Poverty Targeting at Scale: Algorithmic vs. Traditional Approaches*

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

May 2025

Abstract

Innovations in data and algorithms are enabling new approaches to targeting policies and interventions at scale. We compare three paradigms for poverty targeting — proxy means-testing, nominations from community members, and an algorithmic approach using machine learning to predict poverty using mobile phone usage behavior — and study how targeting accuracy and cost-effectiveness vary with the scale and scope of the program. We collect new data from Bangladesh, including mobile phone records from all major telecom operators, community-based wealth rankings conducted in 180 neighborhoods, a census of 100,000 households, and detailed consumption surveys of 5,000 households, to measure the accuracy of targeting methods at identifying poor households. While proxy-means testing is most accurate, algorithmic targeting is more cost-effective for national-scale programs where large numbers of households have to be screened. We explore the external validity of these insights using detailed survey and mobile phone records data from Togo, and cross-country information on benefit transfer programs from the World Bank.

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 University of Dhaka. We are grateful to Leo Selker for excellent research assistance. We thank our implementation partners GiveDirectly and a2i (Aspire to Innovate) and gratefully acknowledge funding for this research from GiveDirectly and the Global Innovation Fund. We also thank BIGD for support with data during an earlier design of the project. Neither the implementation partners nor the funders were involved in the analysis.

1 Introduction

Hundreds of billions of dollars are spent on social protection programs and humanitarian aid each year (ILO, 2021), so accurate targeting of these benefits is vital (Hanna and Olken, 2018). Most programs rely on traditional in-person data collection methods — such as survey-based eligibility verification and community selection of beneficiaries — to determine program eligibility. These in-person targeting approaches can be expensive to implement, and in many cases still result in targeting errors that cause large portions of eligible beneficiaries to be excluded erroneously: Coady et al. (2004) find that a quarter of poverty-targeted programs in low-income countries are regressive (providing more benefits to rich households than poor).

Novel data sources and advances in artificial intelligence have created new opportunities for deploying algorithms to identify beneficiaries remotely, lowering the implementation costs of targeting (Aiken et al., 2022b; Mukerjee et al., 2023; Lopez, 2020; Smythe and Blumenstock, 2022; GiveDirectly, 2022). These new approaches – using, for example, metadata on users' mobile phone usage patterns or satellite images of their homes and neighborhoods — are attractive because digital data can be obtained at a fraction of the cost of traditional in-person visits. Furthermore, these may be the only available data sources for remote and insecure regions where in-person surveys are prohibitively expensive or infeasible.

This paper compares a variety of approaches to identify beneficiaries for a cash transfer program in Cox's Bazar, Bangladesh, to systematically explore if and when it is preferable to target using digital data and machine learning instead of more traditional methods like proxy means tests (PMT) and community-based targeting (CBT). These comparison exercises required us to conduct a census of all 106,000 households in 200 villages, a household survey of a representative random sample of 5,000 households from 180 neighborhoods which included collecting detailed consumption data, and community-based targeting exercises in these neighborhoods. To deploy algorithmic targeting, we also obtained the complete mobile phone call data records from all consenting survey households. We partnered with *GiveDirectly* to deploy substantial transfers of 30,000 Taka (300 USD, or 955 USD PPP) to 22,000 households using the phone-based targeting method, and additional transfers to 1100 Taka based on the CBT exercise. We collect endline data on household satisfaction with the targeting process under each targeting scheme.

We used these datasets to compare PMT-based, CBT-based, and phone-based poverty targeting approaches. The PMT in our context involved predicting consumption in a training sample using household characteristics collected in our survey and deploying machine learning models to project that prediction to all non-surveyed households. Phone-based targeting used phone call and messaging features extracted from users' call detail records to predict consumption poverty. CBT involved asking community members to nominate the poorest, most deserving households in their neighborhood through group meetings. We develop and analyze measures of 'targeting accuracy' achieved by the three methods. Accuracy is measured by benchmarking against the poorest households as identified through intensive consumption expenditures surveys. We also compare the relative costs of deploying the three different strategies based on detailed cost information recorded during data collection exercises. The comparison reveals a key trade-off between the accuracy and cost of traditional versus algorithmic targeting: while the PMT is more costly than phone-based targeting, it is also more accurate. And either method out-performs the CBT in terms of *both* cost and accuracy.

We introduce a framework based on the simple idea that lowering the cost of beneficiary identification leaves more funds for transfers, to provide policymakers and administrators of social protection and humanitarian aid programs some guidance on the conditions under which algorithmic versus traditional approaches to targeting should be prioritized. We supplement these new data from Bangladesh with existing survey data and mobile phone records from Togo (Aiken et al., 2022b) to explore the external validity of our insights on the key tradeoff we identify between cost and accuracy.

Our first main finding — focusing on *accuracy* (not cost) — is that phone data based algorithmic targeting more accurately identifies consumption-poor households than the community targeting approach we test. However, both methods are substantially less accurate than proxy-means testing. Other survey-based targeting approaches — including the Poverty Probability Index (PPI, Kshirsagar et al., 2017) and a decentralized approach based on peer rankings — also outperform community-based targeting in our settings and are comparable to phone-based targeting.

Our second main finding — focusing on identifying the optimal balance between targeting accuracy and cost — is that the welfare-maximizing targeting approach for a specific social protection program depends on its scale and scope. We adapt the social welfare framework introduced by Hanna and Olken (2018) to account for targeting costs, and use this to compare the simulated welfare effects of community-based, phone-based, and PMT-based targeting approaches in both Bangladesh and Togo. We show that for programs with a relatively small budget that screen a relatively large number of households for eligibility, phone-based targeting is the preferred approach. For programs with larger budgets relative to the number of households screened (i.e., larger transfers per household), proxy-means testing is more efficient. Community-based targeting, which is both more expensive and less accurate than phone-based targeting, is never the most efficient targeting approach in these settings.

This paper is related to three main literatures. First, there is extensive past work measuring the accuracy of various "traditional" approaches to targeting social protection, including several papers focused on comparing the accuracy of proxy-means tests (PMT) to community-based targeting (CBT) approaches. This literature generally finds that proxy-means tests are more accurate at identifying the consumption-poor than CBTs (Alatas et al., 2012; Basurto et al., 2020; Premand and Schnitzer, 2021; Schnitzer and Stoeffler, 2022; Trachtman et al., 2022; Sumarto et al., 2025). Consistent with this literature, we find that the PMT outperforms CBT in our setting. Other papers have evaluated alternative 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). Our paper adds to this evidence base, but importantly and distinctively, adds a phone-based algorithmic targeting to the comparison set. This is an important addition, because the rapid spread of mobile phones in otherwise-data-poor regions of developing countries coupled with advances in computing and algorithmic techniques makes cell phone records a promising instrument for cost-effectively improve targeting of humanitarian aid in large scale.

Second, our paper also contributes to a growing literature exploring how "big" digital data sources can be used for targeting (Aiken et al., 2022b,a; Smythe and Blumenstock, 2022). Two prior studies in Afghanistan (Aiken et al., 2022b) and Togo (Aiken et al., 2022a) develop the basic methodology underlying the phone-based targeting approach we deployed in Bangladesh. In this paper we are able to directly compare its performance in the field against competing "traditional" approaches to targeting. We then use that empirical data to identify the circumstances under which phone-based targeting is most efficient. Also related are other papers estimating poverty 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).

Finally, this paper contributes to a nascent literature on cost-effective administration of social protection and humanitarian aid programs in low-income countries. While development programs are frequently evaluated using a cost-effectiveness metric (e.g. Murray et al., 2000), there isn't much systematic evidence on the relative cost-effectiveness of alternative targeting approaches, which is what we attempt to provide here. The tradeoff between cost and accuracy of program targeting we highlight determines cost-effectiveness (Dutrey, 2007; Devereux et al., 2017). Two other studies measure cost-effectiveness of targeting relative to universal distribution: Houssou and Zeller (2011) and Hanna and Olken (2018). The novel contribution of this paper is to provide head-to-head comparisons of the cost-effectiveness of multiple popular approaches to poverty targeting, in addition to the new algorithmic phone-based targeting approach.

2 Data and Methods

The primary empirical 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 in southern Bangladesh — Ramu, Teknaf, and Ukhia.[1] Our main analysis compares proxy-means testing (PMT), community-based targeting (CBT), and phone-based targeting (PBT) for identifying the consumption-poorest households in this setting. We also conduct ancillary analyses of alternative targeting methods, including geographic targeting and simpler variations of proxy-means testing and community-based targeting. In supplementary analyses, we use data from a cash transfer program run by GiveDirectly and the government of Togo in 2021.

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

- A **census** of all households in 200 randomly chosen villages from the three study sub-districts in Bangladesh. The census was conducted in February and March 2023.We collected phone numbers of all adult household members, and basic information about household characteristics and asset ownership necessary to compute the Poverty Probability Index (PPI).[2] The census collected information for around 106,000 households.

- A March 2023 **household survey** which collected consumption expenditures, demographics, assets, and peer rankings. In this survey, we adopted the standardized consumption module from the 2016 Household Income and Expenditures Survey (HIES) implemented by the Bangladesh Bureau

[1]The program was targeted to communities that host Rohingya refugees. These sub-districts host large refugee populations, and there is a sentiment that these poor communities deserve some support for hosting refugees in their midst.

[2]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.

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**. The **peer rankings** module asked each household about eight randomly selected households in their neighborhood. They were asked to report how well they knew the household and to rate how "well-off" the household was on a scale of 1-5. 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 (based on neighborhood size terciles), and the share of households in the neighborhood that were a religious or ethnic minority (no minority households vs. less than 10% minority households vs. 10% minority households or greater). Descriptive measures and summary statistics from the household survey are provided in Figures S1 and S2 and Table S1.

- Household wealth rankings from **community-based targeting exercises** conducted in November 2023 in each of the 180 neighborhoods. Our CBT exercises assembled 12-25 community members from all walks of life from each "neighborhood" to collectively identify the 20% poorest-ranked households who would receive a one-time cash transfer of 1,100 Taka ($35.06 USD PPP) following the meeting. We adopted a protocol regularly implemented by BRAC to determine beneficiaries for their own social safety net programs. The CBT protocol is described in detail in Appendix A.

- Complete **mobile phone metadata** from all consenting survey respondents from March to July 2023, including records of calls, texts, and mobile data usage. We analyzed mobile phone metadata from all four mobile network operators active in Cox's Bazar for all consenting survey respondents. Following the data protection procedures described in our IRB protocol, we pseudonymized or removed all personally identifying information, including phone numbers, prior to analyzing mobile phone metadata.

We use these data to assess several approaches to targeting social protections in the context of southern Bangladesh. The three main targeting methods we study are as follows:

1. The **phone-based targeting (PBT)** approach uses machine learning methods to predict consumption expenditures from roughly 1,500 statistics on each subscribers' mobile phone use (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., 2022b,a, 2023a) and detailed in Appendix A. 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 included 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[3], 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 phone records), and used the matched dataset to train a gradient boosting model[4] to predict log per-capita consumption using mobile phone features. Table S2 shows

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

[4]A gradient boosting model is a nonparametric ensemble machine learning approach. The ensemble consists of a number of decision trees, each of which is trained to predict household poverty from the phone data features. The final poverty prediction for each household is an average of the predictions from each decision tree.

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 most deserving recipients in their neighborhood. See Appendix A for details. Rankings are normalized within each community to a 0-1 range for consistency across communities. This approach implicitly assumes that wealth ranges are consistent across neighborhoods; a more sophisticated approach could make use of data on neighborhood-level poverty to adjust rankings.

