How much do our neighbors really know? The limits of community-based targeting *

Carly Trachtman[1], Yudistira Hendra Permana[2], and Gumilang Aryo Sahadewo[2]

[1]University of California, Berkeley

[1]Universitas Gadjah Mada

January 11, 2022

[Click Here for Most Current Version]

Abstract

A classical motivation for using information provided by the local community to target social benefits in developing countries is that community members may have more current, dynamic welfare information about others than a centralized program implementer. However, there is little direct evidence supporting this claim, which mostly relies on correlations between community-provided information and survey-collected welfare metrics. To understand what information community members have and use in targeting, we conduct lab-in-the-field experiments and community meeting exercises with 300 families in Purworejo, Central Java. Participants individually ranked other community members based on specific welfare benchmarks (consumption, neediness, and assets) and also completed targeting tasks. We find that community-held welfare information is distinct from information captured using standard survey methods, and seems to reflect longer-term fixed attributes, rather than dynamic welfare information. Accordingly, community members use longer-term wealth information to predict dynamic welfare and to target social benefits. Moreover, we find that community information about more dynamic measures does not outperform simple proxy means test scores in predicting more dynamic survey welfare metrics. Finally, we find community members' information sets are fairly concordant, and rankings constructed during community meetings do not seem to more closely reflect survey-collected welfare metrics. These findings suggest that community-based targeting methods may be useful in identifying long-term poverty, but are less useful in identifying acute short-term distress.

*Many thanks to Ethan Ligon, Jeremy Magruder, and Ted Miguel for guidance. Additional thanks to Megan Lang, Elisabeth Sadoulet, Daniel Agness, Alain de Janvry, Chris Blattman, Prachi Jain, and participants of the UC Berkeley Development Lunch Seminar, ARE Development Workshop, and CEGA Targeting Roundtable for helpful comments. Thanks to Adib (Purworejo), Angky Baskoro and our field team for field assistance and local coordination, and to Khansa Fairuz, Jaide Lin, and Drewya Prasasya for excellent research assistance. This research was approved as Protocol ID: 2020-11-13812 by the UC Berkeley Committee for the Protection of Human Subjects. CRediT Roles for the authors can be found in Appendix Section A.2.

1 Introduction

Peer groups, social networks, and communities can serve as vital information sources, for both outside observers and community members. Such community information may be especially important in the small, rural villages prevalent throughout the developing world, where formal documentation of community members' current welfare status and activities is often scarce. Indeed, the ability of individuals in these "tight-knit" communities to observe and monitor others in the absence of formal records underpins many paradigms in development economics, including models of community risk-sharing, group microfinance lending, and network-based technology proliferation.

This local information may be particularly useful to centralized policymakers who seek to target social program benefits to the households currently most acutely in need, such as COVID-19 relief payment programs. Given the difficulty and cost of observing current income in these settings, most standard targeting regimes feature centralized policymakers attempting to predict households' current welfare by collecting data on more easily observable welfare proxies. Notably, such proxies, which usually include demographic characteristics (e.g., family structure) and/or asset wealth information (e.g., ownership of productive assets), change infrequently, and may not respond to transient income shocks affecting current welfare levels. Intuitively, community members may be in a good position to observe such idiosyncratic shocks faced by others; for example, it seems quite plausible that a community member would know if their neighbor experienced a poor harvest or suffered an illness. Hence community-based targeting (CBT) methods, where local officials or community members identify the households most in need of additional assistance, may be able to provide more dynamic welfare information than standard proxy-based methods [Chambers, 1994].

Despite this intuitive argument, there is little direct evidence that community members have high-quality, dynamic information about their neighbors. The literature to date mainly compares households' CBT allocation outcomes (such as their community-assigned poverty status) to survey-measured welfare indicators that are likely responsive to income fluctuations (such as per capita consumption),[1] and interprets significant positive correlations as suggesting community members have and use information about these relatively "dynamic" indicators in making targeting decisions.[2] However, survey reports about any given attribute likely differ from community information about that attribute, and hence this interpretation may not be valid. Hence, what information community members have and use in CBT exercises remains an open question. Specifically, we ask how dynamic community-held information is, how it compares to survey-captured information, and whether it is more predictive of current welfare status than standard proxy-based

[1] As we discuss further in Section 3, welfare is responsive to income fluctuations when households are unable to smooth their consumption in the face of a negative income shock because they lack access to complete credit and asset markets.

[2] See for instance: Alderman [2002], Galasso and Ravallion [2005], Alatas et al. [2012], Basurto et al. [2020], Karlan and Thuysbaert [2019], Stoeffler et al. [2016].

targeting methods.

In this paper, we directly elicit the welfare information that community members have about their neighbors, and compare it to the information about both more dynamic measures (like consumption and shocks) and more fixed measures (like assets) collected via standard survey methods. After exploring the information community members have, we then examine targeting tasks performed by community members, both individually and jointly, to better understand the information community members use in CBT exercises. Specifically, we surveyed 300 families in 10 communities within Purworejo Regency, Central Java, and administer a series of lab-in-the field exercises at the individual and community levels.

After collecting standard survey modules on the participants' own consumption and assets, we asked the them to individually perform a series of tasks in which they ranked the welfare status of other families in their community as they perceived it at the time of the survey visit. We first asked them to complete a standard CBT task called a "participatory wealth ranking" (see Chambers [1994]), where participants rank 10 families from "poorest" to "richest," knowing that a randomly chosen number of families ranked poorest would receive a small cash transfer. This reflects the common incentives of a CBT exercise, where those identified as neediest receive program benefits. We then elicit participants' information about the welfare of others by asking them to rank the same households according to three specific benchmarks: weekly per capita consumption (dynamic), neediness (dynamic), and asset wealth (more static). Here, critically, we incentivized the participant *herself* for providing us with a *"correct"* ranking, which is defined as the ranking most similar to those based on the metrics we calculate using the other participants' self-reported survey data. Specifically, we calculate per capita food consumption using a simple expenditure aggregate, calculate a measure of neediness using the estimated Index of Marginal Utility of Expenditures [Ligon, 2020], and measure asset value using both a principal component-based asset index and total land value. The benchmark-specific rankings provide insight as to the welfare information individuals have about other community members. Comparing these benchmark-specific rankings to the targeting task ranking provides a sense of the information used by community members when providing information for CBT.[3]

Shortly following these household visits, we invited participants to a CBT meeting in their community. Here, we asked participants to *jointly* perform a participatory welfare ranking, where all thirty participants' families were ranked from poorest to richest. Similarly to the individual targeting task, participants were aware that we would award an additional cash transfer to some number of the families ranked poorest. We can

[3]In addition, we also asked respondents to rank the families based on how they think other community members would do so (to elicit second-order beliefs) and in terms of whom they would most like to receive additional money (without a welfare-related context) to elicit other non-welfare-based preferences. The second-order beliefs ranking task was incentivized for accuracy similarly to the other welfare benchmarks above. The preference elicitation task was incentivized similarly to the targeting task, where a random number of those ranked as most desirable to receive additional money received a small cash transfer.

then see whether there are informational gains from collaboration in targeting exercises, as is the practice in many CBT procedures. Finally, in three randomly chosen communities, we re-surveyed participants approximately three months later, and asked them to repeat the three benchmark-specific ranking tasks.[4] This allows us to see whether individuals' information updates over time in response to changes in others' welfare.

Our three key findings are as follows. First, we find evidence that community members have relatively little "dynamic" welfare information. The dynamic information that community members provide about others is quite different than the information recorded in our survey; the average Spearman rank correlation coefficient of individual participants' reports with the survey-based measures is 0.16 for per capita expenditures and 0.27 for neediness.[5] Notably, this cannot be explained by families not knowing each other well, as eliminating the few unfamiliar families participants were asked about does not significantly change these results. Additionally, participants generally were not able to accurately report whether randomly chosen families in their community had faced a negative shock or received a windfall (COVID-19 related benefits), when asked directly.[6] For both shocks and benefits prompts, participants reported that they "didn't know" the answer in around half of cases, with lack of familiarity with the family in question accounting for only 13-14% of these refusals to answer. Among those who did provide an answer, participants accurately identified self-reported shock victims only 13% of the time (which is no higher than the 13% of all cases in which participants indicate a family has faced a shock) and accurately identified COVID-related social beneficiaries 68% of the time (which does constitute some improvement over the 48% of all cases in which participants indicated a family is receiving benefits). Moreover, in the three communities that we re-surveyed three months later, we find that changes over time in participants' rankings of their neighbors did not mirror changes over time in survey-assessed welfare metrics. This suggests participants were not learning and incorporating new welfare information over this three-month period. Instead, it seems participants in the follow-up round attempted to predict welfare using information they had learned during the community meeting three months earlier, despite the fact that this information was no longer current. Finally and perhaps most critically, we find that very simple proxy-based methods (two proxy means score formulas, which each use 10 easily observable proxies to predict a per capita consumption score) outperform "dynamic" community information in predicting survey-measured dynamic welfare benchmark ranks. This notably contrasts with past work which has suggested that a reason community-identified "poverty" reports do not more accurately predict survey-measured consumption is that the community considers an alternate definition of poverty. Taken

[4] This time, participants were incentivized for accuracy with more lottery tickets for a chance to win a larger cash prize.

[5] Similarly to a standard Pearson correlation, this metric can take values from -1 to 1, with 1 indicating the exact same ranking order and -1 indicating the exact opposite ranking order.

[6] We do not think that this was due to COVID-related social distancing hindering the flow of information, as social distancing was not widely practiced in our study area. Both our field team and local officials confirmed this.

together, this evidence suggests community members observe very little dynamic information about others.

Second, we find that the information community members do have tends to mostly reflect long-term welfare, and that they use this more static information both to predict more dynamic welfare metric rankings and to target transfers to the "poorest" families. The average rank correlation between survey-assessed and participant-assessed assets is significantly greater than the average correlations for the dynamic metrics, at 0.45. Notably, participants provide highly similar rankings for both the dynamic and static welfare benchmarks; the average rank correlation between the community-provided rankings of the dynamic measures (consumption/neediness) and more fixed measures (assets) is much higher than is the case for the equivalent survey metrics. Indeed, community-provided rankings based on all three of these benchmarks are significantly more correlated with survey-measured assets ranks than with any other survey benchmark. This suggests that community members have little dynamic information about other households' welfare and hence use their information on more visible fixed assets to predict these more dynamic welfare measures. In targeting tasks, where participants have to identify poor families to receive additional funds, community members also generally seem to employ more static information about assets. The average correlations between the individual's targeting task ranking and participants' perceived rankings of others based on specific welfare metrics were 0.49, 0.76, and 0.89 for consumption, neediness, and assets, respectively. As stated previously, participants' rankings based on different benchmarks are highly correlated, and after controlling for asset rankings, their neediness and consumption rankings have limited explanatory power in predicting individual targeting task rankings. Participants' qualitative responses suggest asset-based targeting decisions, with differences in land ownership being cited more than any other justification.

Third, we find that, when community members are asked to target transfers jointly, the outcomes are very similar to when individuals perform targeting by themselves. Individual participant targeting task rankings and community meeting rankings are very concordant, with an average rank correlation of 0.65, and most of the distribution of individual/community ranking correlations is above 0.5. Interestingly, we note that, when moving from the individual to community rankings, it does not seem to be the case that much information aggregation occurs. Gains in the correlation with survey-measured welfare benchmarks, when going from individual to joint decision, are modest and not statistically significant. To improve external validity, we support this third set of findings using data from community meetings (and individually generated rankings) in 214 Indonesian communities collected by Alatas et al. [2012], and find very similar results.

Taken together, we find little evidence that community members in this context have current, dynamic welfare information about others. Instead, they seem to have and predominantly use information about the same types of more observable, long-term welfare proxies used in traditional survey-based targeting procedures. This is the case whether individuals complete targeting tasks by themselves or jointly as a community.

Hence, community information may not be as dynamic and robust in some contexts as the economic development literature often assumes. Accordingly, CBT methods may not be always be appropriate in targeting households based on current welfare status.

The rest of the paper proceeds as follows. Section 2 provides some background on relevant concepts and literature, highlighting our contribution. Section 3 describes the theory motivating our use of various welfare metrics, and explains how we estimate them in practice. Section 4 gives some background on our experimental data and methods. Section 5 presents results regarding the information that the community has. Section 6 provides results on the information used by the community when generating welfare rankings in CBT exercises. Section 7 provides results on information aggregation at the community level. Section 8 concludes.

