-based
Sierra Leone	2019	$36.28	1.13	$32.24	PMT	Phone-based
Comoros	2016	$4.05	0.12	$34.59	PMT	Phone-based
Benin	2020	$59.61	1.55	$38.38	PMT	Phone-based
Central African Republic	2015	$34.76	0.88	$39.61	PMT	Phone-based
Mali	2021	$117.79	2.60	$45.29	PMT	Phone-based
Congo, Republic of	2021	$63.75	1.32	$48.47	PMT	Phone-based
Cambodia	2015	$142.59	2.90	$49.13	PMT	Phone-based
Mozambique	2021	$310.43	6.29	$49.38	PMT	Phone-based
Tajikistan	2021	$68.82	1.34	$51.54	PMT	PMT
Pakistan	2021	$1,428.92	27.41	$52.14	PMT	PMT
Guinea	2015	$74.75	1.38	$54.08	PMT	PMT
Uzbekistan	2017	$446.99	7.88	$56.73	PMT	PMT
Indonesia	2016	$3,261.57	54.49	$59.86	PMT	PMT
Angola	2021	$325.88	5.16	$63.13	PMT	PMT

Continued on next page

Table E1 – continued from previous page

Country	**Year**	**SA budget (mill. USD)**	**Households (mill.)**	**Budget per HH (USD)**	**Best method (BD data)**	**Best method (TG data)**
Moldova	2017	$105.66	1.62	$65.11	PMT	PMT
Djibouti	2019	$8.96	0.14	$65.95	PMT	PMT
Tunisia	2019	$201.15	3.04	$66.19	PMT	PMT
Afghanistan	2020	$221.51	3.33	$66.57	PMT	PMT
Burkina Faso	2016	$174.53	2.60	$67.21	PMT	PMT
Bangladesh	2019	$2,704.54	38.26	$70.68	PMT	PMT
Nepal	2021	$590.80	7.06	$83.71	PMT	PMT
Sudan	2016	$607.37	6.64	$91.49	PMT	PMT
Philippines	2016	$1,752.45	18.87	$92.85	PMT	PMT
Vietnam	2016	$2,725.22	25.03	$108.89	PMT	PMT
Kiribati	2016	$2.30	0.02	$113.22	PMT	PMT
Senegal	2015	$138.64	1.12	$123.45	PMT	PMT
Kyrgyz Republic	2018	$213.39	1.56	$136.75	PMT	PMT
India	2016	$32,815.60	228.06	$143.89	PMT	PMT
Thailand	2020	$3,903.57	26.24	$148.76	PMT	PMT
Azerbaijan	2020	$256.16	1.69	$151.46	PMT	PMT
Mauritania	2016	$115.82	0.71	$163.42	PMT	PMT
Ecuador	2015	$1,012.76	5.79	$175.05	PMT	PMT
Bhutan	2021	$26.85	0.15	$178.61	PMT	PMT
Fiji	2016	$31.06	0.17	$180.45	PMT	PMT
Jamaica	2018	$193.49	0.96	$201.41	PMT	PMT
Jordan	2021	$462.96	2.10	$220.22	PMT	PMT
Dominican Republic	2021	$942.43	4.10	$229.61	PMT	PMT
Paraguay	2017	$499.16	2.13	$233.91	PMT	PMT
Armenia	2017	$162.54	0.65	$250.14	PMT	PMT
Guatemala	2020	$419.66	1.67	$250.80	PMT	PMT
Serbia	2020	$634.94	2.47	$257.28	PMT	PMT
Lesotho	2017	$128.45	0.50	$258.14	PMT	PMT
Türkiye	2019	$6,468.55	24.81	$260.75	PMT	PMT
Bolivia	2015	$627.00	2.40	$261.16	PMT	PMT
Mexico	2020	$12,440.23	46.91	$265.20	PMT	PMT
Mongolia	2016	$242.64	0.85	$287.04	PMT	PMT
Malaysia	2016	$1,717.16	5.58	$307.54	PMT	PMT
Ukraine	2021	$10,807.33	34.57	$312.65	PMT	PMT
Belarus	2017	$1,269.63	3.98	$319.09	PMT	PMT
Egypt, Arab Republic of	2020	$8,175.32	23.88	$342.39	PMT	PMT
El Salvador	2019	$365.58	1.04	$350.26	PMT	PMT
North Macedonia	2020	$216.36	0.61	$355.75	PMT	PMT
Colombia	2020	$4,430.48	11.94	$370.99	PMT	PMT
China	2016	$117,949.79	314.61	$374.91	PMT	PMT
Peru	2021	$2,192.43	5.73	$382.47	PMT	PMT
Albania	2020	$283.54	0.73	$386.02	PMT	PMT
Algeria	2021	$3,727.17	9.35	$398.57	PMT	PMT
Iraq	2021	$2,679.22	6.62	$404.77	PMT	PMT

Continued on next page

Table E1 – continued from previous page

Country	**Year**	**SA budget (mill. USD)**	**Households (mill.)**	**Budget per HH (USD)**	**Best method (BD data)**	**Best method (TG data)**
Brazil	2018	$24,536.75	57.45	$427.08	PMT	PMT
Montenegro	2020	$83.47	0.19	$436.22	PMT	PMT
Chile	2018	$10,443.77	23.59	$442.73	PMT	PMT
Timor-Leste	2016	$87.75	0.18	$482.64	PMT	PMT
Kazakhstan	2017	$2,702.25	5.23	$516.43	PMT	PMT
Bosnia and Herzegovina	2017	$509.47	0.90	$565.60	PMT	PMT
Morocco	2021	$2,623.63	4.55	$576.49	PMT	PMT
Panama	2015	$448.96	0.78	$578.21	PMT	PMT
Uruguay	2015	$657.56	1.03	$640.20	PMT	PMT
Georgia	2020	$1,059.89	1.09	$971.09	PMT	PMT
Namibia	2018	$384.46	0.39	$975.46	PMT	PMT
Maldives	2021	$87.22	0.08	$1,031.37	PMT	PMT
South Africa	2020	$15,595.23	11.22	$1,389.44	PMT	PMT
Botswana	2019	$496.76	0.34	$1,445.32	PMT	PMT
Mauritius	2015	$391.44	0.23	$1,671.38	PMT	PMT
Trinidad and Tobago	2018	$911.51	0.46	$1,981.55	PMT	PMT

Notes: In Panel A, data on budgets are taken from World Bank (2021) and data on households is taken from the 2022 population and housing census. In Panel B, be start with data on country social protection budgets as a share of GDP in 2015-2021 from the World Bank's Aspire database (`https://www.worldbank.org/en/data/datatopics/aspire`). We match these with data on yearly GDP and population from the World Bank Open Data (`https://data.worldbank.org/`), as well as survey-based data on average household size from the Global Data Lab (`https://globaldatalab.org/`). The intersection of these three data sources contains information for 95 countries allowing us to calculate an estimate of the social protection budget per household per household screened. The preferred targeting methods are determined by on our calculations of cost-effectiveness incorporating only variable costs for targeting methods, as described in Section 4 and shown in Figure 8. The second-to-rightmost column uses our welfare calculations based on Bangladesh data to identify the best targeting method, while the rightmost column uses our welfare calculations based on Togo data.