3. The **proxy-means test (PMT)** predicts poverty from survey-based covariates. We chose 45 covariates from the household survey based on those typically included in 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 (for example, the household size and gender of the household head), and asset ownership. To create a PMT, we then experimented with a number of machine learning models to predict log per-capita consumption from the 45 covariates, including simple linear regression, linear regression with step-wise forward selection, LASSO regression, and a random forest. When evaluated out-of-sample (see Appendix A for details), we found that the LASSO regression was most accurate, so our main results focus on the LASSO PMT. Figure S4 lists the variables that yielded the largest coefficients in our Bangladesh data. In supplementary results we show all four PMT variants. The L1 penalty parameter for the LASSO regression is chosen via cross validation (McBride and Nichols, 2018; Noriega-Campero et al., 2020). See Appendix A.3 for details.

We additionally replicate some less common targeting approaches that are also relevant counterfactuals:

4. **Geographic targeting** at the union (admin-5) level, based on aggregating population-weighted wealth estimates from the global deprivation index

(CIESIN, 2021), which combines subnational administrative datasets and gridded earth observation datasets to produce an index of relative deprivation. 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. We aggregate the GDI at the union (admin-5) 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 simple average of all wealth ratings elicited in the household survey for a given household by their neighbors. This is similar to the CBT in that it seeks to understand the extent to which neighbors correctly perceive each others' relative standing, but it obtains information from households individually and privately rather than through the collective and public process of the CBT. Unlike the CBT, these peer rankings were not consequential and survey subjects were not told that their rankings would affect real transfers.

Appendix A provides detailed descriptions of the construction of each targeting approach.

3 Accuracy of targeting methods

Our first set of results compares the accuracy of the suite of targeting approaches enumerated in Section 2 for identifying the consumption-poorest households in our setting, with a particular focus on phone-based targeting (PBT), community-based targeting (CBT) and proxy-means testing (PMT). In this analysis, we use per capita household consumption expenditures, collected through our household survey, as the primary benchmark against which all targeting methods are evaluated.

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), while the other 25% are used for the evaluation. We repeat this process 100 times on different random train-test splits, and report the mean and standard deviation of each metric over the 100 runs.[5] To illustrate, Figure 1 shows scatterplots from one train-test split of the rankings under each method vs. per-capita personal consumption expenditure as measured in the household survey. Our results on accuracy below can be anticipated by noting that PMT (center) produces the tightest distribution, followed by PBT (left) and then CBT (right).

Accuracy metrics: We use three standard metrics for assessing targeting methods. The first and most intuitive is *recall*: the probability that a truly poor

[5]Some of the targeting methods we simulate do not produce rankings for all households. For instance, in the phone-based targeting approach, 6% of households are not given a wealth ranking (2% of households in the survey do not provide a phone number or do not consent to matching survey data to mobile phone records; 4% of households in the survey provide at least one mobile phone number but no number is associated with transactions appears in our mobile phone metadata). 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 household without a ranking is prioritized for aid after all households with rankings.

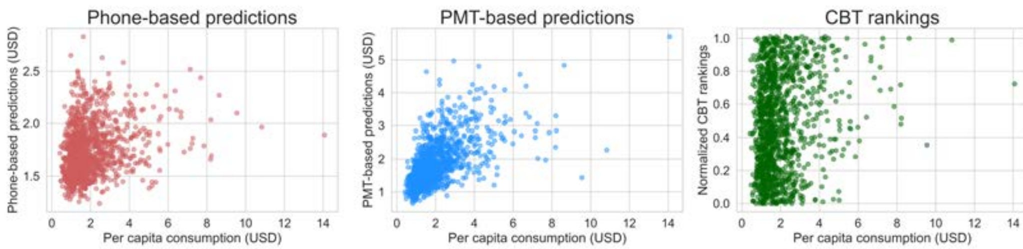


Figure 1: Scatterplots of the three main targeting instruments (phone-based predictions, PMT-based predictions, and CBT rankings) vs. per-capita consumption expenditures. Produced using one train-test split.

household will be correctly classified as poor.[6] This is the simplest metric, but considers only binary errors, not the magnitude of error, and depends on the specific threshold of 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 puts less weight on the exact classification of households near the cutoff, and penalizes large errors in ranking households. The third 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%).[7] A perfect classifier achieves an AUC value of 1, whereas a random classifier (that targets randomly chosen households to fill the quota) achieves a value of 0.5.

[6]Recall, also known as *sensitivity*, is equal to to one minus the type II error rate. Since the program provided transfers to a fixed number of beneficiaries (the 21% quota), *recall* and *precision* – which is the share of households classified as poor that are truly poor (one minus the type I error rate) – are equal in our setting.

[7]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 (which results in high true positives, but also high false positives); by contrast, when the threshold is high, few households will be targeted (low true positives, low false positives). Accurate classifiers yield high true positives and low false positives for a variety of classification thresholds.

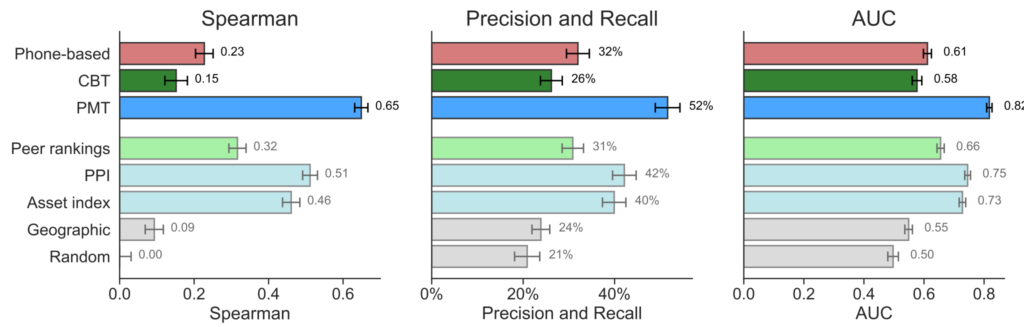


Figure 2: Targeting accuracy comparison, based on Spearman correlation with consumption (left), precision and recall for identifying the 21% consumption-poorest households (middle), and area under the ROC curve (right). Error bars show two standard deviations above and below the mean for each metric.

Main results on targeting accuracy: Figure 2 reports targeting accuracies for each of the targeting methods we evaluate. We observe that phone-based targeting (AUC = 0.61; precision/recall = 32%) is more accurate than CBT (AUC = 0.58; precision/recall = 26%). However, both approaches are substantially less accurate than PMT (AUC = 0.82; precision/recall = 52%). The differences between the three methods are statistically significant based on a Wilcoxon signed-rank test, with $p < 0.001$ for AUC and precision/recall. Other survey-based targeting variants (AUC = 0.72-0.74; precision/recall = 40-42%) also outperform phone-based targeting and CBT but are worse than the PMT. Notably, the decentralized peer ranking approach outperforms the CBT and is comparable in accuracy to phone-based targeting (AUC = 0.66; precision/recall = 31%). Table 1 provides comprehensive targeting accuracy metrics for all targeting methods enumerated in Appendix A.

Binary classification errors do not capture potential differences in magnitudes of errors. That is, two classification methods could have similar error rates for a given threshold, but a method with "small" mistakes (tending to exclude households just below the threshold and include households just above the threshold) is likely to be preferred to a method with "larger" mistakes (tending to exclude households far below the threshold and include households far above the threshold). In Figure S3, we assess magnitudes of errors by

Table 1: Accuracy metrics for all targeting method variants

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: Comparison of targeting accuracy metrics for all targeting variants described in Appendix A. Standard deviations across 100 bootstrap simulations are shown in parentheses.

showing the distribution of consumption per capita for households included and excluded by each targeting approach. Figure S3 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? 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. For comparison, Table S2 shows the features of mobile phone use that are most correlated with per-capita consumption. These include "recharge behavior", which indicates how much money the subscriber adds to their SIM card when they buy phone credit,[8] 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.

Table S3 presents multivariate regressions that identify the household and community-level characteristics that are predictive of inclusion for each of the three main targeting methods (phone-based targeting, community-based targeting, and PMT). Most notably, the community is more likely to select widows/widowers for transfers than either the PMT or PBT, a result that is consistent with the community-based targeting in Indonesia studied by (Sumarto et al., 2025). This suggests that the community may be making use of private, local information about the idiosyncratic disadvantages faced by specific households, which may not be reflected in surveys or in patterns of

[8] 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.

phone use. Both phone-based targeting and the PMT are better at identifying households that spend a large share of their budget on food (a proxy for the household's subsistence risk - see (Bryan et al., 2014)), although this variable is a positive predictor under all three methods.

Table S3 also identifies some of the biases inherent in phone-based targeting. Phone ownership is curiously a *positive* predictor of selection, since households without phones were mechanically excluded by our phone-based selection process. However, conditional on ownership, both phone-based and the PMT exclude households with more frequent phone usage (those with larger number of calls and messages) – which may be a hidden proxy for deprivation that community targeting fails to pick up on. At the neighborhood level, the PMT targets more unequal communities with lower average consumption levels. Phone-based targeting directs transfers to households with fewer social connections; this suggests that the phone data may help reveal the extent to which households are socially isolated.[9] At the neighborhood level, the PMT targets more unequal communities with lower average consumption levels. None of the targeting strategies disproportionately favor or disfavor minority households or minority-dominated neighborhoods.

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, it is possible that the aggregate results mask important heterogeneity – for instance, that CBT's would work better in certain types of neighborhoods (e.g., more homogenous neighborhoods), or that phone-based targeting would work best with certain types of subscribers (e.g., active phone users). However, we find little evidence that the relative performance of different targeting methods

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

varies systematically by neighborhoods or household type. In Panel A of Figure S5, we observe that the PMT generally performs better than phone-based, which performs better than CBT, across all different types of community — including when disaggregating by community size, the share of non-Muslim or non-Bengali minority households, etc. Panel B of Figure S5 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)). Across all types, PMT performs best, and phone-based generally beats CBT , for all types of households and neighborhoods. Phone-based and CBT are statistically comparable, but phone-based targeting almost always outperforms CBT, except within the top quartile of household size.

Figure S5 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 PBT and CBT are a bit more accurate *within* the set of non-minority households.

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 best the goals of a social protection program designed to identify the poorest households among those screened. However, some programs may seek to identify the poorest households within each community, with a quota assigned at the community level. Importantly, in the CBT approach we implement, 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

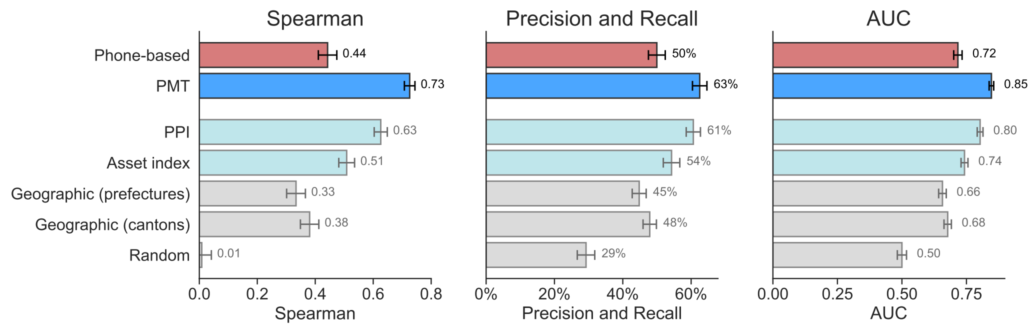


Figure 3: Targeting accuracy comparison for the country of Togo, reproducing results in Aiken et al. (2022b). As in our analysis in Bangladesh in Figure 2, accuracy is calculated over 100 random train-test splits, and error bars show two standard deviations above and below the mean for each metric. This bootstrapping procedure explains very slight differences to the results presented in Aiken et al. (2022b), where 1,000 train-test splits were used.

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 S6, 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 a useful signal for targeting), the quality of targeting approaches relative to one another is unchanged: phone-based targeting is still more accurate than CBT, and less accurate than PMT.