2 Background

Targeting social program benefits to the poorest households is a long-standing challenge in developing country contexts. In such settings, income can be particularly difficult to observe, given high rates of participation in sectors lacking clear income documentation, such as small-scale agriculture and informal sector employment. Thus, economists have traditionally relied on comprehensive survey methods to identify the poorest households, where household welfare status is determined by per capita consumption expenditures collected via a detailed consumption module [Deaton, 1997]. The rationale is that income and consumption expenditures are closely related.[7] Many economists consider survey-collected per capita consumption to be the "gold standard" welfare proxy in such settings. However, detailed consumption surveys are generally expensive to implement and susceptible to measurement error. Hence, alternative, less-expensive methods, that rely on collecting data on more easily observable welfare proxies, are also frequently used.[8] Generally, these more observable proxies consist of durable assets owned by a household and other relatively fixed demographic characteristics, such as household size and composition.[9] These proxies are then combined to form a single welfare metric, sometimes through a dimensionality-reducing procedure like principal components analysis [Filmer and Pritchett, 2001]. Perhaps more common is the proxy means test (PMT) approach, where these proxies are used to predict per capita consumption, with a formula (estimated using a supplemental data source) mapping the relationship between the proxies and consumption [Grosh and Baker, 1995].

[7] Some papers like Chen et al. [2006] consider programs where targeting is based on "income" like the *Di Bao* program in China. However, even in this case, the authors note that survey-reported income is likely measured with significant error, and that local officials participating in the selection of beneficiaries consider "other factors such as financial assets, consumer durables and housing conditions."

[8] For a review of common targeting methods in developing countries, see Coady et al. [2004].

[9] Big data methods that have been developed in recent years have enabled the uses of other types of proxies, such as "night lights" [Henderson et al., 2012], cellular call data records [Blumenstock et al., 2015], and other spatial indicators of welfare observable from satellites [Jean et al., 2016].

CBT methods are theoretically distinct from proxy-based methods, and do not generally require survey data collection. Instead, some subset of local community members (often either any interested participant or local leaders) is asked to provide input about the welfare status of other households. Such exercises can take many forms given their wide use; in fact, over 50% of cash transfer programs in Sub-Saharan Africa have some community targeting component [Garcia et al., 2012]. (We choose to set our study in Central Java given the breadth of targeting work in Indonesia (see Alatas et al. [2012, 2016a,b, 2019], Bah et al. [2019], Banerjee et al. [2020]), for easy comparability to other literature.) These programs often include tasks meant to assess the relative poverty status of households in a local community, such as providing a list of households that are considered "poor", placing all households into various poverty strata "bins", or providing a complete welfare ranking of all households in a community. As mentioned previously, the complete ranking task is often referred to as a participatory wealth (or welfare) ranking, and is the type of CBT exercise considered in this paper. We consider this exercise type because it is commonly used [Banerjee et al., 2009] as part of the "Participatory Rural Appraisal" methodology disseminated in the late 1990s [Chambers, 1994] as part of a wider push for community-driven development methods. Most CBT tasks require joint decision-making of all participants (under the guidance of a facilitator), with the idea that this may decrease incentives to provide biased information reported by any individual. However, there are also iterations in which exercises are completed separately by multiple individuals or in smaller groups, and information is aggregated by policy implementers [Premand and Schnitzer, 2018, Hussam et al., 2017, Dupas et al., 2021, Bloch and Olckers, 2021].[10]

Perhaps unsurprisingly, the households identified as poorest under proxy-based regimes are often different than those identified under CBT [Temu and Due, 2000, Alatas et al., 2012, Premand and Schnitzer, 2018, Stoeffler et al., 2016, Basurto et al., 2020, Beaman et al., 2021, Karlan and Thuysbaert, 2019]. The targeting literature has pointed out many reasons why this may be the case, which we will condense into three main categories.[11] The first category of deviations arises due to differences in preferences over program benefit recipients between centralized program implementers and local community members. While we assume that program implementers have the goal of targeting the poorest or neediest households, community members may have other preferences over who receives benefits, perhaps based on familial connections (nepotism) or political objectives (elite capture). The evidence as to whether such deviations actually create significant differences in targeting outcomes is mixed. Traditional literature [Bardhan and Mookherjee, 2000, Platteau, 2004, Crook, 2003] on participatory techniques is extremely skeptical of local leaders, using their input in the

[10]There is an emerging literature on how to best aggregate rankings of the same individuals made by multiple respondents; however, we abstract from this issue and mostly look at the choices of each participant separately.

[11]There is also a difference between the processes in how errors may arise within the data-generating process, but we set that concern to the side for the moment.

targeting process for their own political gain. There is some experimental support for this as well: Schüring [2014] shows that individuals in Zambia were much more likely to allocate transfers to themselves than to the average participant when given the opportunity. However, more recent evidence shows the distortionary effects of such preferences to be much more understated. For instance Basurto et al. [2020] find that local chiefs in Malawi were slightly more likely to allocate benefits to their kin, and Alatas et al. [2019] find evidence that elites are more likely to allocate benefits to themselves and their relations. However, in both cases, the distortionary effects are minimal; in the former case this is because the kin referred by local chiefs tended to indeed be poor, and in the latter case because the magnitude of the effect is so small. Of course, there may be other strategic preferences over who receives transfers; Basurto et al. [2020] also find that chiefs allocated the fertilizer subsidy benefits (from a different program) to those with the highest returns to farming, which could be efficiency enhancing in places where resource-pooling is common. In both of these settings, we notably do not know much about the information that community targeting participants have about other households. What could seem like nefarious behavior by those with decision-making power could actually reflect individuals having more accurate information about those with whom they are connected socially. Indeed, [Alatas et al., 2016a] find evidence of more "correct" targeting for closer social connections in the Indonesian context.

The second class of deviations arises because community members and centralized program implementers may have different definitions of what "poverty" or "welfare" means. In most CBT exercises, community members are not given a strict definition of "poverty," and are often asked to define it themselves. In other words, communities may not use the types of information used to construct the PMT in their targeting, nor might they have the objective of predicting individual consumption (as PMT scores often do). The third class of deviations arises due to differences in the underlying information set of community members and the information collected via survey. If such deviations exist, even if policymakers and communities have the same definition of "poverty," there could still be differences in targeting outcomes.

We discuss these two classes of deviations together, because most research to date has not attempted to identify them separately. Specifically, evidence regarding what information communities have and use in targeting mostly relies on correlations between CBT allocation outcomes and survey-measured welfare metrics. Perhaps the most seminal work of this form is Alderman [2002].[12] In that paper, cash transfer allocation amounts decided by local Albanian officials are regressed on the welfare information that can be observed by the central government (related to asset wealth), as well as a less observable component of

[12] Galasso and Ravallion [2005] is a classic paper from the same time period supporting the general view that the community may have better information than centralized government targeting bodies, studying the "Food for Education" program in Bangladesh, though they consider village-level targeting outcomes and characteristics. Additionally, more recent papers have directly or indirectly made similar arguments to Alderman's to explain what information is used in CBT processes, for instance: Alatas et al. [2012], Basurto et al. [2020], Karlan and Thuysbaert [2019], Stoeffler et al. [2016].

consumption (collected via survey) which is orthogonal to assets. Because this orthogonal consumption component is significant in predicting transfer amounts, Alderman concludes that local officials have information about consumption, a dynamic welfare measure, beyond that observed by the central government, and then use this information in allocating transfers. Yet this result does not necessarily imply that local officials or community members have and use such community information in targeting, as survey and community information is likely different.[13] A perhaps equally likely explanation of this result is that, like centralized policymakers, community members also mostly observe less dynamic welfare information about others, such as asset ownership. Current welfare status depends on both transitory and permanent income, and, in that sense, current survey-measured expenditures and community reports of (more permanent) assets may also be correlated. Additionally, if the more permanent information captured by the community is not exactly the same as the more permanent information captured via survey, Alderman's results could be explained by a correlation between more permanent community-reported welfare information (that is not captured by more permanent survey welfare measures) and survey-measured consumption.

A limited set of work seeks to identify the information the community has and can use to make targeting decisions. Much of this literature entails validation of key informant interview and rapid rural appraisal techniques used in the late 1990s and early 2000s [Adams et al., 1997, Bergeron et al., 1998, Takasaki et al., 2000, Macours, 2003]. These papers generally compare the information gathered from a few key informants regarding asset holdings and other relatively fixed characteristics of other households with information collected via survey, finding reasonably low rates of discordance (though far from perfect agreement).[14] More recently, (and with a larger data set), Alix-Garcia et al. [2021] asked local leaders in Mexico to answer 10 questions regarding others' ownership of assets, other relatively fixed demographic characteristics, and participation in the community. They find relatively high, significant correlations between leader and household reports, with correlations for indices of the 10 collected variables ranging from 0.69 to 0.77. Additionally, Hargreaves et al. [2007] attempts to directly elicit the kind of information community members actually use in targeting. In this paper, small groups of South African participants were asked to place other community members into various welfare strata piles, and also to provide characteristics associated with each pile. They find that in this context the topics of employment, schooling, and housing were mentioned most frequently. Additionally, they find statements about not having soup, being mentally ill, and being an orphan to be highly associated

[13] Some work using similar methods does acknowledge this limitation in interpretation of the results. For instance, Alatas et al. [2012] highlights this issue, but takes the strong correlations found between various survey-measured attributes and targeting ranks as suggestive evidence that targeting deviations between CBT and PMT are likely caused by a difference in the community's definition of poverty. Basurto et al. [2020] also note this issue, but state that, when asked, chiefs claimed that they did have relevant welfare information about the households in question. Of course, even if chiefs do have perfect information, this only implies that the information that chiefs use in targeting is correlated with the significant explanatory variables.

[14] An exception is Bergeron et al. [1998], which compares more dynamic "food security" ratings by various sub-groups of local farmers in Western Honduras, and finds low levels of concordance between sub-group ratings. This is despite the fact that these farmers all belong to a common farmers' group, and know each other well.

with being poor, and statements about having a big, strong, and renowned business most associated with being more well-off. Of course, such generalizations about welfare strata piles do not necessarily explain how participants make targeting decisions about a particular household.

Our paper contributes to the literature on community information and targeting because we directly elicit the welfare information that many community members have about others in an incentivized manner, and then compare this information to actual targeting outcomes (both at the individual and community level) to get an idea of how this information is used. We believe this is the first paper to complete such an exercise, and is amongst the first to clearly explain the welfare information that community members have and use in targeting.[15] Moreover, we specifically collect community welfare information regarding both relatively dynamic and relatively fixed welfare measures within the same context, whereas most previous work only collects information on one or the other. Additionally, we are able to do this in a scenario where we find little evidence of any deviation in preferences between participants and the program implementers (experimenters in this case), meaning that elite capture and nepotism do not seem to confound our results. This allows us to arrive at a clearer understanding of the information that the community has and uses in community-based targeting.

3 Measuring Welfare: Theory and Practice

3.1 Capturing Welfare Fluctuations Using Observable Metrics

Given that it is difficult to observe fluctuations in income in developing country settings, we focus on three more easily estimable welfare benchmarks: per capita consumption, neediness, and asset wealth. We have claimed thus far that per capita consumption expenditures and neediness are better than assets in capturing intertemporal fluctuations in welfare (transient income). However, it is essential to explain why we think this is the case, as this claim rests on assumptions regarding individuals' access to various markets. Namely, we assume that individuals in our sample generally 1) lack access to credit markets and 2) lack

[15] Notably, Dervisevic et al. [2020] completes a somewhat similar exercise to ours, though the projects were developed and carried out completely independently. They asked 85 leaders in Lao PDR to each rank 15 individuals in order of whom they would most like to participate in a road-building public works program (targeted at poor, able-bodied women), and then subsequently asked those leaders to rank those same individuals in terms of land ownership, food security, and being in need (due to a shock). However, we believe our work deviates from theirs in several important ways. First, while the targeting task in Dervisevic et al. [2020] is framed in the context of a real social program, the ranking task was not actually the targeting procedure for that program; in fact, the actual targeting task was carried out shortly before the ranking exercises. The additional three information-based ranking tasks were also not incentivized, which may exacerbate any desirability bias that elected leaders may face in wanting to make their decision look "fair". In our work, both targeting and information-gathering tasks were clearly incentivized. Additionally, the context in which we elicited this information is quite distinct; we asked community members instead of leaders to provide information, and the benefit being targeted was a general cash transfer instead of a specific public works program, in which local leaders may have other considerations beside poverty status. Finally, we are able to collect (survey and community) information on consumption (which is often economists' preferred welfare proxy) and assets (the primary component of most proxy means scores), as well as to compare our neediness ranking to a more theoretically motivated measure of neediness (the MUE).

access to asset markets/cannot quickly sell off assets given liquidity constraints. A failure in both of these markets is required for the result. Here, for purposes of intuition, we assume individuals lack access to credit markets, and show which metrics we would expect to be more dynamic with complete and missing asset markets.