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 3, we replicate the results of Figure 2, instead using data from Togo (Aiken et al., 2022b). In both settings, we find that the PMT is substantially 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.[10]

[10] Our finding that PMT is also more accurate than CBT is consistent with most other papers that have compared the two methods (Schnitzer and Stoeffler, 2022; Premand and

More generally, across all targeting methods, targeting accuracy is relatively low in our setting (AUC = 0.52-0.82; precision and recall of 23-52%). We compare our results to three other published targeting evaluations (which primarily focus on PMT and CBT) to see whether this is unusual: (1)Aiken et al. (2022b), which calculates targeting accuracy nationwide in Togo for a PMT with a 29% targeting quota; (2) 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%, and (3) Brown et al. (2018), which simulates PMT-based country-level targeting in eight African countries with 20% and 40% targeting quotas. Figure 4 plots the precision and recall of CBT and PMT in each of these studies as a function of the targeting quota used. It shows that targeting accuracy under both methods increase linearly with the size of targeting quota. The fit is remarkably tight despite large variations in data, program implementation, and study contexts. Importantly, our results appear to be well within the range of results reported in past work.

Another possible reason for the low targeting accuracy in our setting is the narrow geographic scope of the program we study. Our study is limited to 120 neighborhoods in Cox's Bazar district. As a result, there is likely substantially less variation in poverty in our setting than in the settings of national-scale social protection programs. To test this hypothesis, Figure S7 simulates targeting more homogeneous subsets of our study population by poverty. The results confirm that, for all methods except for random and geographic targeting, targeting simulations that are restricted to poorer subsets of the households in our survey result in lower targeting performance than evaluations conducted with the full set of households in our survey.

Combining targeting methods: It is possible that the targeting data sources could complement each another, such that a combined approach im-

Schnitzer, 2021; Alatas et al., 2012). However, the difference in our setting is relatively more extreme: 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).

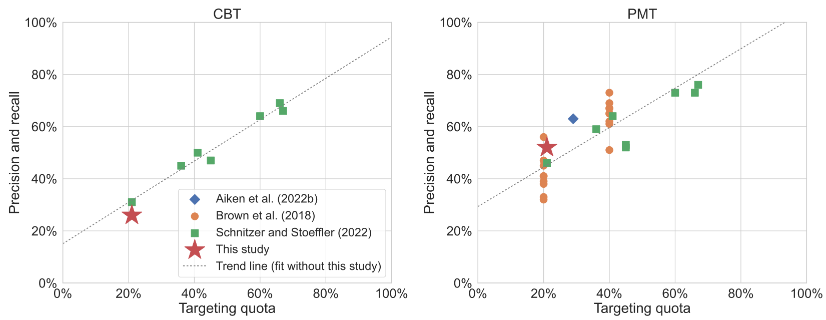


Figure 4: Comparison of our results on targeting accuracy (red stars) in comparison to past studies that also use a quota approach to targeting evaluation (green squares for Schnitzer and Stoeffler (2022), blue diamonds for Brown et al. (2018), and orange dots for Aiken et al. (2022b). Targeting error rate is shown as a function of the targeting quota.

.

proves overall targeting accuracy. Figure S8 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 household deemed worthy of a transfer under method A with the poorest household identified 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 S8 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. The phone + CBT approach does improve precision and recall very slightly (0.5 percentage points) over pure phone-based targeting.

An alternative approach to combining targeting data sources is to include variables from multiple data sources in the ML models used to train the PMT and phone-based targeting methods.[11] Figure S9 shows that the ML-

[11]Specifically, for the phone + PMT-based approach, all phone features and PMT features are included in the model. For the phone + CBT-based approach, phone features and the CBT rankings are included in the model. For the PMT + CBT-based approach, PMT features and the CBT rankings are included in the model.

based approach to combining data sources also does not substantially improve accuracy relative to using single data sources individually.

4 Cost-accuracy trade-offs

Section 3 shows that proxy-means testing is more accurate than phone-based targeting. However, phone-based targeting can be much cheaper than proxy-means testing – especially for large scale programs – because it does not require in-person primary data collection for screening. This creates a trade-off between cost and accuracy. This section introduces a framework for identifying the conditions under which each of the two targeting methods would be more "cost-effective", adapting a framework introduced by Hanna and Olken (2018).

We assume that the implementer has a total budget B and chooses between targeting methods to identify and send 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 equal payments b_m to each recipient, with $b_m = (B - C_m)/I$.[12] 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 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

[12] 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 most real-world social protection programs.

unweighted sum of household utilities

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

Due to diminishing marginal utility, the implementer prefers to allocate transfers to households with lower pre-program y_i, but faces a tradeoff when 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 the value of the implementer's objective function in the absence of the program, and $V_{1\mathrm{B}}$, the "first-best" value, i.e., if the implementer could costlessly obtain the exact ranking of all households and target perfectly. The gain in this first-best scenario, then, is $V_{1\mathrm{B}} - V_0$. We then calculate V_m, the value of the objective function for each method m (i.e., PMT, PBT, CBT), and report G_m, the gain relative to the first-best:

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

Using this framework, we calculate how G_m varies across different targeting strategies m as we vary program budgets, transfer size, and the size of populations screened. This allows to compare the cost-effectiveness of algorithmic versus traditional targeting approaches as a function of the scale of the benefit transfer program.

Data on Costs. Screening costs is 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. All targeting methods require a detailed household consumption survey for benchmarking and calibration. In Bangladesh, this cost $46,600 for 5,000 households.[13] This element of cost is common to all targeting methods and does not meaningfully affect our cost-effectiveness comparisons.

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.[14] 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.[15]

Cost-effectiveness Results for Bangladesh. We begin by computing G_m generated by the GiveDirectly program described in Section 2, which had a budget of roughly $5 million and screened around 100,000 households. Given our data on the fixed and variable costs of implementation, we can also compute changes to G_m as we vary the program budget and the number of people screened.

Figure 5 (left panel) shows the gains from the GiveDirectly transfers in southern Bangladesh under the assumption of CRRA utility. This program

[13]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.

[14]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 that the median reported PMT cost is $4.00 (Aiken et al., 2022a). We also present results assuming this higher PMT cost to examine sensitivity.

[15]Our CBT cost is similar to other costs reported in the literature. Aiken et al. (2022a) report a median per-household CBT cost of $2.20.

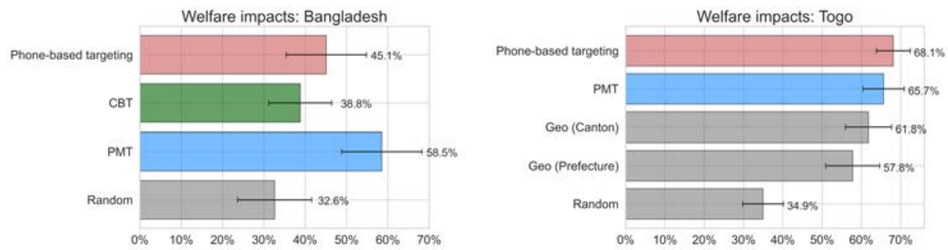


Figure 5: Welfare impacts of GiveDirectly's cash transfer program in Southern Bangladesh (left) and the GD-Novissi program in Togo (right). Welfare impacts are calculated based on the screening costs described in Section 4 and the parameters of the two programs. In Bangladesh, these parameters are a $5 million budget for 100,000 households screened, targeting 21% of households. In Togo, these parameters are a $5 million budget for 207,000 households screened, targeting 29% of households. Utility impact is calculated as the ratio of post-program utility to pre-program utility.

had a budget of roughly $5 million and screened roughly 100,000 households. Using the cost estimates from our surveys, the screening costs for a PMT in this setting are estimated at $177,900, leaving around $4.8 million for cash transfers. The screening costs for phone-based targeting were $46,600, leaving nearly the entire $5 million for cash transfers. Despite the higher costs of the PMT, the higher targeting accuracy of the PMT results in a larger gain to V than implementing these transfers using phone-based targeting (58.5% of the gains of costless perfect targeting vs. 45.1

Generalizability to Togo To understand the generalizability of our findings, we repeat this calculation for the GiveDirectly-Novissi (GD-Novissi) program in Togo described in Aiken et al. (2022b). GD-Novissi also had a budget of roughly $5 million, and screened roughly 207,000 households. GD-Novissi, like GiveDirectly's program in Bangladesh, targeted transfers using mobile phone metadata. 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.

Three key differences between the two research settings in Bangladesh and Togo are worth noting: First, no community-based targeting data was collected in Togo, so we can only compare PMT and phone-based targeting there. Second, the nationally representative survey data in Togo are 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 Bangladesh, households without phones are included in the analysis, and assumed to be targeted *last* under the phone-based targeting approach. Third, in Togo we analyze a national-scale aid program using nationally-representative survey data. In Bangladesh we focus on three sub-districts, resulting in a study population with substantially less geographic and socioeconomic heterogeneity.

The right panel of Figure 5 repeats the same calculation as in Bangladesh for the GD-Novissi program in Togo. Screening costs for the PMT in Togo would be \$258,750, leaving around \$4.7 million for cash transfers. The larger screening costs for the PMTand the better accuracy of phone-based targeting in Togo jointly imply that the gains in V from phone-based targeting (68.1%) slightly exceed that of a PMT (65.7%) in Togo.

These contrasting relative cost-effectiveness results in Bangladesh versus Togo illustrate how the relative efficiency of phone-based targeting and proxy-means testing depends on both the scale of an aid program (in terms of budget and screening costs) as well as the relative accuracy of the targeting methods being compared. We now study this tradeoff more extensively by varying the scale and scope of hypothetical transfer programs, and calibrating against real-world social protection programs tracked by the World Bank's ASPIRE database.

Comparing methods' performance as a function of program scale. The GiveDirectly programs we analyze are fairly small scale — 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.[16]

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

Figure 6 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). 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 welfare gains. 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, phone-based targeting is preferred. This is mainly because the marginal cost of screening additional beneficiaries using mobile phone meta-data – once an algorithm is already developed – is essentially zero.

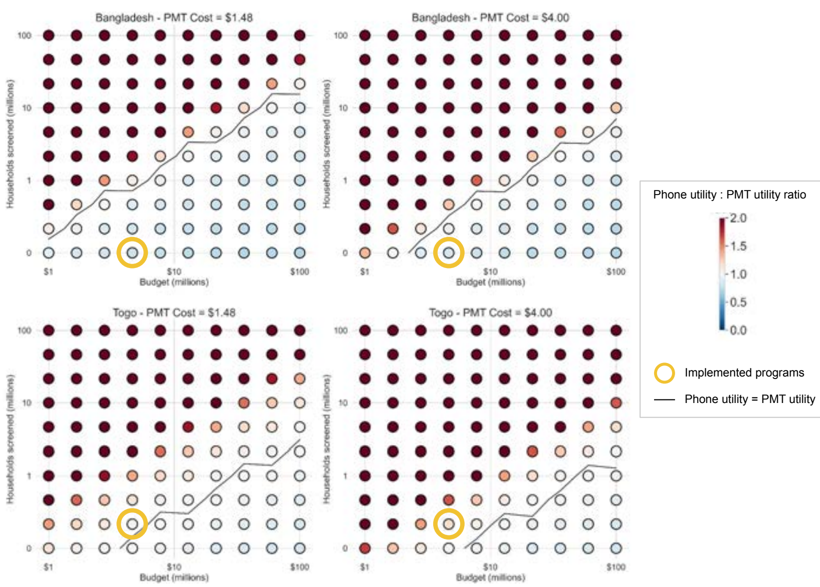


Figure 6: Ratio of utility impacts between phone-based targeting and proxy means testing as a function of a hypothetical social protection program's budget (x-axis) and households screened (y-axis). Red shades represent program scales at which phone-based targeting is preferred, blue shades represent program scales at which PMT is preferred, and the line identifies the "decision threshold". Left: PMT variable cost of $1.48 per household screened (based on the costs of our surveys in Southern Bangladesh). Right: PMT variable cost of $4.85 per household screened (based on the median of values reported in the literature). Above: Using data from Bangladesh. Below: Using data from Togo.