If individuals have perfect access to asset markets, standard intertemporal models of consumer choice suggest that a consumer will attempt to smooth consumption over time. Specifically, if a consumer faces a negative income shock, they will borrow from their assets, in order to avoid fluctuations in consumption. More formally, suppose a consumer with utility function $U(C_t)$ (with $U'(\cdot) > 0$ and $U''(\cdot) < 0$) faces a choice between consumption today (C_0) and future consumption (C_1). They also earn some income today (Y_0). For simplicity, suppose the individual has a discount factor of 1 for future consumption, and that they cannot borrow. An individual also has a stock of productive assets worth A. As a researcher, we can observe or estimate C_t, $U(C_t)$ and A, but cannot observe or estimate Y_0.

With complete asset markets, the consumer will solve:

$$\max_{C_0, C_1} U(C_0) + U(C_1)$$

$$\text{s.t. } C_0 + C_1 \leq Y_0 + A$$

which will yield a solution of $U(C_0^*) = U(C_1^*)$, and therefore $C_0^* = C_1^* = \frac{Y_0 + A}{2}$. Suppose this individual experiences a very low realization of $Y_0 < \frac{Y_0 + A}{2}$ today. It will then be the case that $C_0^* > Y_0$. Since this person can sell their assets to finance consumption, they will do so such that they can maintain $C_0 = C_1$. Hence in such a market context, we would not be able to detect an (unobservable) fluctuation in income by looking at fluctuations in consumption or the marginal utility of consumption. Indeed, assets would actually be more reflective of an income shock, as a household drawing down on its assets would potentially indicate distress.

However, in many places throughout the developing world, asset markets are incomplete, making it challenging to quickly sell off one's assets. For the purposes of our model, we assume the consumer cannot sell assets today, but can sell them in the long-run/future period. Under these market conditions, while the consumer's problem remains the same, the constraints differ; now $C_0 \leq Y_0$ and $C_1 \leq A + Y_0 - C_0$. Notably, the solution under these conditions only implies that $U'(C_0) \geq U'(C_1)$. If the household experiences an income shock and gets a very small draw of Y_0, the former constraint may indeed bind. As a result, we would observe relatively low $C_0 = Y_0$ and relatively high $U'(C_0) = U'(Y_0)$. Because the household cannot sell its assets, assets would remain unchanged. The graphical intuition of how this smoothing becomes

impossible with a low enough Y_0 can be seen in Figure 1.

Which model do we think is most relevant in the context of this study? Notably, the most economically important assets for individuals in our sample, as can be seen in the results to follow, are things like agricultural land, household dwellings, motorcycles, and refrigerators. These stores of household wealth are relatively illiquid, and generally do not not have complete sale (or even rental) markets. Hence, in our setting, households cannot rely on selling assets as a consumption smoothing strategy. Therefore, in our setting, we believe that consumption and neediness are more "dynamic" and responsive to income shocks than are assets.

3.2 Capturing Welfare Information via Survey

We have noted that we can observe or estimate per capita consumption expenditures, the marginal utility of consumption expenditures/neediness, and asset value. Here we briefly discuss how we collect and estimate each survey benchmark in practice. The distribution of these four survey-calculated welfare measures can be seen in Figure 2. In all cases, there is sufficient variation to be able to identify families that are much richer or poorer than others.

3.2.1 Per Capita Consumption Expenditures

Our survey asks participants to report expenditures on up to 205 food categories in the past week. We choose to focus on food, as food expenditures may be relatively easy for respondents to conceptualize (as opposed to general weekly expenditures) and survey consumption panels tend to concentrate on food items. As is common practice, we simply add reported expenditures in all food categories together and divide by the number of household members. We chose to use a simple consumption aggregate as this is often the "status quo" benchmark used in the targeting literature to compare the accuracy of various methods.

3.2.2 Marginal Utility of Expenditures

In practice, consumption expenditure modules can be sensitive to measurement error, and hence we choose to also use another metric capturing dynamic welfare: the estimated index of marginal utility of expenditures (MUE) metric from Ligon [2020]. Notably, we present the first application of this measure in the targeting literature. [16] Besides a lower sensitivity to measurement error (which results from being

[16] We do not discuss construction of this measure in great detail here, but it can be found in Ligon [2020] and can be estimated using the "CFEDemands" python package. In terms of a basic description, there are two major estimation steps. Step 1 involves estimating a household/good demand system using seemingly unrelated regression (SUR), which allows the econometrician to estimate and account for household characteristic-specific preferences for a given good type (where household characteristics may be number of men, number of women, etc.). Step 2 involves decomposing the residuals from Step 1 using a Singular Value Decomposition (SVD), which allows the econometrician to separate and identify the most explanatory orthogonal components of the variation stemming from the good type and of the variation stemming from the given household. The latter becomes the

able to extract the "signal" from observations of expenditures of different categories of goods), this metric has some theory-related benefits as a measure of dynamic welfare, compared to per capita consumption expenditures. First, there is a clear link from this measure (which is meant to capture the marginal utility of additional expenditure relative to other households) to economists' utilitarian concept of welfare, and it maps nicely to the solution of the potential choice problem a community member faces in allocating transfers (see Appendix Section A.3). Notably, in order for total per capita consumption expenditures to vary one-to-one with utility, it must be true both that the household's utility function is additively separable and that its demands for goods are homothetic. The MUE measure, based on a Frischian demand system, is able to relax the latter homotheticity assumption. There is significant evidence that homotheticity, which would imply that expenditure of every additional dollar is distributed the same amongst the consumers' chosen bundle of goods, does not hold in practice [Banks et al., 1997]. Moreover, because the MUE allows for non-homothetic demands, it can use the additional information given by the types of goods bought by those with higher and lower levels of total expenditures to learn more about the current welfare of all households. For instance, in the absence of observing true income, seeing households buying large amounts of very "total expenditure"-elastic goods likely implies they are well-off, and vice versa. Additionally, the MUE measure has been shown to be quite responsive to income shocks, perhaps more so than total expenditures, and hence may be better at capturing acute vulnerability or neediness due to dynamic shocks [Ligon, 2020].

Practically, to calculate the MUE measure for the baseline survey, we count the period as one time period and all of the communities as one market, facing similar prices. This market aggregation is done both because these villages are geographically proximate and likely face similar prices, and to preserve degrees of freedom for estimation purposes (as we only have 300 observations at baseline). In a similar vein, we aggregate similar types of goods into broader categories to use for estimation, such that we have enough degrees of freedom to get reliable estimates on different types of goods. The categories that are used in estimation, and their estimated expenditure elasticities, can be found in Table A1. These estimates can serve as a sanity check of our MUE result estimates; foods like meat and fruit, which tend to be higher income elasticity goods, have higher expenditure elasticities, whereas staple starches and soy products have lower elasticities. Our estimates control for good-specific household demand that can be explained by household structure, namely: number of men, women, girls, and boys in the household, and log household size.

household's MUE parameter. An SVD process can be thought of as a more generalized form of principal component analysis. Because only the most explanatory components (associated with having the highest singular value) are used, idiosyncratic measurement error will have relatively low distortionary effect on the estimation.

3.3 Asset Value

Given the difficulty in calculating the total value of a household's wealth, we estimate an asset index, as well as total value of land (which is a critical asset of which participants may be able to estimate the value in this agricultural context). The asset index is constructed using a combination of 45 continuous and categorical values, using factor analysis for mixed data methods. These variables include the number of various household durables and assets owned, per capita land area owned, and various characteristics of the household dwelling. A more detailed breakdown of these variables and summary statistics can be seen in Table A2. Notably, only observations containing data for all 45 variables were usable, so not all observations could receive a score. To estimate total land value, we collected a detailed module asking about ownership, land area, and value of (1) irrigated rice fields, (2) rain-fed rice fields, (3) dry land, (4) other household land, and (5) other land owned by the household for the purpose of doing business. We take the total reported value of land from all five categories to get total land value. In cases where participants report owning a certain type of land but not the value of that land, we impute the median value for that type of land (of the five types mentioned previously). If anything, this probably biases us slightly against finding a large correlation between the survey land measure and participant-generated asset rankings, as we ignore any variation in land value between the plots without value reports.

4 Data Collection and Methods

4.1 Context and Sampling

This study was conducted in the Purworejo Regency of Central Java, Indonesia, within the villages (*desa*) of Dlangu, Lugurejo, and Wareng. Ten communities from these villages were randomly chosen to participate in experimental activities. For the purposes of this study, a community was generally defined as a *Rukun Tetangga* (RT), or sub-village hamlet. To select these communities, we first obtained rosters of all families from the village administrations in the three villages. Given that the roster-recorded populations of Dlangu (687 families; 13 RTs) and Wareng (846 families; 18 RTs) are roughly twice the population of Lugurejo (375 families; 8 RTs), we stratify our sampling, choosing four communities from both Dlangu and Wareng, and two communities from Lugurejo. In cases where the number of families in a given RT was less than 40, we combined this RT with the closest adjacent RT that was within the same sub-village division (*Rukun Warga*/RW) to form a sampling unit. This resulted in 30 total sampling units, of which 10 were selected. The number of families in the 10 communities selected range from 46 to 80, with a mean of 58.9. In practice, only one of the chosen communities included a combination of two adjacent RTs.

From here, we randomly selected 30 families and 10 back-up families from each community's village roster, for a total of 300 experimental families (and 100 back-up families).[17] Additionally, we allowed for replacement of experimental families from the list of (randomly ordered) back-up families in cases when we could not reach the selected families. Replacement occurred when we were unable to reach the sampled family after three attempts, or were informed that the family moved recently (and the village roster had not yet been updated). In practice, we substituted in an average of 3.8 of the randomly chosen back-up families in each community to get to our 300 family sample. While this could induce some selection into our sample, we are not deeply concerned, given that these hard-to-reach households are less likely to participate in community decision-making processes more generally. We mandated that survey participants be at least 18 years of age, and requested that the participant be the head of the family or their spouse.[18] The average ages and genders of survey respondents can be seen in Table A3.

For the follow-up survey, we randomly chose three communities (without any village-level stratification), and attempted to re-sample all of the families that we previously surveyed. In practice, we re-surveyed two communities in Dlangu and one in Wareng. We were able to reach 89 out of 90 of the families that we had previously reached, as one family had moved away between rounds. Table A3 shows a comparison of means for various baseline characteristics of both the original survey sample and the sample that was re-interviewed for the follow-up survey. The observable characteristics of the two samples generally do not differ.

Table 1 shows detailed characteristics of the full sample of families. Most households are employed in agriculture, with rice-farming as their primary income-generating activity. The sample is also quite homogeneous ethnically, religiously, and linguistically, with almost all households being ethnically Javanese, Muslim, and Javanese speakers. Most of these households have lived in their current community for many years, with the household head having been born within the village in about 85% of cases. We might think such homogeneity provides a "best case scenario" for possible information sharing, as we might expect a high level of integration and interaction within a community. Indeed, 97% of households report participating in at least one community organization, and 98% report knowing at least one local official.[19] Additionally, we note that there are well-organized local government structures, and making joint decisions at a community

[17] Notably, our sample is of (nuclear) families rather than of households. This was done because in community meetings/affairs in this context, there is generally representation at the family level, and there are some cases where multiple family units reside in one household. The relevant grouping for government documentation is also the family; each has a "Kartu Keluarga," or family card. Therefore, it seemed that, to hold a proper community meeting, representation should be at the family level. However, when we conducted standard survey modules (member roster, consumption module, etc.), we asked questions at the household level to be comparable with other surveys.

[18] In practice, the survey respondent is a family head in 52% of cases, their spouse in 44% of cases, and another family member in the remaining 4% of cases.

[19] Given that this study took place during the COVID-19 pandemic, one might be concerned that information-sharing might have been hindered by social distancing practices. However, reports from our survey team and local officials suggest that COVID safety practices were not generally followed in this area over the period of study activities (except during study activities; relevant COVID-19 protocols are listed in Appendix Section A.4). For instance, entire communities would still gather closely together at the mosque for Friday prayers every week.

meeting would be familiar in this context, even if participants were not necessarily familiar with the specific community targeting task in the experiment.[20]

In terms of welfare, weekly per capita expenditures on food are about $15.57 per week (median), and total weekly expenditures (averaging all annual purchases over 52 weeks) are about $26.74 (median).[21][22] The villages in our study tend to be on the poorer end of villages in Purworejo.[23] Moreover, 92% of sample families report receiving government benefits of some type (which may or may not be directly related to poverty), and 35% specifically mention receiving benefits associated with the COVID-19 pandemic. While most of the study households are not wealthy by any means, it is worth noting that most are living slightly above conventional $1 per day or $2 per day poverty lines. Most participants put themselves on about the third step of a six-step welfare ladder (where one is poorest and six is richest). Participants do own assets like small plots of land, motorcycles, and livestock, and practices such as risk pooling during the lean season are not common. Hence we should note that the welfare information environment in this context may not be comparable to some other developing contexts where asset-ownership is less common or sharing food and income is more common.

4.2 Data Collection

4.2.1 Baseline Survey

The main survey was administered through household visits in March to April 2021. During this visit, a participant would complete the survey and perform the experimental tasks directly afterward. The survey collected a roster of current household members, basic demographic characteristics, information about participation in the community, a detailed consumption module, a detailed asset module, information about shocks faced, and information about benefits received.[24] The consumption module is quite comprehensive, enquiring about expenditures over the past week (and other self-produced or gift consumption) on 205 different food categories and 28 nonfood categories. The asset module included sections on land value/area, housing characteristics, and ownership of 39 different durable items.

Notably, in the modules about shocks experienced over the past year and benefits currently received

[20] Notably, as communities get more experience with such targeting activities, they may begin to act differently. For example Schüring [2014] finds that those who have more experience with the targeting task in question tend to act more selfishly in their experimental targeting tasks in Zambia.

[21] We use a rounded 2020 purchasing power parity exchange rate of about Rp. 4,700 for US$1 [OECD, 2021].

[22] For context, these figures are quite comparable to 2020 estimates of national per capita expenditure in Indonesia of about $19.40 per week [CEIC, 2021], and of 2015 estimates of per capita expenditures in Central Java (around $35.31) and Purworejo Regency(around $30.35)[Ubaidillah et al., 2019].

[23] While we do not have statistics on household expenditures at the village level, a village-level welfare proxy is the percentage of households with an official government poverty letter. The average for Purworejo is 6.6%, while it is 9.6% in our study villages [Indonesian Central Statistical Body, 2021].

[24] Modules were heavily based on the survey used in Alatas et al. [2012] and Wave 5 of the IFLS [Strauss et al., 2016].

from the government, participants were not only asked about themselves, but also about other families. We randomly and independently generated two three-family sets. In the shocks section, participants were asked whether each family in the first set had faced any shock in the past year, and, if so, what type. Similarly, in the section on benefits, participants were asked whether each family in the second set was receiving benefits specifically associated with the COVID-19 pandemic.

4.2.2 Individual Experimental Tasks

Directly following the main survey, the participant would complete the experimental game section. To begin, the participant was shown 20 index cards in a pre-randomized order, each showing the name of another family (participating in the experiment) in their community. For each family, the participant was asked how well they know this family, with six possible responses: 1) "Not familiar at all," 2) "Know of/can recognize someone in the family," 3)"Have talked to a member of the family before infrequently," 4) "Have talked to a member of the family before frequently," 5) "Close friend or family member," 6) "Close colleague at job/position." The breakdown of responses is shown in Table A4. We see that there is a high level of familiarity with other community members; about 2/3 of all responses fall in categories 4, 5, and 6, which we consider to mean the participant "knows" the household well.[25] Additionally, Figure 3 shows the distribution of the number of community members known well by each participant. The average participant knows about 13.2 (median 14) out of the 20 families presented. 18% of participants claim to know all 20 families well, and only about 8% of all participants report knowing 5 or fewer of the families presented.

After this, the enumerators selected 10 families out of the 20 presented to be used for the rest of the experimental exercises, using a pre-determined algorithm (as it likely would have been cognitively taxing to rank more than 10 simultaneously). Enumerators were instructed to pick the first 8 family cards from the pre-randomized order that the participant claimed to know well (in that they gave a response of 4, 5, or 6 to describe their relationship), and the first 2 family cards that the participant claimed to not know as well.[26][27] The reasoning behind this selection procedure was as follows. We are mostly interested in understanding what information community members *can* provide about others. Various work has already established that unknown households tend to be targeted less accurately than known ones [Alatas et al., 2016a, Premand

[25] At this stage, participants were not aware whether or how this information would be used within the experiment, so strategic misreporting seems unlikely.

[26] In cases where fewer than 8 families were known (16% of all cases), all known family cards were used, and the rest of the 10 families were unknown families. In cases where 19 or 20 families were known (22% of all cases), all unknown families were used, and the rest of the 10 families were known families.

[27] This procedure was implemented correctly by enumerators in 95% of cases, and the median number of known households ranked was 8 (mean= 7.7). Notably, participants' own family names were sometimes on one of the 20 cards, with the intention that households might actually have to rank themselves. However, there was confusion over this point, and enumerators did not include a participant's own family card in the final 10 if it was one of the first 8 known households. This is not being counted as an error in the rate above, as this was systematically done; no participants ranked their own families.

and Schnitzer, 2018], and logically it is likely that participants do not have better information about families they do not know compared to those they do. Given this, we chose to mostly focus on known households. However, more widely known families are also likely to be a selected set; to partially counteract the resulting bias, we also include some unknown families. Using this process, each family was ranked an average of 9.6 times by other families (median=10), though this ranges from 0 to 20 times. Only 2 families were ranked 0 times.[28]

After the 10 cards were chosen, participants completed various tasks involving ordering the family cards in response to different ranking prompts. Cards were shuffled by enumerators between each task to ensure that participants had to make a new ranking each time. First, participants were given a practice task to make sure they understood the idea of ordering the cards. This task asked participants to order the cards alphabetically (with the help of a visual reference to the alphabet if necessary). If this was done incorrectly the first time, the enumerators reiterated the instructions and had the participants re-rank the cards until the ranking was generally correct (one or two mistakes were allowed).

The participants then started the six main experimental ranking tasks. Table A5 presents the key elements of each task for easy reference. We call the first task the Individual Targeting Task. This task asked participants to rank the 10 family cards from "poorest" to "richest" without any further structure on how participants should define welfare. The incentive associated with this task (which was clearly explained to participants) is that we would draw a random number after the experimental session between 0 and 10, and that this number of people that they ranked "poorest" would receive an additional payout of Rp. 5,000 (about US$1.06).[29] These payout recipients would not be informed that this money was based on the participant's ranking (or that it was a payout from another participant at all).[30] This task was designed to essentially mimic the instructions and incentives of a typical participatory wealth ranking community targeting task, where some unknown number of families marked poorest receive a benefit of some type.

We call the second task the Preference Elicitation Task. For this task, the incentives were essentially the same as the prior task; participants were told that some random number of families would receive an additional Rp. 5,000 payout. However, this time, they could make a ranking using any criteria that they wanted, and were simply to rank families from "most preferable to receive an additional Rp. 5,000 payout" to "least preferable to receive an additional Rp. 5,000 payout." The idea of this task was to try to elicit any non-welfare related preferences that might influence the ranking process, such as nepotism.

[28]There are 66 participants that ended up ranking at least one non-experimental family because the family ended up later being replaced. These are not included in the above statistics, but there are 109 cases (out of 3000 participant by ranked family pairs–about 3.6% of all possible pairs) where a participant ranked a non-participating family.

[29]Rp. 5,000 is not a large sum of money, but participants confirm that it is not an amount of money that they would regard as trivial.

[30]All such payments were simply delivered as a lump sum with other incentive payments from the experiment, of which participants also did not know the amount they would receive.

The final four tasks were presented to respondents in a randomized order, to mitigate the effects of framing biases or experimental fatigue on the results of any one task. Three of these remaining tasks were information elicitation tasks corresponding to specific welfare benchmarks: per capita food consumption, asset wealth, and neediness. Specifically, participants were asked to rank families from lowest per capita weekly expenditure (lowest value of assets owned/most needy of additional money) to highest per capita weekly food expenditure (highest value of assets owned/least needy of additional money). Critically, in these tasks participants themselves were paid based on the "accuracy" of their provided response; participants were told that they would be paid more the closer that their answers were to what we calculate in our survey data. This incentive structure was designed to elicit the information about these benchmarks that participants actually had, devoid of the other potential incentives to misreport in the prior targeting tasks. While we didn't provide any additional information on how calculations with the survey information would be done, individuals were at least aware of the type of information collected, as they had also taken this survey just before completing these tasks.[31] The remaining task was meant to elicit second-order beliefs of respondents. Participants were asked to rank the families in a way that they think other participants would have done in the initial targeting task, and again were told they would be paid based on how close their rankings were to other respondents' rankings.[32] This task was meant to help tease out any sort of individual-specific preferences over transfer recipients, which participants did not think would be shared by the community at large.

In general, participants completed all of these tasks successfully. Ties were not allowed in any task, nor were households allowed to be excluded from a ranking. Hence, for all but three participants (in which one or two tasks were accidentally skipped), we have six complete rankings of 10 households. The total amount of incentive payments received by each participant, combining any transfers received from the first two rounds and any incentive payments from the latter two rounds, ranged from Rp. 15,000 to Rp 180,000 (about US$3.19-$38.29).[33] The average payment was Rp. 80,183 (US$17.06) and the median was Rp. 75,000 (US$19.56). These are of course significant amounts in this context, with the median payment value surpassing median weekly per capita food expenditures. All task-related payments were provided in a lump sum to each participant at the community meeting that followed, with no information as to the composition of payments stemming from each task.

[31] In practice, payments are scaled by the Spearman rank correlation between our survey information and participant responses for the same welfare benchmark.

[32] In practice, we did this by calculating the most common pairwise ordering among all participants' rankings between each pair of families ranked in the community, and scaled the payments by how many of each participant's pairwise orderings matched the most common orderings by all participants.

[33] This is in addition to Rp 25,000 (US $5.31) for survey participation.

4.2.3 Community Meeting Task

In each of the 10 communities, we held a community meeting exercise, where all thirty participants (or members of their family) were invited to participate. We visited each family to inform them about the date and time of the meeting in their community. Meetings were held on evenings and weekends at local community centers in order to maximize turnout. To encourage attendance, besides an additional Rp. 25,000 payment for participating in the meeting, we also used the meeting to dispense any incentives earned during the household visit. Moreover, on the day of the meeting, if there were participants who had not yet arrived around 15 minutes after the designated time, we attempted to contact them via a phone call or visit. Given these efforts, 90% of families sent a member to this community meeting; 63% of households sent the same respondent as had participated in the survey and individual tasks; and 27% sent another member of the family instead. Table A3 shows that there is no clear statistical difference between families who participated in the community meeting and those who did not. However, a family's community meeting participant was more likely to be male than its survey/individual task participant. In both the survey and the meeting, most respondents were women, but in the community meeting almost half of the participants were men, versus about 40% of survey respondents.[34]

At the beginning of each meeting, community members were tasked with brainstorming characteristics they associated with the words "poor" and then "rich" as a group. An enumerator guiding the meeting wrote all suggestions down on a large notepad in the front of the space, to be in easy view of all participants. Such an activity is a common activity used to frame participatory wealth ranking exercises. After this, participants were tasked with ranking all thirty experimental families in their community from poorest to richest, regardless of whether a representative from that family was present, together as a group. We explained to respondents before starting that at the end of the meeting an enumerator would draw a number between zero and thirty from an envelope, and that number of households ranked poorest would receive an additional Rp. 50,000 (US$10.64) payout. Enumerators who were facilitating the meeting provided as little input as possible, allowing the community to discuss freely for as long as they needed to reach a consensus. When a full ranking was completed, the community members were then tasked with voting to approve this ranking. If it was not agreed upon by a simple majority, then participants had to resume discussion. Additionally, if communities reached an impasse on a particular card's placement at any time during the exercise, that would also be put to a vote. In general, these instances were minimal. The average number of total votes per meeting was 2.3 (median= 2), including the required final vote. Discussion at these meetings was generally quite lively, and the ranking activity took about 1 hour and 36 minutes on average (median

[34]The fraction of meeting participants who are male ranges quite a bit between communities from about 23% to 67%.