The top-left panel of Figure 6 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 outperformed phone-based targeting in terms of V. This occurs partly because the variable per-household screening cost for the PMT in Bangladesh ($1.48) was unusually low, reflecting uniquely low costs of data collection in Bangladesh (see footnote 14). The right two panels of Figure 6 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., 2022a). For more typical PMT screening costs, the scope and scale of programs where phone-based targeting is preferred to PMT expand.[17] More broadly, Figure 6 highlights how 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 that attempt to screen a large number of individuals to make smaller per-household transfers.

Figure 7 illustrates the thresholds at which different targeting methods provide the largest increase in utility, as a function of the *program budget per household screened*.[18] In Bangladesh (Panel A), phone-based targeting (red line) is preferred to the PMT for programs with budgets under $4 per household screened if the PMT costs $1.25 (solid green line); however, if the PMT costs the "industry-standard" $4.00 (dashed green line), phone-based targeting is preferred for budgets 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 the more typical 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 scenar-

[17]Figure S10 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.

[18]In constructing Figure 7, 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.

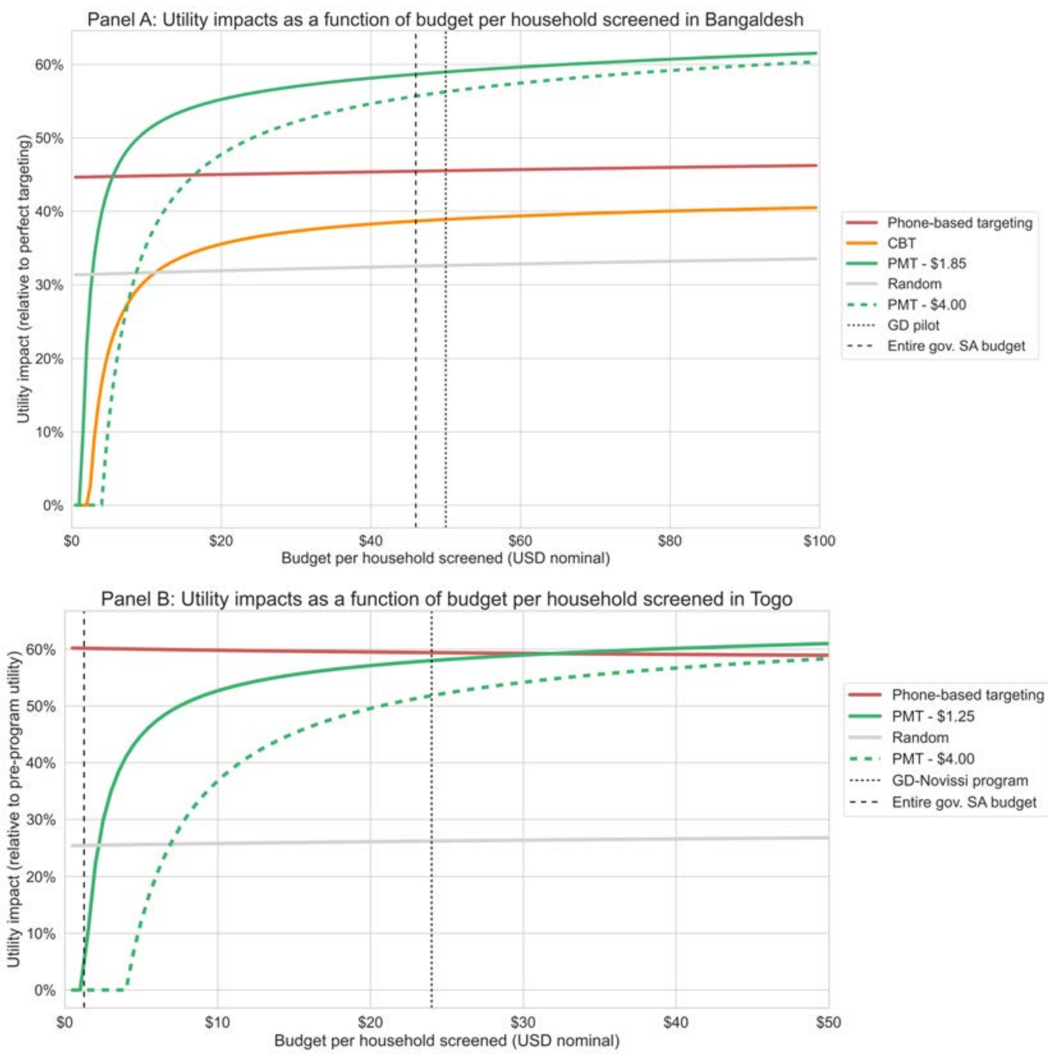


Figure 7: Above: Utility impacts of phone-based targeting, CBT, and PMT as a function of the budget per household screened, using our data from Bangladesh. (This analysis requires fixed costs to be dropped from calculations, implicitly assuming that fixed costs are negligible when the number of households screened are sufficiently large). Utility impact is calculated as the ratio of the estimated post-program aggregate utility to the estimated pre-program aggregate utility. The solid green line uses the PMT variable cost of $1.48 per household screened (based on the costs of our surveys in Southern Bangladesh), and the dashed green line the value of $4.85 per household screened (based on the median of values reported in the literature). Below: Utility impacts of phone-based targeting and PMT as a function of budget per household screened, using our data from Togo. In each figure, the dashed lines show the budget of the aid program by GiveDirectly and the entire government cash assistance budget.

ios, Figure 8 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. Figure 8 shows, on a log scale, our two cutoffs 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 sufficiently low (less than \$15/hh) that phone-based targeting is preferred 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.[19]

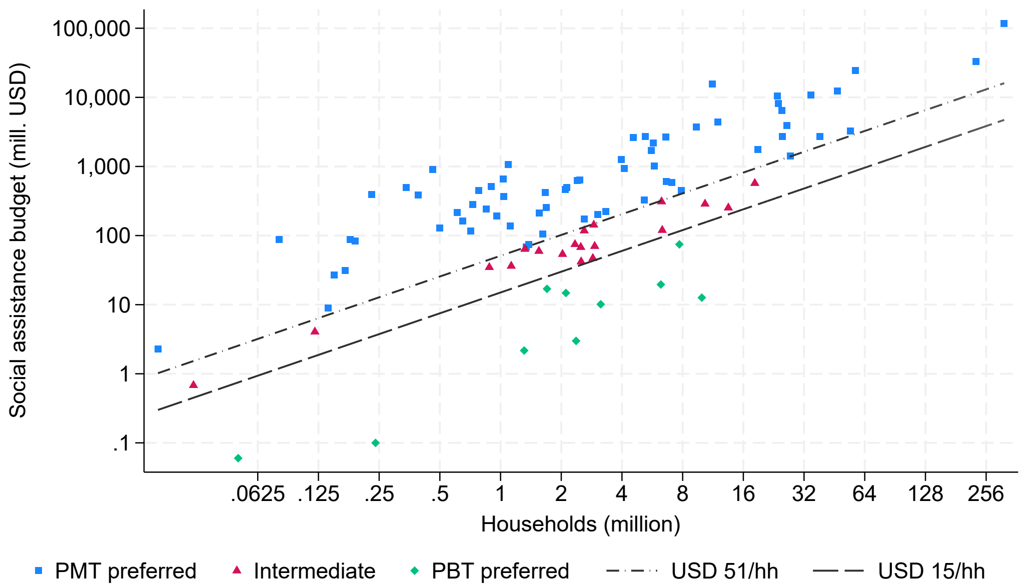
Sensitivity. Our analysis thus far highlights how the choice of the "best" targeting method (i.e., the one that maximizes W_m) depends on the size of the program relative to the number of households screened. Another critical factor in that determination is the relative accuracy of each targeting method. In settings like Togo, where the accuracy of phone-based targeting is higher, phone-based targeting will be preferred for a larger range of programs. The role of targeting accuracy is shown in Figure S11 Panels B and C — which replicate the results shown in Figure 7 but simulates more accurate phone-based targeting[20] — and in Table S4, which indicates the point at which phone-based targeting would be preferred to a PMT, as a function of the accuracy of the phone-based targeting method. 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. Appendix B provides more details on these simulations of the welfare implications of improved phone-based targeting accuracy.

Other factors can also influence the performance of phone-based targeting

[19] We provide country-by-country details in Appendix D.

[20] Figure S11A displays similar analysis for CBT targeting of simualted higher accuracy levels.

Figure 8: Social assistance: total budgets vs households screened



Notes: this figure plots countries' budgets for cash-based social assistance transfers versus the number of households that need to be screened in order to implement. 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.

relative to a PMT. Several of these are highlighted in Figure S10, which shows how relative gains G_m vary with the variable cost of the PMT, the share of the population targeted, and the coefficient of relative risk aversion σ. While the share of the population targeted affects the total utility associated with each targeting method (with wider targeting leading to higher impacts), the ranking of different targeting methods does not change: PMT is still substantially better than phone-based targeting, which is in turn better than CBT. Panel B shows that qualitative differences across methods are somewhat more sensitive to the coefficient of relative risk aversion σ: at very small values of σ differences between methods are less, while at very high values of σ gaps increase. Not surprisingly, differences between the PMT and other methods are most sensitive to the screening costs of the PMT (Panel C): as the PMT variable cost becomes very high (approaching and exceeding $15 per household), the welfare impacts of the PMT approach that of phone-based targeting.

Caveats. The cost-effectiveness analysis we conduct abstracts away certain aspects of real-world social protection programs. A particularly important aspect is the frequency of targeting reassessments: We assume that targeting costs are incurred every year, and that all households are screened each year. In reality, PMT and CBT targeting sweeps are typically conducted only every few years (Barca and Hebbar, 2020), lowering per-year screening costs. However, when a PMT is not up-to-date, its targeting accuracy also decreases (Aiken et al., 2023b; Hillebrecht et al., 2023; Brown et al., 2018) — our analysis does not account for these dynamics; nor does it account for on-demand registration models where households can self-select into providing PMT or other data relevant to screening, either instead of or in addition to a comprehensive country-wide survey process.

5 Discussion and Conclusion

Our paper produces two key findings. First, in the context of 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). This result is consistent with past work in Togo showing that phone-based targeting is less accurate than PMTs Aiken et al. (2022b). Second, we provide the first head-to-head comparison of targeting approaches based on cost-effectiveness, building on the welfare framework introduced by Hanna and Olken (2018). We show that social protection programs with large budgets relative to the number of households screened for eligibility should invest in proxy-means testing for accurate targeting, but cheaper phone-based targeting is preferred for programs with thinly stretched budgets (below $10-50 per household screened).

Our results have important implications for real-world social protection and humanitarian aid programs run by governments and NGOs. Using data on real-world government-run social protection programs, we find that most government-run social protection programs have sufficiently large budgets that proxy-means testing is the most efficient targeting approach, however 10-30% countries from the World Bank ASPIRE database have sufficiently low social protection budgets relative to the size of their population that phone-based targeting is preferred. Intuitively, phone-based targeting is particularly relevant for one-off programs that aim to deliver aid to the poorest households among a large population, where screening costs would become prohibitively high.

A number of abstractions and limitations to our analysis are worth noting. First, our analysis relies on household survey and administrative data from Bangladesh and Togo. The cost-accuracy trade-off among targeting approaches is likely to be country-specific, depending on the distribution of poverty, accuracy of targeting approaches, and cost of conducting field activities, among others. Future research to calibrate costs and accuracy of targeting approaches — particularly digital approaches like phone-based targeting — is needed to make specific recommendations in individual countries, or to generalize across a larger set of countries.