= 1 hour and 41 minutes). The number of families drawn to receive the additional payout for being ranked poorer ranged from 6 to 13, with a mean of 8.4 (median = 8).

4.2.4 Follow-Up Survey

The follow-up survey was conducted in the three randomly chosen RTs in June-July 2021. In this round, participants were administered an abridged version of the original survey, containing only the consumption, asset, and shocks modules, with the shocks module focusing on shocks reported between surveys.

We then asked participants to repeat the three benchmark-specific information elicitation ranking tasks only (based on consumption, asset value, and neediness), with the same set of families that they had ranked in the first survey. The goal was to see if participants would change their rankings to reflect any changes in welfare reported between surveys. This time we incentivized the provision of correct answers with additional tickets in a drawing toward a prize of Rp. 600,000 (around US$128), awarded to one participant in each of the three communities, instead of with cash payments. This was done for logistical purposes; during the first round, incentive payment amounts were calculated and then distributed at the community meeting to follow. In the second round, no such meeting occurred, and hence we would have had to re-visit all households to disburse payments. Of course, this may have some effect on the amount of effort that individuals exerted in the ranking tasks in this second round.

4.3 Analytical Methods: Measuring Concordance in Ranked Data

Here we provide a brief description of the analytical methods used throughout the results section, for easier interpretation of our results.

4.3.1 Spearman's ρ

Participants in the experiment were asked to construct ordinal rankings, and hence measures of concordance that are designed for cardinal survey measures may not be appropriate. Further, we make no assumptions about the underlying distributions of individuals' information sets and rankings produced. Hence, we quantify concordance between two rankings using a non-parametric statistic, Spearman's rank correlation coefficient (ρ). Given two ranked vectors of (the same) n elements, the formula to calculate ρ is:

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

where i indexes the element to be ranked, and d_i indicates the difference in rank positions that the same element takes in the two ranking vectors.

While there are various ways to compare the concordance between two vectors of ranked data, we choose Spearman's ρ for the following reasons. First, it is simply a special case of the standard Pearson correlation (where differences in rank position instead of values are used in computation), and hence interpretation is fairly intuitive. For example, two identical rank vectors will have $\rho = 1$ and a vector with the exact opposite rank ordering will have $\rho = -1$. Second, the Spearman correlation essentially measures the monotonicity of a relationship between two ranked variables, and is relatively insensitive to small deviations. Specifically, ρ is decreasing in the squared distance between various elements, and we can think of it as explaining how well we can fit a line between variables' ranks, in an OLS-type sense. Moreover, because Spearman coefficients are decreasing in squared rank, the value of ρ is relatively insensitive to small deviations in rank ordering, and quite sensitive to large deviations in rank.[35] This is an attractive property for our application, as we care less about specific ordering between similarly poor or similarly rich families, but rather about large deviations, such as a poor family being ranked as relatively rich and vice versa. Additionally, this means that our concordance measures will be less significantly affected by small measurement error in our survey welfare metrics than for other rank concordance statistics (like Kendall's τ); as long as the noise in the values of these metrics does not induce large changes in rank order, bias should be relatively minimal.

4.3.2 Hypothesis Testing

For each individual participant i on a welfare metric j, we can calculate $\hat{\rho}_{ij}$, with participant-assessed ranks of j, and ranks of j based on the values of the survey-collected data. What we attempt to test for welfare metrics $j = 1, 2$ is the null hypothesis that $H_0 : \mu_{\hat{\rho}_1} = \mu_{\hat{\rho}_2}$.[36] Because we are assuming nothing about the underlying distribution of our estimated $\hat{\rho}_{ij}$ or any of its underlying components, we use a bootstrapped test of means. Specifically, we re-center the distributions of $\hat{\rho}_{i1}$ and $\hat{\rho}_{i2}$, so that they share a common mean, and then (for 5,000 iterations) randomly draw 2 subsets of 300 observations each. The p-value of the test is the fraction of times in which the difference in the means of the two redrawn samples is more extreme than the true one we observe. If the p-value is less than or equal to 0.05, we reject the null hypothesis. Throughout the paper, whenever we note that two average estimated Spearman correlations are statistically different, we are referring to rejection of the null hypothesis using this procedure.

[35] For example, if we have a vector of 10 elements, and compare this to a similar vector, where the rankings of one adjacent pair is switched, $\rho = 0.99$. If we increase the distance between elements to swap from one to two, then $\rho = 0.94$. On the other hand, if we swap the rank position of two very far away elements (say a distance of 6 away, such as swapping elements 2 and 8), ρ drops dramatically to 0.56.

[36] We also compare Spearman correlations that do not capture concordance between survey and participant-assessed welfare, but this example is illustrative of the general idea.

5 Results: Information Held by Community Members

5.1 Rank correlation between participant-assessed and survey-assessed welfare metrics

The average Spearman ranking correlations (among survey participants) between participants' benchmark-specific rankings and the rankings suggested by our survey calculated metrics are displayed in Table 2. For all three welfare benchmarks (expenditures, neediness, and assets), participant-assessed and survey-assessed rankings are far from perfectly concordant, and the average correlation coefficients are statistically significantly different from one. Yet, encouragingly, the average correlation coefficients are positive and statistically different from zero for all benchmarks. This means that on average there is at least some agreement between benchmark-specific welfare rankings reported by participants and those suggested by our survey. Notably, the average concordance for our less dynamic welfare metrics (0.45 for the asset index and 0.40 for land value) are significantly higher than the correlation for our more dynamic welfare metrics (0.16 for expenditures and 0.27 for MUE/neediness). This suggests that participant-assessed asset information is much more similar to survey-assessed asset information than participant-assessed expenditure (neediness) information is to survey-assessed expenditure (neediness) information. Interestingly, participant and survey neediness information is more highly correlated on average than is the case for expenditures. It is possible that participants have a general sense of who is most in need, while having less of an idea of actual consumption expenditures.

Additionally, if we calculate these average correlations using only families that are reported as "well-known" by participants, we do not find any significant changes in the results. (See bottom panel of Table 2.) Hence, low concordance between survey and participant welfare reports cannot be explained by participants not knowing other families well. We also might wonder whether considering only averages disregards some underlying trends in the distribution of correlations. We present the distribution of the same participant-assessed and survey-assessed benchmark-specific welfare rankings in Figure 4. These distributions do not provide any notable challenges to this story; the distribution of per capita expenditure information correlations is visually much more symmetrical than the other distributions, which have more probability mass at higher correlation levels. When comparing the neediness information correlations to the asset/land-based correlations, there are notably fewer "near perfect" correlations in the neediness distribution. We conclude that participants seem to have a better sense of other community members' long-term welfare (correlated with assets) than of their more dynamic welfare status (correlated with consumption expenditures/neediness), though their knowledge of both types of welfare is limited.

5.1.1 Robustness

The low correlations above may not actually indicate limited welfare information about other households, if there are data (or procedural) issues biasing these correlations downward. Here, we think through some alternative explanations for the results and rule them out as the main source of low correlations. First, these results could be sensitive to our choice of survey-based metrics. This issue is explored in Tables A6, A7, and A8, for consumption, neediness, and assets respectively. For per capita consumption, we first consider an alternate definition of household size using an adult equivalency measure, to control for the different consumption needs of men, women, and children. This does not significantly change the correlations between survey-assessed and participant-assessed expenditures. We also consider per capita consumption value, which includes not just expenditures on food, but also self-production and gifts/transfers, which could be relevant if neighbors can easily observe how much a family consumes, but not how much is purchased. Yet this does not yield different results either. Finally, we can get fitted values for total expenditures as part of the estimation process of the MUE measures. The intuition is that we can use the estimates for the households' MUE and the estimated good-specific elasticities to get a prediction of a household's total expenditures. This predicted expenditure value is less sensitive than the raw aggregates to random noise in consumption reports, because the estimated good-specific elasticities are fairly insensitive to outliers. This method ends up producing total consumption values that are quite similar to the actual consumption values (Pearson correlation is 0.86), and hence the survey-assessed versus participant-assessed information correlations remain fairly unchanged.

To test robustness for neediness information correlations, we recalculate the MUE using the value of total consumption, rather than just expenditures, because it's possible that households with other sources of food might be less needy. However, the correlation between survey-assessed and participant-assessed neediness information rankings remains essentially unchanged. In terms of the robustness of the assets results, we have already presented two asset measures (an asset index value and total value of land), and have shown that the results are quite similar. To be thorough, we also consider per capita land value, land area, a raw unweighted count of the number of household assets owned, and an index consisting only of housing characteristics. The survey-assessed versus participant-assessed asset ranking correlations for these measures are a bit lower than the case when we use our preferred asset measure. Yet, these measures tend to use less information than that considered by the asset index, and hence this is unsurprising. In any case, these alternative asset metric correlations are all significantly higher than our preferred per capita expenditures and neediness correlations (at $\alpha = 0.05$), except for the asset count measure correlation, which is not significantly different from the MUE correlation.

We also may wonder whether we are asking for ranking information from too large a set of community

members, and whether some identifiable individuals may have better information. Specifically, it's possible that more central or "well-known" people might systematically have significantly better information than others or might be much easier to rank than others. Though we are underpowered to do much analysis at the individual ranker level, we can do some analysis of this question at the ranking participant by ranked family level. While we do not have any sort of detailed data on social networks in these communities, we do know what percentage of times families are reported to be well-known by participants when they are presented on one of the 20 index cards. Hence, in Table A9, we regress the absolute difference between a family's relative survey-assessed and participant-assessed rank order positions for our three main welfare benchmarks (per capita consumption expenditures, neediness, and assets) on the fraction of the responses in which both ranking participants and ranked families were reported to be in one of the well-known categories. We also control for the number of times each family was asked about, because the set of families for each participant to rank was generated randomly. The average family was reported as well-known by other participants 65% of the time. We do not see any evidence that more well-known individuals provide rankings that are closer to those suggested by the survey, nor that more well-known families are ranked more closely to their surveyed rank. If anything, more well-known rankers provide less concordant expenditure rankings with the survey than less well-known rankers.

We may also worry that asking participants to provide complete rankings of all families is a more complicated task than necessary if the goal is to find the poorest families based on a given welfare benchmark. It could be the case that participants can identify which of the families are generally poorest by each benchmark, but have more trouble figuring out the exact ranking between families. To see if this is true, we look at the poorest households as assessed by each survey-assessed welfare measure, and see how many of those were also ranked poorest by the participant. The results of this exercise can be seen in Table A10. We see that, regardless of whether we define the "poverty line" as being ranked as in the 2, 3, or 4 poorest families and regardless of the benchmark considered, participants do not successfully identify anywhere close to all of the poorest ranked families. Even for assets, which is the most familiar benchmark, participants correctly ranked the survey-assessed poorest 4 participants as such in 58% of cases. This percentage is even lower for our dynamic welfare benchmarks: about 47% for expenditures and 53% for neediness. These numbers seem especially low given that, if we chose at random, there is a 40% chance that any given family would be ranked as one of the bottom 4 poorest. So, while the community does provide information (their answers are better than random selection), these statistics do not seem to challenge our previous assertions.

5.1.2 Survey Measurement Error

While we have in some sense been conflating survey metrics with the "truth," we know that in practice survey metrics can be very sensitive to measurement error, and that participants might make some mistakes. Hence, we may be worried our rank correlations are artificially low due to such errors. While we can't rule out measurement error playing a role in deflating these correlations, we will argue that it is unlikely to fully explain the low correlations. First, while survey-based expenditure aggregates are notoriously noisy (because measurement error from each reported food category is additive), both the MUE and asset index are much less sensitive to random errors. The MUE and asset index are both estimated with dimensionality-reducing procedures that identify the strongest common signal that emerges from multiple correlated pieces of information (for MUE, expenditures over multiple categories of food, and for the asset index, ownership of various asset types). As long as the errors between these various pieces of information are uncorrelated, measurement error for any one piece will have little impact on the value of the estimated metric.

While this provides some credence to our findings, it would be useful to have some sense of the magnitude of the measurement error's impact in deflating these rank correlations. As noted above, the Spearman correlation coefficient has some attractive properties in reducing the impact of measurement error on outcomes. Namely, Spearman coefficients are relatively insensitive to small deviations in rank position. So, we might expect that relatively little measurement error would only create negligible bias.

To test this, we perform some simple simulation exercises to quantify the magnitude of potential bias due to measurement error. Noting that the empirical distributions of our measures of interest are roughly normal (see Figure 2), we suppose the distribution of "true values" of a welfare metric are normally distributed with mean μ and standard deviation σ.[37] We then repeat the following procedure 5,000 times:

1. Randomly draw 300 values from a normal distribution with mean μ and standard deviation σ, to represent the true value of the welfare metric value for the 300 participants.

2. For each participant, randomly take two independent draws from uniform distributions over the following support ranges: 1) [$-\sigma/2$, $\sigma/2$], 2) [$-\sigma$, σ], 3) [-2σ, 2σ], and 4) [-3σ, 3σ]. All of the draws represent "noise levels", where larger ranges of potential values indicate more noise. One draw from each distribution represents potential noise reported in the survey metric and the other draw represents noise in a participant's report of any given metric. Add the "true value" to the various measurement error draws. Each participant now has one "true value", four estimated survey metrics (at four noise levels), and four estimated participant metric values (at four noise levels).

[37] Given the nature of the exercise, the values of μ and σ will not affect the results. But in practice we use the mean and standard deviation associated with the empirical distribution of log consumption.

3. Assign each of these 300 participants to a community, and randomly draw 10 other participants in their community for them to "rank." Calculate relative ranks for these 10 ranked participants from 1 to 10, under both the "true" metric and the noisy observations of the metrics.

4. Calculate the average Spearman coefficients (over the simulated ranking participants) between the noisy "survey" and noisy "participant" rankings for all combinations of noise levels. (We also calculate the Spearman correlations at different levels of survey noise, in the case where the community report does not contain error.)

Looking at the average (over iterations) of simulated average (over participants within an iteration) Spearman coefficients under various levels of survey can give us a sense of how much noise would be needed for our low correlations to be fully explained by noise. The results of this exercise can be seen in Table A11. We see that, in order for our low correlations for the "dynamic" survey measures to be explained fully by measurement error, there would have to be two to three standard deviations of noise on both survey and participant reports. For our asset measure, slightly less noise would be needed, but either the participants or survey still would have to have at least 2 standard deviations worth of noise. We also see that relatively low levels of noise (one standard deviation or below) have a relatively small effect on the rank correlation. Hence, while it is likely that some of the low correlation levels could be attributed to measurement error in survey or participant reports of our three welfare benchmark metrics, there would need to be a pretty large amount of error for this to completely explain our results.

5.2 Correlations between rankings produced under different welfare benchmarks

The above evidence suggests that participants may not have much information about the dynamic welfare status of others in their community. If this is the case, we want to understand how participants are constructing the rankings for these more dynamic measures when they complete the ranking tasks. We present suggestive evidence here that participants are using their knowledge of assets/long-term welfare in order to attempt to predict shorter-term welfare metrics.

Our first piece of evidence supporting this argument is the high level of correlation between participants' more and less dynamic welfare metric rankings, which is higher than the correlations within similar survey-based welfare metric rankings. Table 3 shows the rank correlation between different survey-assessed welfare measures (left panel), and the rank correlation between participant-assessed welfare rankings (right panel). Notably, we see that the average correlation between participants' expenditure and asset rankings is 0.52, compared with a correlation of 0.42 using our survey calculated measures. Similarly, the average correlation

between participants' neediness and asset rankings is 0.77, compared with a correlation of 0.44 using our survey calculated measures. Hence, it seems that participants are providing very similar information for both more and less dynamic measures, more than we would expect the correlations between these metrics to be, based on our survey assessments. In fact, about 20% of participants provide exactly the same ranking for all three welfare benchmarks (despite the cards being shuffled between each round). Also of note is that our two more dynamic welfare measures (expenditures and neediness), are highly correlated with each other in the survey-based data, as expected, but are much less correlated in the participant-based ranking data.

If participants are generally using the same underlying information in reporting both more and less dynamic welfare metrics, we may wonder what this information is. Notably, participants' rankings of the two more dynamic metrics (expenditures and neediness) are actually significantly more correlated with our survey-based asset index than they are with their respective counterpart measures that are supposed to capture the same type of welfare. This is shown in Table 4. These correlations are almost as high as the correlation between the asset index and participants' asset rankings. This suggests that the common information participants are using when reporting both more and less dynamic welfare measures is more reflective of longer-term welfare information. Hence, it seems likely that participants are using their knowledge of others' less dynamic welfare attributes to predict their more dynamic welfare attributes.

5.3 Knowledge of shocks faced and benefits received by others

We might think that participants are able to observe concrete "shocks" faced by other households, even if they do not know how such shocks would affect the welfare benchmarks upon which we asked them to rank others. Hence, after asking survey participants whether their own family had faced various types of negative welfare shocks in the past year (such as death, illness, harvest failure, etc.), we randomly chose three families from their community and asked them to report directly on shocks faced by those families.[38] With this random selection process, participants were asked about shocks suffered by other families, with 2.8 inquiries on average per family (median= 3 times), though this number ranges from 0 to 7. Only one family was asked about zero times. For each of the three randomly chosen families, we first asked the participant whether that family had faced a negative economic shock in the past year. If the participant responded in the affirmative, we then asked them to clarify the type of shock faced. Respondents should have had a good idea about the types of negative shocks we were referring to, given that they had just answered similar questions about their own family.

In about 30% of cases, survey respondents report that they have faced a negative economic shock of some

[38]This was done before all of the experimental ranking tasks began, and the draw of names was fully independent from the randomly chosen families used in the ranking exercises.

kind. The majority of these reported shocks have to do with reaping a small harvest or total harvest failure, which denotes a significant loss of income for agricultural households. Also reported in smaller percentages are employment loss, and illness or death in the family. The few respondents facing "other" shocks mainly report losses of business income due to the COVID-19 pandemic (either due to lower demand for their goods or having to watch children instead of work).

Generally, we find little evidence that participants can accurately report shocks faced by other families. First, unlike the ranking tasks, where we strongly encouraged respondents make complete rankings for every task, we allowed respondents to state that they did not know whether or not a family had experienced a shock. As a result, out of 900 possible pairs of a participant and an "asked about" family, participants declined to provide an assessment in 386 cases (roughly 43%) because they did not know. We may think that this could reflect participants being unfamiliar with the families that they are being asked about. Yet, out of all 386 cases where participants declined to provide an assessment, in only 55 (or 14%) did the participant say that they could not provide an assessment because they did not know the family.

The participants that did choose to respond provided very different reports about shocks faced than those reported by survey respondents. These results can be seen in the top rows of Table 5. Overall, participants are much less likely to report that others faced shocks (13.4% of cases) than those families reported themselves via the survey (30.5% of cases). However, community information could still be potentially useful if participants are more likely to report a shock for a family that indeed reported a shock than to report a shock for a family that did not report a shock. Overall, this does not seem to be the case either; families who reported facing a negative shock were identified as such by other participants in 12.7% of the time. Given that participants reported that other families faced shocks about 13.4% of the time (for reports about both those who actually faced shocks and those who did not), it does not seem that the community is providing much information. Moreover, participants seem to be unable to identify others who faced harvest shocks, which are the most common in this setting, with 0% of those experiencing harvest failure correctly identified, and 1.3% of those facing a small harvest correctly identified.[39] If we look at the overall match rates in the "% Match (All)" column, the picture looks much better; the percentage of times where the survey respondent and participant report the same shock status (not conditioning on the household having faced a shock) is much higher in most categories. This is primarily driven by the fact that most families have not faced any given type of shock, and by the relative infrequency of participants reporting that other families faced a shock. If nothing else, it does not seem that participants report other families facing shocks that the family itself does not report.

Similarly, we want to know whether community members can observe positive shocks or windfall gains

[39]If we consider these two types of harvest shocks as one shock type, the accuracy improves, but only to 2%.

received by others that may mitigate the effects of negative shocks. Given the salience of the COVID-19 pandemic at the time of the study, we asked participants about whether three (independently) randomly chosen families had received any type of additional social benefits related to the COVID-19 pandemic.[40] Similarly to the shocks question, a participant was asked about another family 2.8 times on average (median=3 times), with the number of inquiries per family ranging from 0 to 6. For these COVID benefit questions, even more participants declined to answer; in fact, 471/900 (52%) of participants provided no assessment. Again, this is not because the families asked about are not well known; this explanation was only given in 59 cases, or about 13% of all refusals to answer. However, when participants were willing and able to make a report, they could provide some information about COVID benefits received. This results can be seen in the bottom row of Table 5. Participants accurately identified COVID relief beneficiaries in 68% of cases, compared to 48% of all assessment cases (whether benefits are received or not) where a family is identified as a beneficiary. These percentages are statistically different, and hence the community is providing some additional information on the beneficiary status of COVID relief programs. Yet, of course, this 68% of beneficiaries is also statistically different from 100%, and far from perfect targeting. We also note that, unlike the case of negative shocks, participants generally overestimated how many other families were receiving benefits; receipt of benefits was reported by the families themselves via surveys in only around 37% of cases. Moreover, families who did not report receiving COVID relief benefits were identified by participants as beneficiaries in 37% of cases. Indeed, because of this over-reporting, we note that the overall match rate among all observations is almost identical to the match rate for identifying families that faced any type of negative shock.[41]

In general, we see participants having little information about shocks faced and benefits received by others, or at least information that is quite different than what is collected via survey. For both shocks and benefits, in around half of cases, individuals do not have enough information to provide an assessment of the family asked about, and this cannot be explained by the participant not knowing the family. Participants tend to underestimate shocks faced by others and overestimate benefits received. Notably, participants do have some information on which families are receiving additional benefits due to the COVID-19 pandemic, in that they are significantly more likely than random guessing would suggest to correctly identify a family that is actually a self-reported benefit recipient. Yet, this information is still far from perfect, and the overall match accuracy rates for having faced a shock and for receiving any COVID-19 benefits are around 63% and 62%, respectively.[42]

[40] Various social programs disbursed additional (mostly cash) benefits as part of Indonesia's National Economic Recovery Plan. These programs included the Family Hope Program (PKH), Raskin rice subsidy program, and Pre-Employment Card Program (PKP).

[41] For context, if households had randomly chosen 48% of cases to report receiving COVID benefits, the match rate would have been around 51%. Hence, even taking into account this over-reporting, this targeting is better than random by about 11 percentage points.

[42] It is worth noting that, as in any survey, families may have tried to misrepresent themselves as poorer than they actually

5.4 Information Updating in Response to Changes in Welfare over Time

If community members hold dynamic welfare information, we would expect their welfare rankings of families to change over time, in response to changes in families' welfare. We explore whether this is the case by comparing the participants' benchmark-specific welfare rankings reported in the main (baseline survey), and those reported in the follow-up survey around three months later. In the three communities where we administered the follow-up survey, each participant was asked to re-rank the same 10 families that they had ranked in the first survey round, for clear comparability.

First, we note that welfare rankings, both as reported by participants and as calculated by our survey, do change over time. There are many reasons that individuals may have experienced idiosyncratic changes in welfare over this three-month period: many families participated in large religious celebrations, harvested their crops, and experienced the effects of the worsening COVID-19 pandemic (including deaths in two sampled families). We see this confirmed by the survey-collected welfare benchmark information; though the average change in per capita consumption (accounting for household roster changes), MUE, and the asset index were all centered around zero, individual changes in these measures range from -1.3 to 2.1, -2.1 to 5.0, and -2.2 to 1.8 standard deviations of the baseline sample metric values, respectively (see Table 1 for the standard deviation values and Figure 5 for the distribution of changes in each measure).[43] Table A12 shows the rank correlations between rounds for both participant-assessed and survey-assessed welfare rankings. In both sets of welfare rankings, we see that the asset rankings are significantly more stable between rounds (with between-round correlations around the order of 0.7) than the expenditures and neediness rankings are (with between-round correlations around the order of 0.4), which lends support to our claim that the latter measures capture more dynamic changes in welfare. Yet, encouragingly, we also see some persistence of these more dynamic measures between rounds, which suggests we are indeed capturing some welfare information and not only noise. Comparing participant- and survey-generated rankings, we note that survey measures of assets are significantly more correlated between rounds than rankings provided by the participants. But the qualitative patterns are generally similar.

Given that participant-assessed rankings do change between rounds, we want to understand whether these changes capture ranked families' "true" changes in welfare. This is critical, because these changes between rounds could simply be due to noise, or changes in experimental task effort between rounds (given that the

are in order to qualify for more social benefits. We were very explicit from the beginning of the survey visit that these activities were purely for research purposes, we were not affiliated with the government in any way, and their answers would not be used to directly inform any policy decisions. But of course such a motive can never be fully ruled out.