Second, we assume that households are re-targeted for eligibility each year. Less frequent re-targeting would lower costs, but would also likely result in lower accuracy (Aiken et al., 2023b; Hillebrecht et al., 2023; Brown et al., 2018). Future research to identify how quickly the accuracy of proxy-means testing, community-based targeting, and phone-based targeting decay would provide a stronger understanding of the cost-accuracy trade-off over time.

Third, our measurement of costs focuses only on the financial costs of screening households. Other costs associated with targeting — such as private costs to households participating in screening activities like household surveys and community meetings, social costs from tensions arising from selective inclusion of beneficiaries, psycho-social costs incurred by screening processes, and political costs to support for social protection programs (Devereux et al., 2017) — are not accounted for in our analysis, primarily because the data necessary to account for them were not collected in our field activities and not recorded in the literature.

Fourth, our analysis focuses on the accuracy and cost of targeting poverty at a single point in time, and does not account for changes in poverty over time. Of the methods assessed in this paper, phone-based targeting could most easily be adapted to identify households experiencing economic shocks or poverty transitions. Future work on the extent to which phone-based targeting can measure changes in poverty over time, in comparison to traditional approaches, could provide new opportunities for adaptive targeting of program benefits.

Fifth, and finally, we have abstracted away from the "Lucas critique" that, if eligibility decision criteria (including phone-based targeting PMT, and CBT) are used systematically, households might strategically alter their behavior to game the targeting regime. However, the "black box" nature of the phone-based targeting approach may be a virtue in this setting, since it would be difficult for households to predict which features of phone use will be selected, and thus difficult for them to successfully game the regime. Furthermore, it is not clear that manipulating phone use is less costly than answering PMT survey questions strategically or hiding information from neighbors.

In conclusion, we find that while phone-based targeting is less accurate than proxy-means testing, its lower cost still makes it the more efficient targeting approach for social protection programs with low budgets relative to their scale.

Proxy-means testing is the preferred approach for higher-budget programs, and community-based targeting is both less accurate and more expensive than the other options (at least in our setting in southern Bangladesh). More broadly, our results provide a framework for trading off cost and accuracy in poverty targeting which can be adapted across countries and additional targeting methods beyond those studied here.

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., Bedoya, G., Blumenstock, J., and Coville, A. (2022a). Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan. *arXiv preprint arXiv:2206.11400*.

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. (2022b). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603(7903):864–870.

Aiken, E., Ohlenburg, T., and Blumenstock, J. (2023b). 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.

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., 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?

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.

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

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.

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.

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.

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.

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.

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*.

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.

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*.

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.

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., 2022b,a). 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[21] 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

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

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 metadata from the most senior member of the household is used.[22] 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

[22]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 S12 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%).

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 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 wealth of all households in the community by placing index cards representing each household on a string in the order of wealth. 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 wealth 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 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 one-hot encoded; 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 S13 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.[23]

[23] 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., 2022b; Smythe and Blumenstock, 2022) because RWI data are missing for

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 from lowest to highest resolution:

- **Unions:** We use publicly available union shapefiles[24] 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.

much of the eastern portion of Cox's Bazar.

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

- **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 S14 shows the poverty maps produced through this technique, at the union, village, and neighborhood level.

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 S13 Panel B shows the distribution of the PPI in our household survey.[25]

[25]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.

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 S13 Panel C shows an example asset index distribution from one of the train-test splits.

A.7 Peer rankings

The community ranking module in our survey collected two types of peer rankings: an absolute welfare estimate, where households were asked to rate other households' welfare on a scale of 1-5, and a relative welfare estimate, where households were asked to order the welfare of other households. Each household is asked to assess eight randomly selected other households in their neighborhood, as well as themselves. Each household is asked to rank a different set of eight households from among the other households surveyed in their neighborhood, drawn such that every household appears on the to-be-ranked list provided to eight other households. We also elicit how well each household knows the households they are asked to rank. From these various configurations we obtain six variants of peer rankings: two options for the ranking type (absolute or relative), and three options for which rankings to include (all rankings, just neighbor rankings — dropping the self-ranking — or just high-confidence neighbor rankings plus the self ranking). For the absolute poverty rankings, we also test using only the self ranking, without any community input.

In the survey, if a household reported not knowing one of the households it was supposed to rank at all, they were not required to rank that household. As such, most households are not ranked eight total 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 S15 shows the distribution of the number of times households are ranked.

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 ranker 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 HodgeRank algorithm.

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 S16 shows the distributions of the six targeting rankings.

B Simulating Counterfactual Targeting Performance

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 S11 Panel A reproduces Figure 7 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 ρ = 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).

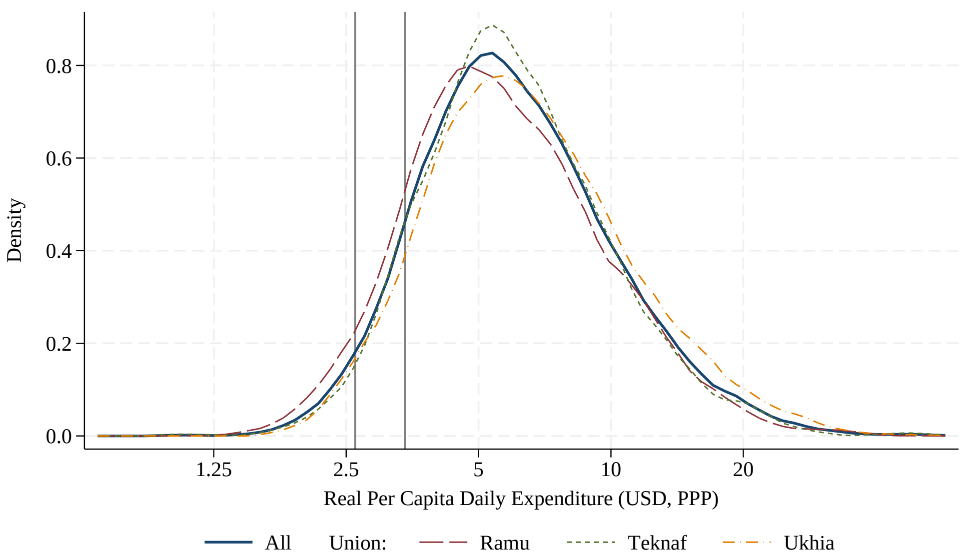
B.2 Simulating Improved Phone-Based Targeting

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 (ρ = 0.23) — there is a broader scope of programs for which phone-based targeting achieves a higher utility impact than PMT (Figure 7). 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 CBT rank and its true consumption rank, weighting the consumption rank progressively higher to move CBT rankings closer to the correct rankings. Figure S11 Panels B (Bangladesh) and C (Togo) reproduce the results from Figure 7 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 S4 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.

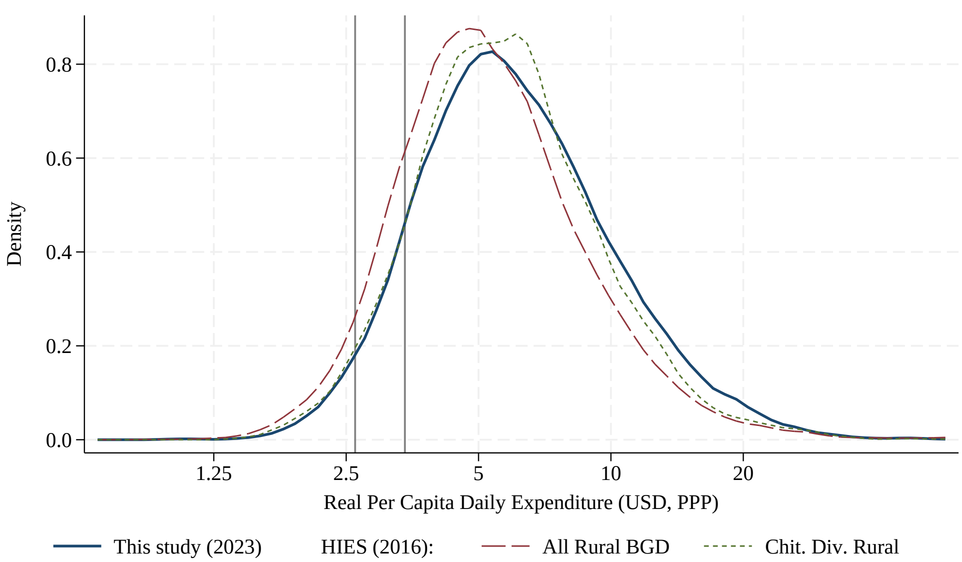
C Supplementary Figures and Tables

Figure S1: Density of household real per-capita daily consumption

(a) By union



(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. The 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 union. 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 (long dash) and rural households in Chittagong division (short dash). Our study was conducted in three sub-districts of Cox's Bazar district in Chittagong. HIES observations are weighted using the HIES household inverse probability weights. 2016 nominal consumption in BDT is converted to 2023 BDT using the Bangladesh CPI, and then to USD at the mean 2023 PPP exchange rate for personal consumption of 30.7 BDT/USD.

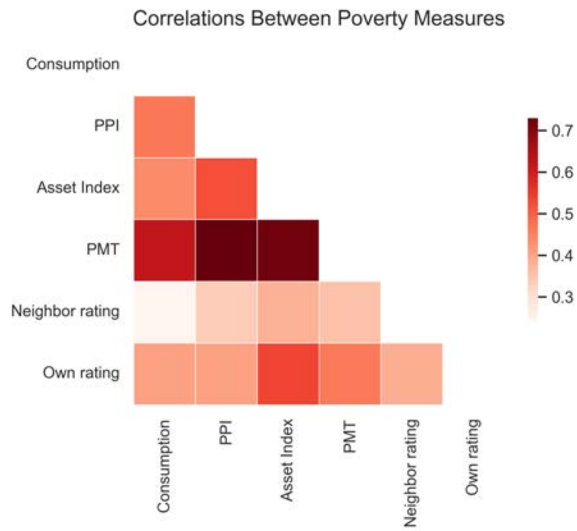


Figure S2: Correlation between key poverty outcomes in our household survey.

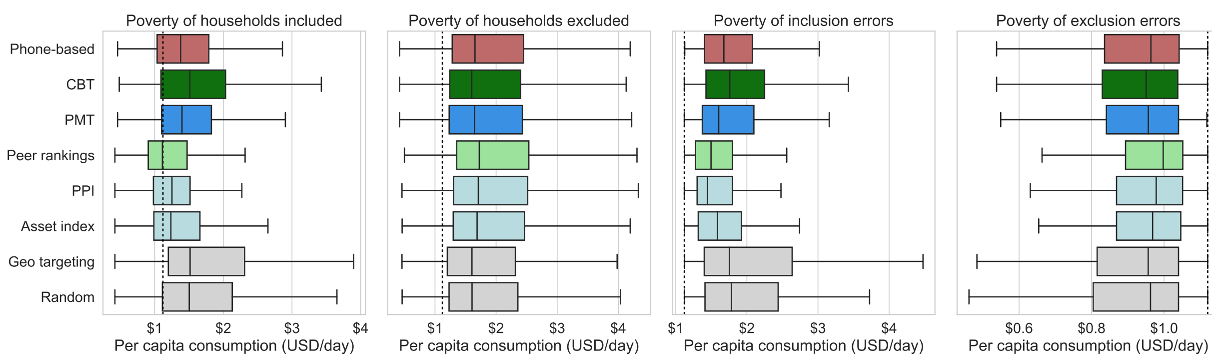


Figure S3: Consumption expenditure distributions for households included by each targeting method (left), excluded by each targeting method (center left), wrongly included by each targeting method (center right), and wrongly excluded by each targeting method (right).