[43] Note that the MUE and asset index values for the main survey were estimated before data from the follow-up survey was collected, and hence incorporate none of the follow-up survey data. However, especially given the small sample size for the follow-up survey, baseline data was used to estimate the MUE and asset index in the follow-up survey. The correlations between the baseline MUE and asset index measured in this estimation were very highly correlated with those in the first estimation (Pearson correlations of greater than 0.99), hence this choice does not affect any results in any important way.

incentives changed from payments to chances to win a prize). As a first-pass validity check, we again calculate the rank correlations between participant- and survey-assessed benchmark-specific welfare rankings in the the follow-up survey (as well as re-calculating the baseline correlations for the three communities re-surveyed only); this can be seen in Table A13. Notably, the correlations between participant- and survey-assessed neediness and assets are strikingly similar to those found in the baseline survey (see Table 2), and also generally similar to the baseline correlations for only those three communities in question. The per capita expenditures correlation, on the other hand, is lower than before, and even becomes negative. This suggests either that individuals truly do not have a good sense of others' expenditures or that expenditures are not well measured by our survey.

One way to understand whether changes in participant-assessed rankings reflect true welfare changes is to see whether survey-measured welfare and participant-measured welfare change in tandem between surveys. To see whether changes in participant- and survey-assessed welfare benchmarks are correlated, we regress changes between rounds in the ranking order positions that participants assigned to a given family on the analogous rank position changes of those families as defined by our survey measures. Results can be seen in Columns 1-3 of Table 6. A positive coefficient would indicate that changes in participant-assessed welfare ranking positions between rounds move simultaneously with changes in survey-assessed ranking positions of the specified welfare metric. As we can see, none of the coefficients are positive and significant, suggesting that survey-assessed and participant-assessed welfare rankings are not capturing similar changes between rounds. We may worry that some ranking changes will be small, and that participants may only be able to observe large changes in welfare. To address this concern, in Columns 4-6 of Table 6, we regress changes in benchmark-specific participant rankings between rounds on the change in the *value* of the survey-assessed metrics. In this specification, we *do* see a positive and significant correlation between changes in survey-assessed and participant-assessed neediness. Hence, it is possible that participants can observe large changes in neediness between surveys, and update their rankings accordingly.

Given the lack of evidence that changes in participant rankings reflect true changes in welfare, we wondered what additional information participants may be incorporating. Notably, the community meeting exercise (the results of which are discussed in Section 7) happened shortly after baseline surveys were administered in all communities, and anecdotally enumerators suggested that participants had been referencing what they learned in the meeting to inform their rankings in the follow-up survey. This meeting, where the welfare status of others was discussed openly with community members, certainly may have informed participants' welfare information about others. At the same time, this information would have been "old" by the time of the follow-up survey, and not necessarily reflective of current welfare status. Yet, if it is the case that participants really do receive and hold very little dynamic welfare information (as suggested by previ-

ous results in this section), the information shared at this meeting could have been the most recent welfare information about others that participants had learned. Hence, while participants tried to predict dynamic welfare with assets and long-term welfare correlates during the first round, they might have been rationally updating to incorporate the most recent welfare information they had, which was what they learned at the meeting.

To explore this hypothesis, we regressed participants' second-round rankings on their first-round rankings, changes in survey-assessed welfare, and community meeting rankings. The results can be seen in Table A14. We see here that, even controlling for individual's baseline rankings and survey-assessed welfare changes between rounds, the ranking at the community meeting has significant predictive power in determining participants' rankings based on neediness and asset benchmarks in the follow-up survey. Indeed, the predictive power of the relative rank position given to a family by the community is very similar to that of the predictive power of the participant's first-round benchmark-specific ranking. In fact, in all of the columns considering the neediness benchmark (2,5) and asset benchmark (3,6), the coefficient on the benchmark-specific ranking in the baseline round is not statistically different from the coefficient on the community ranking position.

We also wonder whether survey-measured welfare changes between the rounds may mostly be capturing noise, rather than actual welfare changes. To address this concern, in columns 4-6 of Table A14, instead of including the change in welfare between rounds, we instead control for survey-assessed benchmark-specific welfare measures in the follow-up survey. Notably, this does not substantially affect the results. Column 5 suggests that the neediness rankings captured by the survey may help explain participant-assessed neediness, providing further evidence that community members may observe some changes in neediness over time.

Taking this evidence together, participants generally do not seem to receive or hold very current welfare information about other families (except perhaps for large changes in neediness); the most recent salient signal about welfare information that participants may have systematically incorporated was information shared in a community meeting around three months earlier. This evidence is only suggestive.[44]

5.5 Comparison to Simple Proxy-Based Methods

Simply showing that the community-held dynamic information seems sparse does not necessarily mean the community does not capture more dynamic information than standard proxy-based methods. To explore this, we compare the correlations between both community-reported metrics and very simple proxy means test scores with survey-based metrics. As noted in Section 2, one of the standard arguments for why

[44]Of course, we cannot rule out an explanation where individuals think that the experimenters are basing the accuracy of their responses in the round 2 exercises on what was shared in the community meeting. However, enumerators explicitly told the participants as part of their script that we were using the information collected in the current survey round in order to calculate their accuracy.

community-based targeting rankings do not outperform proxy means scores in predicting survey-measured consumption is that the community might be targeting based on some non-consumption welfare benchmark. Here we can make a direct comparison, since we directly asked the community for consumption information (as well as dynamic neediness information).

The proxy means tests formulas that we consider for this comparison are the Poverty Probability Index (PPI) for Indonesia [IPA, 2020] and the Simple Poverty Scorecard (henceforth referred to as SPS) for Central Java [Scorocs, 2019]. Both of these tools use data from Indonesia's National Socioeconomic Survey (SUSENAS) (the 2016 round for PPI and the 2018 round for SPS), to identify 10 easily observable proxies that predict per capita consumption well. They then create a simple formula that maps these 10 variables to a predicted consumption "score," which should allow others in similar contexts to estimate relative consumption between households, collecting only information on these 10 variables. Using these formulas, we are able to roughly calculate both PPI scores and SPS scores for each family in our sample.[45] Notably, the proxies used in the two formulas overlap significantly but not completely, and are mostly fairly static demographic and asset ownership characteristics. Some summary statistics of the calculated scores for our baseline sample can be found in Table A15. Despite the fact that these scores are calculated with less than 10 variables, there is indeed variation in calculated scores for both formulas.

Once we calculate these proxy means scores for our sample, we can then calculate their rank correlation with our various survey-based benchmarks for each set of participant-ranked families (for direct comparison to participants' responses). The average rank correlations (and comparable statistics for the community) are displayed in Table 7. In each row of the table, we see how well the rankings based on both PMT scores and participant-assessed rankings (in the targeting task and each information elicitation benchmark) correlate with each survey-based metric. Notably, we see that none of the community-based rankings outperforms either proxy means test score in predicting any of the three survey-based metrics.[46] This is regardless of the welfare benchmark that communities are asked to rank; it is not the case that participants' "dynamic" welfare rankings of consumption and neediness contain significantly more dynamic predictive information beyond what is captured by a proxy means test score. This is important because it shows that differences

[45]The PPI formula uses province of residence, total household members, household members ages zero to five, household members ages six to ten, ownership of a refrigerator, ownership of a motorcycle, floor material, toilet type, cooking fuel source, and recent internet use. As a practical matter for our sample, province is the same for all sampled families and recent internet use is omitted because we did not collect this information during the main baseline survey. The SPS formula uses regency of residence, total household members, total currently working household members, total members with "permanent" jobs, receipt of the *Rastra* program rice subsidy, ownership of a refrigerator, ownership of a motorcycle, toilet type, cooking fuel source, and whether the female head of household has a cell phone. Again, there is no variation in the "regency" variable in our sample. We proxy the female head's ownership of a cell phone with whether the household owns more than one cell phone, since we did not collect information on whether the female head owns a phone. While we are not able to calculate these scores perfectly because of this imperfect match in data, if anything this should bias these measures against predicting consumption better than the community.

[46]Table A16 also shows that the community does not have a predictive advantage in identifying the "poorest" individuals as indicated by each survey benchmark.

in community definitions of poverty cannot explain away differences in the community's and proxy means scores' abilities to predict consumption outcomes. Notably, in terms of predicting neediness, there are some community ranking benchmarks for which the PMT does not significantly outperform the community; the community's targeting and asset rank correlations are statistically similar to those for both proxy means scores. However, it is important to remember that PMT scores are designed to predict consumption rather than the MUE measure. It seems likely that, if we were able to recalibrate the PMT formula to predict MUE, the community could lose its predictive advantage. In predicting assets, on the other hand, the proxy means test scores have a mechanical advantage over the community, in that some of the variables in the asset index (or very closely related variables) are also used to predict the proxy means test score (i.e., refrigerator ownership, motorcycle ownership, cell phone ownership, floor quality, toilet type). Given this, we should consider comparisons regarding assets with a grain of salt, and note that community information may still be a reasonable and low-cost way to predict more static welfare characteristics.

6 Results: Information Used by Individuals in Targeting

6.1 Relationships between Participant-assessed Welfare Information and Targeting Task Behavior

Now we can compare community-assessed information to the targeting task outcomes to gain some insight on the type of welfare information being used by participants to target transfers. First, we calculate the rank correlations between the participant-assessed welfare rankings and their targeting task rankings. These results can be seen in Table 8. We note that the average Spearman coefficients associated with all of the benchmarks are positive and significantly different from zero, suggesting that participants are indeed considering these types of welfare information (or correlates of these types of information) when providing reports in the targeting exercise. Indeed, Figure 6 shows that, for all benchmarks, there is a significant mass of participants who provide identical targeting and benchmark-specific welfare rankings. The results suggest that participants' knowledge of others' assets (and correlated variables related to longer-term wealth) explain most of the variation in participants' rankings in the targeting task. The average correlation is significantly higher than the correlation between the targeting task rankings and both the neediness benchmark and consumption benchmark rankings.

We may wonder whether the reliance on asset information (which tends to be more visible than consumption information) to determine welfare status is a technique participants are more likely to employ if they do not know a family well or interact with them infrequently. Looking at the bottom rows of Table 8

however, we see that the correlations between these types of information for known families are extremely similar to those for the full sample. Hence, even with well-known families, asset information seems to be quite important.

Finally, as we noted in the previous section, participants' information on short-term and long-term welfare metrics are highly correlated. Hence, we want to understand whether information from the more dynamic rankings can explain additional variation in targeting task rankings, after controlling for the influence of asset rankings. To explore this, we perform some simple regression analysis. The results can be seen in Table 9. Similarly to our analysis in Alatas et al. [2012] and Dervisevic et al. [2020], we simply regress the targeting ranking order positions on other welfare information; in this case, we regress these positions on the ranking positions made by the same participant about the same family in the welfare benchmark-specific ranking tasks. In column one, we see that all three benchmark-specific rankings seem to provide some explanatory power in predicting the individual targeting task ranking. However, the coefficient on the asset-based ranking is both economically and statistically significantly (via F-test) much larger than the coefficients on the two dynamic welfare benchmark rankings. Because the various benchmark-specific rankings are highly correlated with one another, to more clearly identify the additional information contained in the dynamic welfare measure rankings, we perform another specification in column 2. Here we first regress the individual targeting ranking on the asset-specific ranking. Then we take the residuals from this regression and see whether this additional variation can be explained by the dynamic welfare measure rankings. This specification suggests that per capita food expenditure rankings provide no additional information to explain participants' targeting rankings. While the neediness rank still provides some additional information, the coefficient is much smaller than in our first regression. Additionally, looking at the R^2, we see that these dynamic welfare rankings explain less than 4% of the part of the targeting decision that is not already explained by the asset ranking. So, while it seems some more dynamic information is being incorporated into targeting, it is relatively much less important than longer-term welfare information.