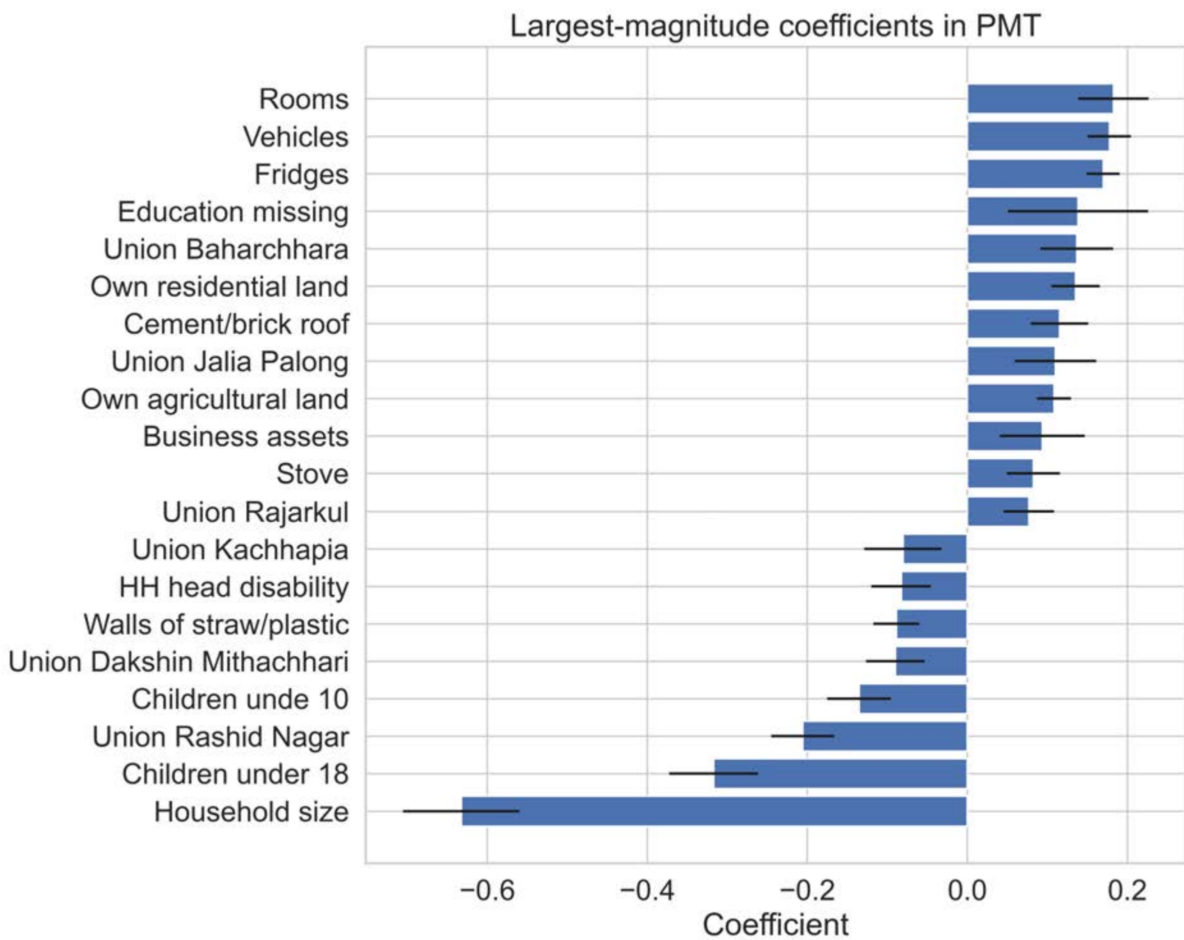


Figure S4: Top 20 PMT variables with the highest magnitude coefficients. Coefficients are averaged over all 100 train-test splits, with error bars showing two standard errors above and below the mean coefficient across the 100 splits.

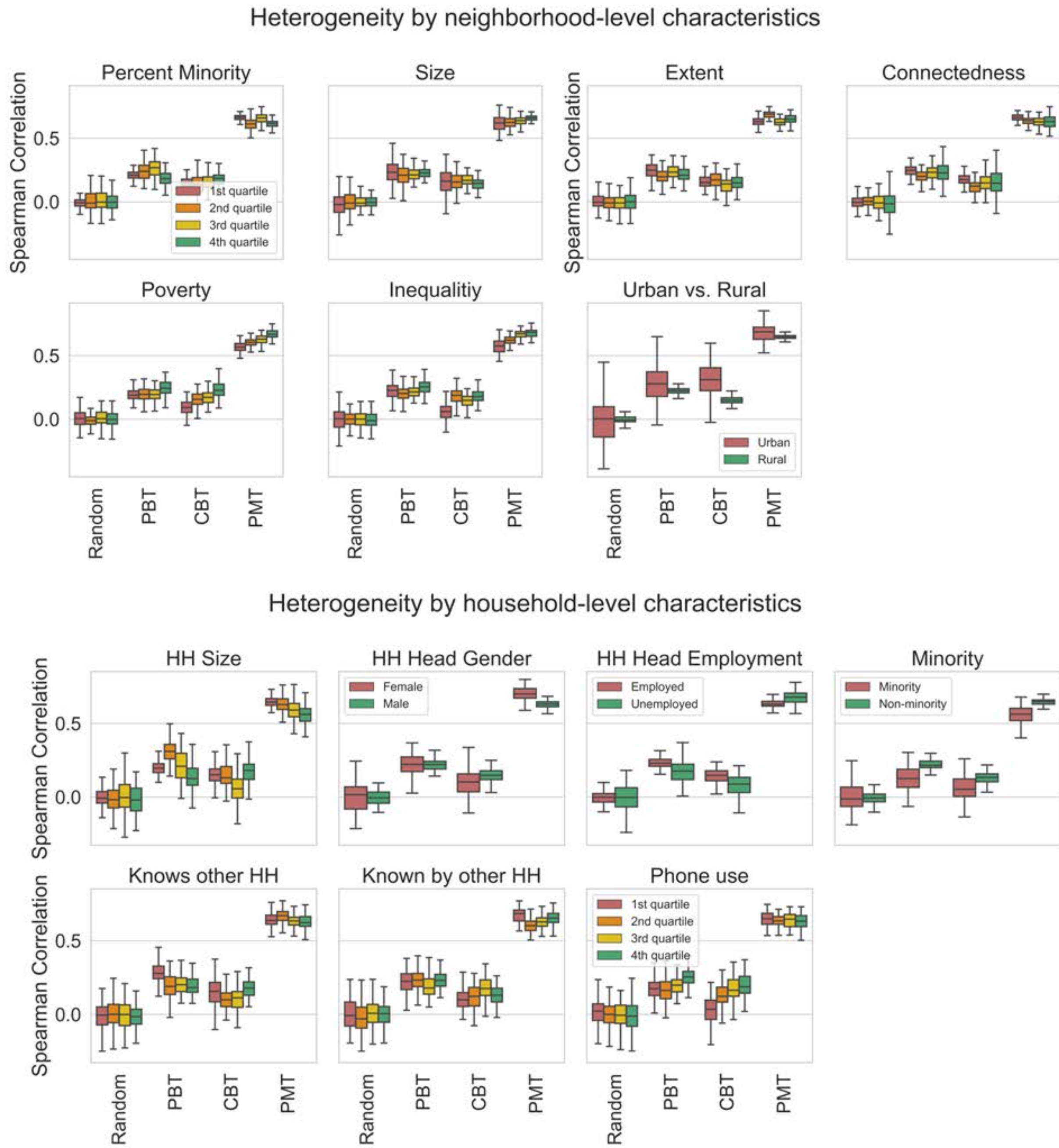


Figure S5: Heterogeneity in targeting accuracy by neighborhood-level characteristics (top row) and household-level characteristics (bottom row). Each plot shows the distribution of Spearman correlations (over the 100 random train-test splits) for each group.

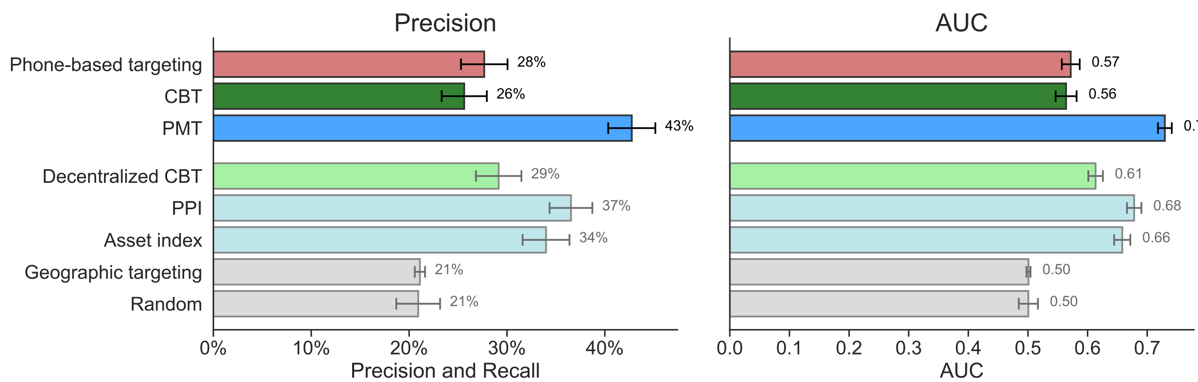


Figure S6: Targeting accuracy comparison for identifying the poorest households *in each neighborhood (rather than across the entire population, as in Figure 2). Accuracy based on precision and recall for identifying the 21% consumption-poorest households in each neighborhood (left), and area under the ROC curve (right). Error bars show two standard deviations above and below the mean for each metric.*

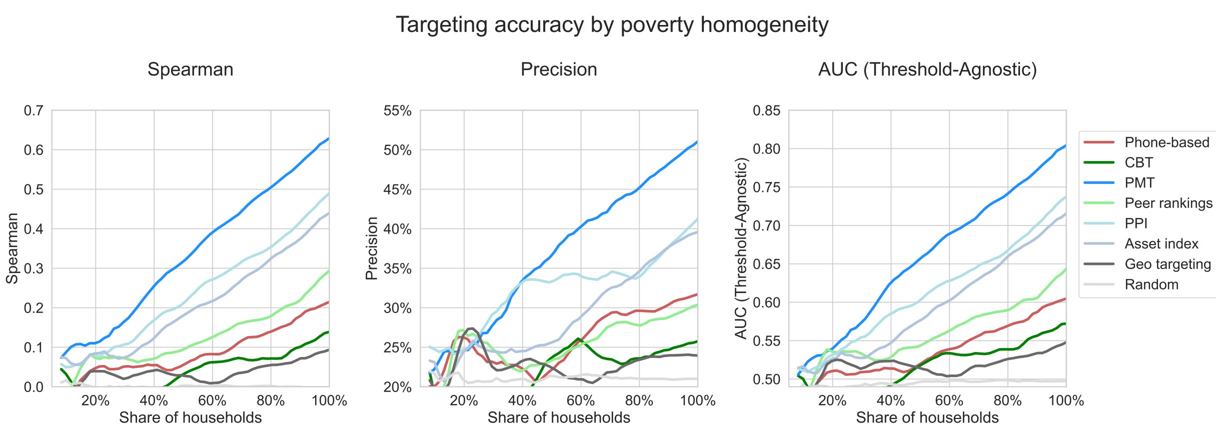


Figure S7: Targeting accuracy of each targeting method as a function of the poverty homogeneity of the population. The x-axis represents the share of households from our survey included, ranked by poverty: thus 20% indicates restricting the targeting evaluation to the 20% poorest households in our survey.

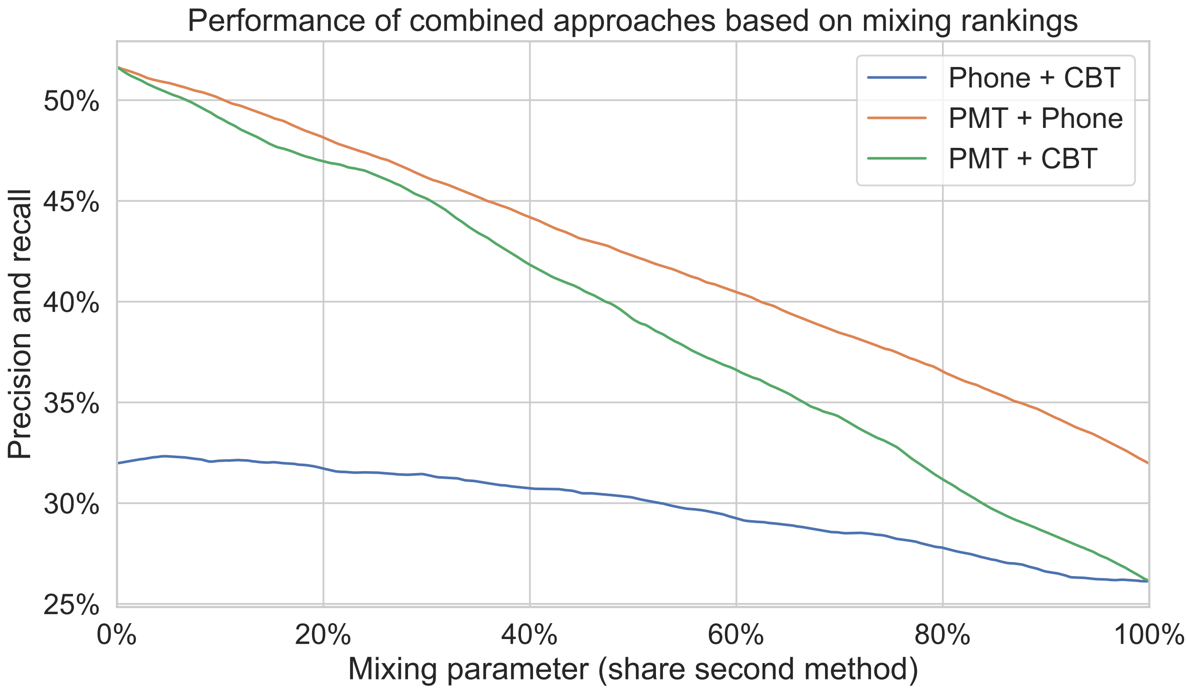


Figure S8: Accuracy of approaches that combine rankings from multiple data sources for targeting following the methods described in Section 3. The x-axis represents the "mixing parameter": the share of rankings that are taken from the second method in the pair (as opposed to the first). Three combined methods are tested: phone + CBT rankings (so the x-axis represents the share of rankings taken from the CBT), PMT + phone rankings (so the x-axis represents the share of rankings taken from phone-based targeting), and PMT + CBT rankings (so the x-axis again represents the share of rankings taken from the CBT). Precision and recall measures are the average over 100 bootstrap simulations.

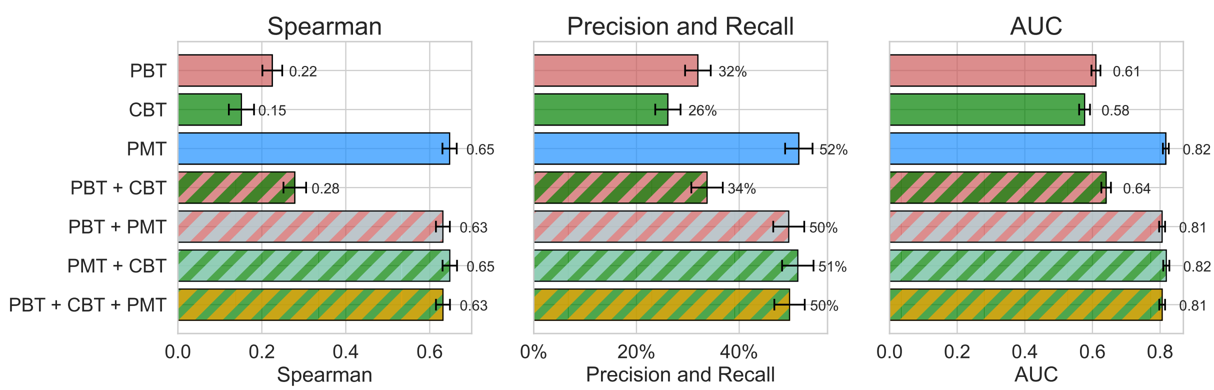


Figure S9: Accuracy of ML-based approaches for combining two data sources into a single targeting approach, following the methods described in 3.

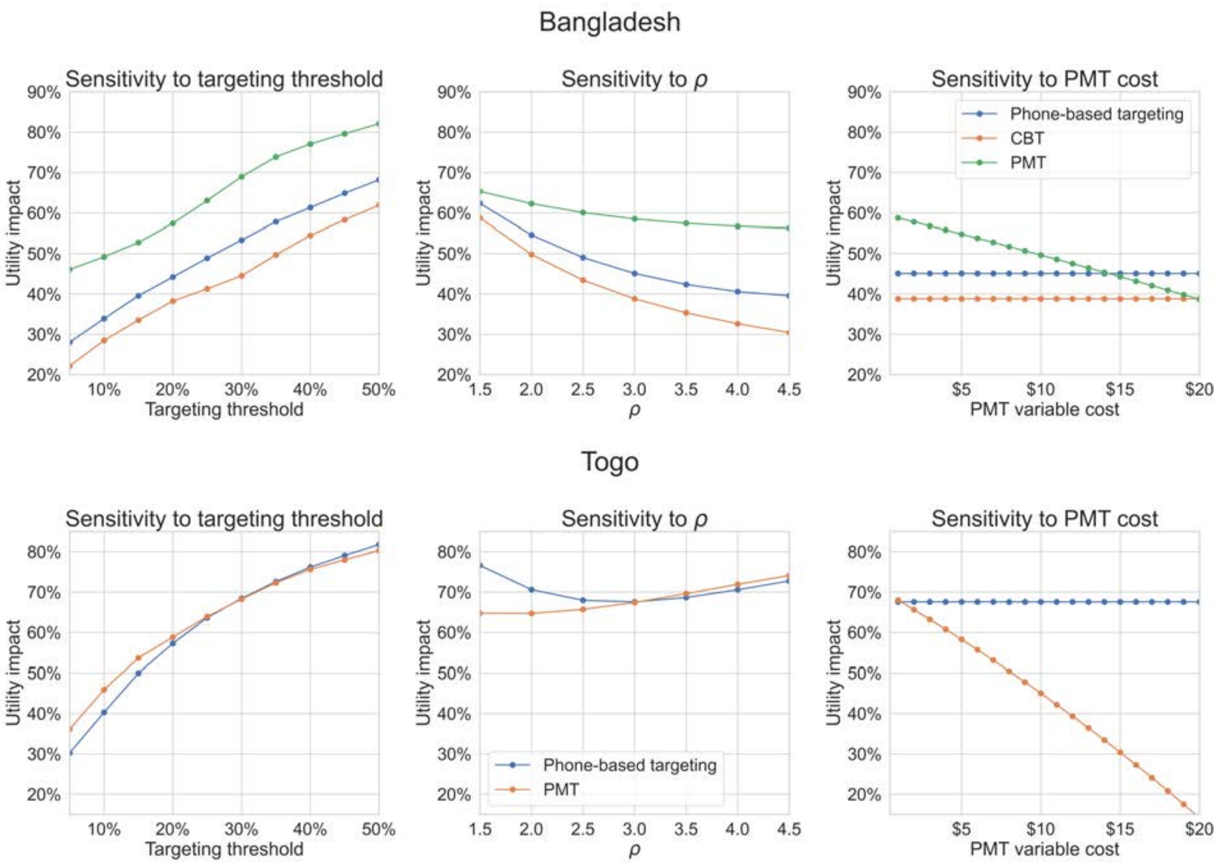


Figure S10: Sensitivity of welfare results for the GiveDirectly programs to additional parameters: the targeting threshold (left in each row), the coefficient of relative risk aversion for the CRRA utility function (center in each row), and the PMT variable cost (right in each row). Top: GiveDirectly program in Bangladesh (using the same data as the left panel of Figure 5). Bottom: GD-Novissi program in Togo (using the same data as the right panel of Figure 5).

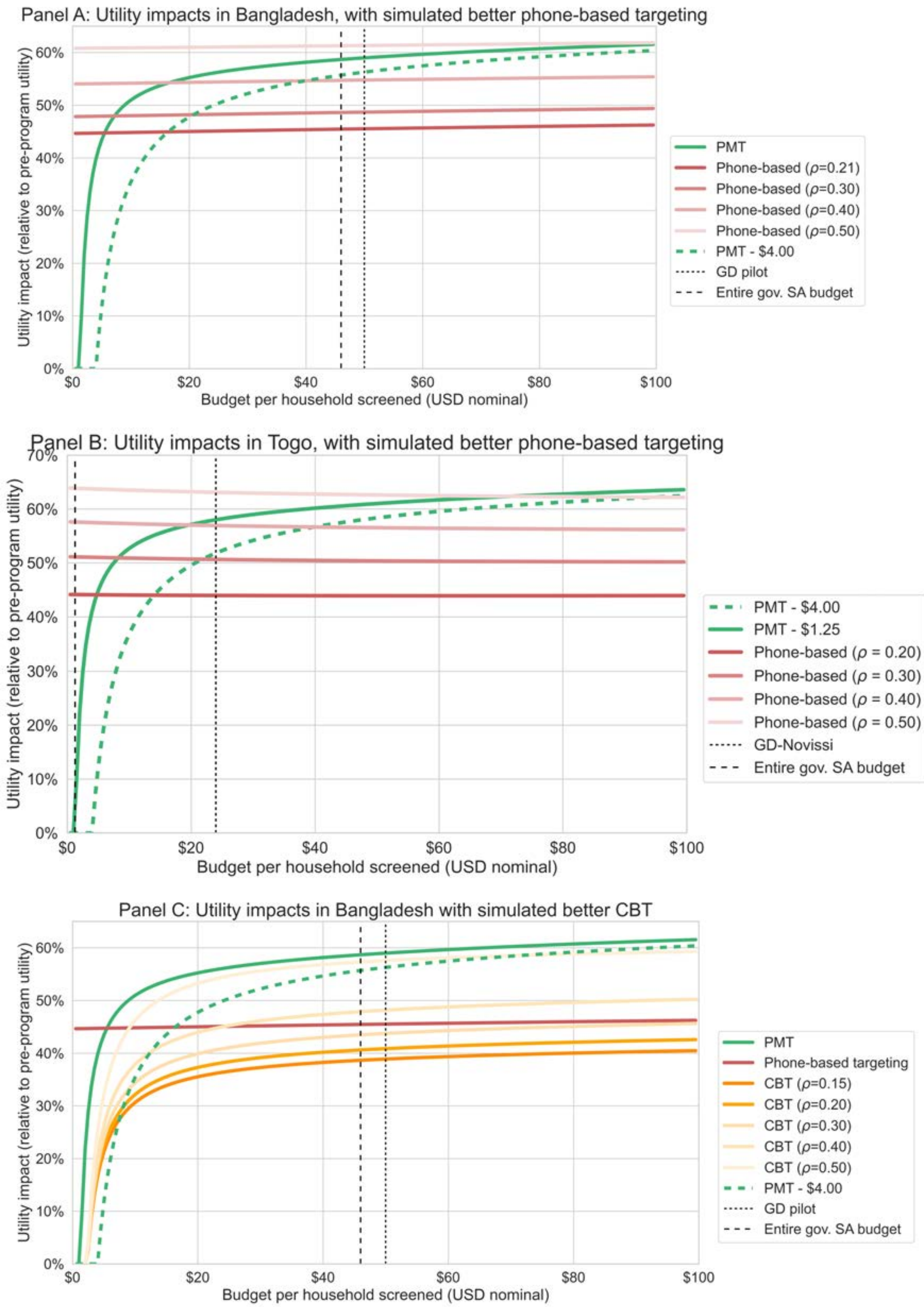


Figure S11: Replication of Figure 7 with the addition of simulated "better" CBT targeting (top panel) and phone-based targeting (middle and bottom panels). See Appendix B for details on how higher-accuracy CBT and phone-based targeting methods are simulated.

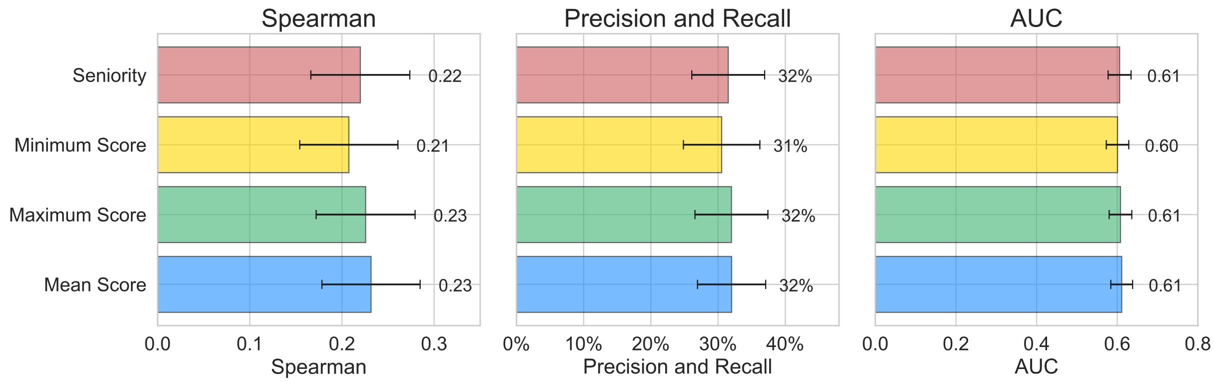


Figure S12: Targeting accuracy metrics for four different approaches to aggregating phone-based predictions for households with multiple phones. Targeting accuracy is calculated using the household survey dataset, as in the main targeting evaluation (Figure 2), and the approach for households providing 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.

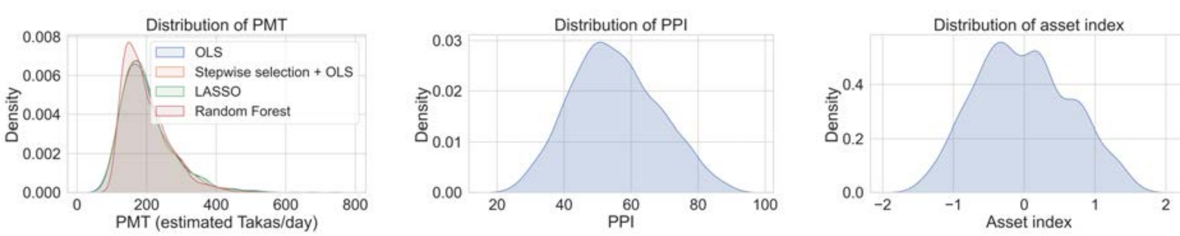


Figure S13: Kernel density estimates showing the distribution of the PMT (left, with four versions corresponding to the four machine learning models tested), PPI (middle), and asset index (right), for one example train-test split.

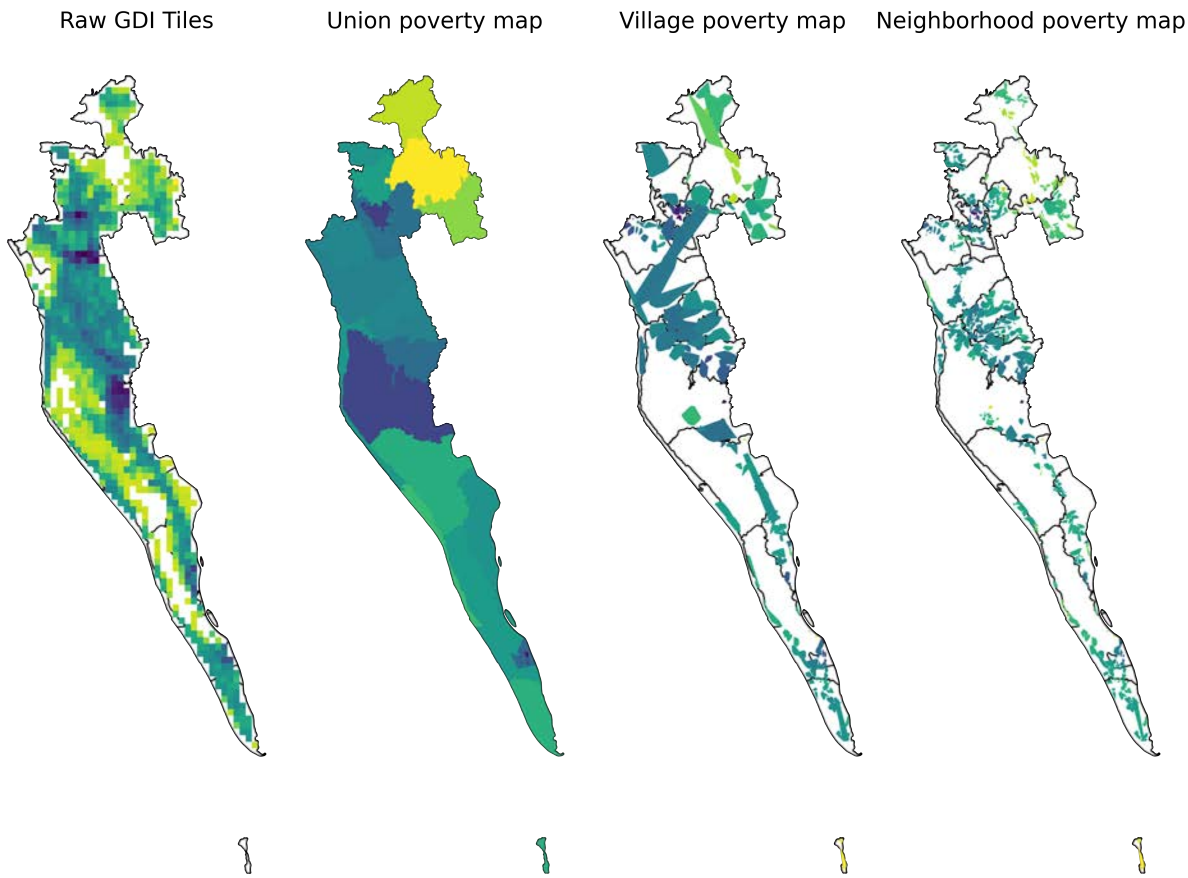


Figure S14: Poverty maps produced by aggregating the Global Deprivation index (GDI) at the union, village, and neighborhood level, as described in Appendix A.

Figure S15: Distribution of rankings per household obtained in survey, when keeping all rankings (left), only peer rankings (middle), and only high-confidence rakings (right).

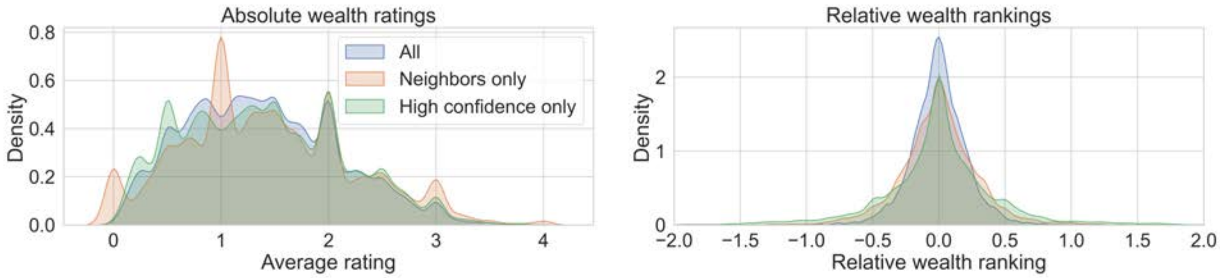


Figure S16: Distribution of aggregated peer rankings produced by averaging absolute ratings of wealth (left) and using the HodgeRank algorithm to aggregate relative rankings of wealth (right).

Table S1: Summary statistics from our 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: Summary statistics from our 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: Drivers of inclusion and exclusion for each targeting method

	Phone-based	**CBT**	**PMT**
Panel A: Household characteristics			
HH head female	0.011 (0.042)	0.035 (0.042)	0.065 (0.038)
HH head age	0.001 (0.013)	0.004 (0.013)	-0.065 (0.012)***
HH head employed	0.023 (0.032)	0.010 (0.033)	0.038 (0.030)
HH head minority	-0.061 (0.072)	-0.003 (0.072)	0.003 (0.066)
HH head widow/widower	-0.018 (0.057)	0.140 (0.057)*	0.005 (0.052)
HH size	0.006 (0.012)	-0.025 (0.012)*	0.154 (0.011)***
Connectedness (in)	-0.036 (0.015)*	0.003 (0.015)	-0.008 (0.014)
Connectedness (out)	-0.020 (0.009)*	-0.004 (0.009)	0.023 (0.008)**
Own phone	0.319 (0.062)***	-0.068 (0.062)	0.006 (0.057)
Phone transactions	-0.086 (0.012)***	0.002 (0.012)	-0.043 (0.011)***
Food consumption share	0.060 (0.012)***	0.029 (0.012)*	0.071 (0.011)***
Panel B: Neighborhood characteristics			
# of Households	0.064 (0.015)***	-0.005 (0.015)	0.024 (0.013)
Land area (square km)	-0.030 (0.014)*	0.012 (0.014)	
Density	0.012 (0.016)	-0.004 (0.016)	-0.026 (0.015)
Urban	0.025 (0.098)	0.018 (0.098)	0.057 (0.090)
% Minority	-0.006 (0.025)	-0.003 (0.025)	-0.042 (0.023)
Connectedness	0.059 (0.019)**	0.005 (0.019)	0.024 (0.017)
Average consumption	-0.013 (0.014)	0.002 (0.014)	-0.089 (0.013)***
Inequality (Gini)	0.007 (0.014)	-0.001 (0.014)	0.041 (0.013)**
Constant	-0.102 (0.067)	0.253 (0.068)***	0.150 (0.062)*
N	1,252	1,252	1,252

Notes: Results of regressions for which types of households are selected by each targeting method (using one train-test split). The dependent variable of each regression an indicator for whether a household was targeted by the method in question. Regressions are run jointly with all explanatory variables in the first column. All explanatory variables are standardized. Connectedness (under neighborhood characteristics) 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. Connectedness (in) under household characteristics represents the average knowledge that other households had of the household in question during the peer ranking exercise; connectedness (out) represents the average knowledge that the household in question had of other households in their community. Regressions are run using data from a single train-test split. Standard errors are in parentheses.

Table S4: 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 7, separately for a PMT with variable costs of $1.25 per household screened (left) and $4.00 per household screened (right).

D National Social Assistance Budgets and Scope

Table D1: 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
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

Continued on next page

Table D1 – continued from previous page

Country	Year	SA budget (mill. USD)	Households (mill.)	Budget per HH (USD)	Best method (BD data)	Best method (TG data)
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
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

Continued on next page

Table D1 – continued from previous page

Country	**Year**	**SA budget (mill. USD)**	**Households (mill.)**	**Budget per HH (USD)**	**Best method (BD data)**	**Best method (TG data)**
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 7. 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.