In columns 3 and 4 of Table 9, we perform exercises similar to the first two columns, using a different type of dynamic welfare information: whether a family faced a shock.[47] We may also think that, while a participant cannot directly observe the consumption or neediness of a family, they may be more easily able to observe a shock (like the death of a family member), which might affect these less observable dynamic welfare measures. We did not ask participants whether each of the 10 families that they ranked experienced a shock; rather, we asked about three random families that may or may not have been the same ones ranked. There are a total of 253 cases where a participant ranked a family and also was asked if that family had

[47] We could also do this with the benefit measure, but it's unclear whether we would expect receipt of benefits to improve or reduce a family's welfare ranking. Because benefits are generally targeted at poorer households, but then improve their welfare, these results would be more difficult to interpret.

experienced a shock. In 147 of those cases, the participant provided a yes or no answer to whether a shock was experienced. Because families in both processes were chosen independently, we do not think the sample of 253 cases would induce any selection bias. However, there may be some bias as to which families were known as having suffered a shock, and that should be kept in mind when interpreting these columns. We note here that this shock information does have a strong predictive impact on a participant's ranking decision, even after controlling for asset rank. A family that experienced a recent shock was ranked as approximately 0.7 positions poorer. If we interpret the willingness of participants to report whether or not a family faced a shock as an indicator of whether they feel confident that they have certain information, then these results suggest that participants *do* incorporate more dynamic welfare information when they have it; it is simply the case that they generally do not have said information.

Considering all of this analysis, it seems to be the case that participants' information about others' assets (and correlates of asset ownership relating to longer-term wealth) are most often employed in the targeting process. However, there does seem to be some use of information about neediness or shocks that is not related to assets. Hence, there may be some dynamic welfare information incorporated in individuals' targeting reports when participants have it.

6.2 Participants' Explanations of their Targeting Decisions

Qualitative justifications that participants provide about their targeting decisions support the assertion that they mostly consider asset-based welfare information in creating targeting rankings. After participants completed the individual targeting task, they were asked to provide open-ended verbal explanations of their relative ranking positions of three randomly chosen pairs of families. Importantly, this was before we asked participants to provide information about any specific welfare benchmark, so that their later rankings would not inform answers to this question.[48] Some very basic text analysis of the responses to these answers can be found in Tables A17 and A18. Table A17 shows the most commonly used words in the sample of explanations, after eliminating common stop words that provide little insight.[49] [50] Words describing land are mentioned significantly more than any other words. "Rice fields" ("sawah" in Bahasa Indonesia) alone is mentioned 430 times, which is over 3 times more than any other word is mentioned. The second and fourth most common words also have to do with land, and together are mentioned about half as many times as rice fields. Hence it seems that comparison of land ownership is a driving factor behind many participants' answers; the word "rice fields" is mentioned by 68% of participants in at least one of their

[48] Participants also were informed that this would be the only round in which such questions would be asked, and that in other rounds they would not have to explain their process.

[49] Examples of stop words in English are "is," "and," "about."

[50] Two participants accidentally did not complete this task, so there are 298x3=894 unique observations.

explanations. Besides land, other agricultural characteristics, such as being a farmer or a farm laborer, are mentioned frequently. Notably, these employment distinctions are also closely tied to land ownership status; farmers farm their own land, while laborers work on the land of others. Other than this, we note that houses (another asset) and terms describing family structure also appear in many explanations. We consider family structure a relatively stable, longer-term welfare characteristic, given that families do not seem to frequently gain and lose members in this context. Words such as child and widow are among the top six most commonly mentioned words. However, we also see some other terms generally associated with income and jobs on this list, suggesting that there may be some consideration of dynamic income flows. Importantly, we do not see mentions of families consuming more or skipping meals, being unable to meet their daily meal requirements, having recently experienced a welfare shock, or any other types of dynamic welfare measures that traditionally have been hypothesized to be considered by the community.

Additionally, we may be concerned that, by looking at single words, we may miss the nuance of some of the responses, and may be incorrectly interpreting the presence of certain words. Hence, we also look at the most common three-word phrases in Table A18. These phrases confirm the importance of ownership of land, as well as the ability to work on one's own land. Elements of family structure emerge here as well, such as being a widow or having dependents. In general, most participants are explaining their welfare ranking decisions by referencing land ownership and family structure, both of which are associated with longer-term wealth. These explanations provide confidence that our interpretation is valid: participants rely on more visible, less dynamic welfare information when providing information for targeting.

6.3 The Limited Role of Other Preferences in Targeting

In previous sections, we discussed how various other preferences (unrelated to families' welfare status) may enter into the targeting process, which also may distort identification of a clear relationship between the information community members have and their ultimate targeting decisions. We argue here that we do not see evidence of other preferences entering into the ranking process, allowing for a clearer look at the relationship between information and targeting behavior. We attempt to catch the effect of any individual preferences on the ranking process with two additional ranking tasks. First, the preference elicitation task asked participants to rank families simply based on whom they would prefer to receive extra money, with no explicit welfare context. The idea is that comparing this ranking to the targeting ranking might give us useful information on the influence of preferences in the targeting decision. We can see in Table 10 that the average Spearman correlation between the targeting and preference ranks is 0.92. This correlation is higher than that between the individual targeting task ranking and any one of the information benchmark ratings. Yet,

this high correlation on its own is hard to interpret; it could support both an explanation where targeting rankings are completely dependent on non-welfare related preferences and an explanation where individuals' preferences are indeed to give more money to the poorest households, with no confounding preferences. In our experiment, we believe the latter explanation is most likely the case, because both the targeting task and the preference elicitation task on average have statistically similar, relatively high correlations with the participant-provided welfare rankings based on the various benchmarks. We do not take a stand on whether this "preference" to award money to the poorest households is induced by desirability bias from the experimental setting, as this does not matter for the purpose of interpretation of the other results.

Second, we ask participants to rank the families in a way that is most similar to how they think other participants would do so in the individual ranking task, in the second-order beliefs elicitation task. The idea is that comparing this ranking (which is incentivized based on accuracy) to the participant's own individual targeting task will reveal whether the participant incorporates idiosyncratic preferences that are not shared by other community members. Yet, again in Table 10, we see that the individual targeting task rankings and beliefs about how others would complete the ranking task are highly correlated at an average of 0.94. Because these second-order belief rankings also are highly correlated with participants' welfare benchmark-specific rankings, we conclude that most participants are not using idiosyncratic preferences in forming their rankings, and that they believe that other participants will also use welfare information to make their ranking decisions. Hence, we again find little evidence that other preferences are influencing targeting in the context of our experiment.

In addition to these correlations, we do some more standard "nepotism"-based tests to learn whether participants may be allocating transfers to individuals they value more (or think are likely to share the money with them), such as close family and friends.[51] To do this, we regress a participant's ranking of a given family on an indicator of whether the ranked family and the participant's family are close friends or family members. (This identification of the type of relationship between the participant and each family was done before any mention of the ranking tasks, and hence incentives for strategic misreporting of relationships are unlikely.) A negative coefficient on the indicator variable would suggest that family/close friends are ranked systematically poorer than other participants, and hence more likely to get additional money. The results of these regressions are in Table A19. In column 1, we can see that the coefficient on the family/close friend indicator is positive and insignificant; if anything, these family and friends are being ranked as less poor than others. However, individuals who are family/friends with another participant could be systematically richer, and there could still be some bias in ranking them lower. However, in columns 2 and 3, we see that, when

[51] We do not do an "elite capture" test looking at the role of government officials, because a very small fraction of the sample actually are government officials of some type (3%), but most are connected to an official in some capacity (98%).

controlling for either participant-assessed welfare benchmark-specific rankings (of expenditures, neediness, and assets) or rankings established by our survey, the coefficients stay positive and insignificant.[52] We also may be worried that the indicator of being a family member/close friend may not capture all of a participant's important connections in a community, so we repeat the exercises in the first three columns, instead using an indicator of whether a family is well known by the ranking participant (either close family/friend, close work colleague, or someone the participant speaks with frequently). These results can be seen in columns 4 through 6 of Table A19. We see the same qualitative pattern here as in the first three columns; when we control for survey-assessed welfare ranks, participants ranked individuals that they know as systematically richer than others. In fact, holding the survey-assessed welfare benchmark-specific rankings constant, a family that is known well is ranked about 0.25 positions richer in the targeting task (and hence less likely to receive a transfer). Taking this evidence together, we find it improbable that non-welfare related preferences are significantly affecting our results. This lends additional confidence that we can trace out a clear relationship between individuals' information sets and the information they use for making targeting decisions.

7 Results: Information Aggregation at the Community Level

7.1 Community Meeting Data

Given that many CBT exercises involve joint decision-making by many participants, we now consider how targeting outcomes may change individuals' information and how preferences are combined at the community level. We do this using two sources of data. The first source is the experimental data we have been describing thus far; we compare participants' individual targeting task rankings to the rankings produced by their community during the CBT exercise meeting. Because we only carried out 10 community meetings (which may produce noisy results), conducted these meetings in a small, homogeneous geographical area, and think framing from the previous experimental exercises may have informed participants' conceptions of welfare, we may worry about the external validity of these findings. To alleviate such concerns, we consider a second data source: the data collected for Alatas et al. [2012], as well as various follow-up papers by this group of authors. Alatas et al. [2012] performed the following experiment. Using the same administrative community division that we use (the RT), 640 communities throughout Northern Sumatra, South Sulawesi, and Central Java were randomly assigned to use different targeting methods to distribute a set of one-time Rp. 30,000 (around US$ 6.38) transfer payments. The targeting methods were PMT targeting, CBT targeting, or a hybrid of the two methods. Relevant for our purposes are 214 of these communities, in which experimenters

[52] The fact that this relationship holds in column 3 implies that this effect cannot be explained by family members also being ranked as poorer than they actually are in the welfare benchmark-specific rankings, perhaps in participants' attempts to justify their individual targeting task rankings.

organized a community meeting exercise similar to ours.[53] Additionally, nine surveyed households in each of those communities were asked to complete a version of an individual targeting task, ranking the eight other surveyed households from poorest to richest. Hence we can compute a similar analysis comparing individually-assessed and community-assessed rankings in this data set. Given that this data source has a larger sample (over 1,900 households) in more communities, covers a larger geographical scope, and was not framed by experimenter-imposed notions of welfare, finding similar results in our experimental data can support the external validity to our results.

The Alatas et al. [2012] sample is generally similar to ours, but differs in a few key ways. The mean household size is slightly larger in their sample (4.3 compared to 3.3 in our sample) and household heads are slightly younger (48 compared to 56 in our sample). Our participants are also more likely to live in agricultural households (about 48% of households in their sample versus 57% in ours), which makes sense as their survey did not exclusively cover rural agricultural areas. Household head gender, education levels, and marital status are very similar between samples. Per capita expenditures are nominally slightly higher on average in our sample than in theirs. However, this is sensible, given the over ten-year gap between the two data collection exercises.

There are a few differences in procedures worth noting between the Alatas et al. [2012] experiment and ours. First, the sampling unit in their work is households rather than families. While the individual targeting task survey was similar to ours – participants were asked to rank other randomly chosen households from poorest to richest, without further directions – the households to be ranked were not selected to potentially overrepresent households that were well known by the survey respondent. Perhaps as a consequence, participants were not as strongly encouraged to rank all eight households. Rather, they could claim that they "did not know" for any of the eight households and could choose not to rank them. Hence, about 46% of the households in the communities we consider chose to rank fewer than 8 households, although the median number of households ranked is still 8 (mean=7.2). Additionally, unlike our experiment, there were no clear consequences associated with the exercise's outcomes (it was simply a question in a larger survey), and hence it is less clear what objective function respondents may be maximizing when providing their ranking list. In terms of the community meetings, in the Alatas et al. [2012] experiment, all community members were invited to the community meeting, and all community members were ranked in the CBT exercise, regardless of whether they were surveyed. Hence, there are many people who participated in these community meetings whose individual rankings we do not know. There also may have been larger effects of exercise fatigue on final CBT ranking positions in their context, because, in many communities, more than 30 households were

[53]There were other communities that had meetings in Alatas et al. [2012], but only local leaders were invited, which does not map as clearly to our work.

ranked. In fact, the average number of households ranked at these meetings was 53.6 (median: 48), and more than 30 households were ranked in about 86% of these communities. Finally, the framing exercise performed in our community meetings was slightly different than theirs; we asked meeting attendees to list characteristics associated with being poor and characteristics associated with being rich, while they asked attendees to list specific characteristics that can help differentiate households' welfare, and then to select which of those are most important to the community.[54]

7.2 Concordance of Preferences

Working with the data in Alatas et al. [2012] as well as from our experiment, we first consider how concordant individuals' ranking decisions are with the ranking ultimately crafted by the community. To do so, we take the set of families or households ranked by each individual and construct the relative rankings of those families/households made at the community level, disregarding any households/families that were ranked by the community but not by the individual participant. We then compute Spearman rank correlations between the individual targeting task ranks in our experiment (or individual rankings in the Alatas et al. experiment) and the community meeting-produced relative ranking. Average correlations are displayed in Table 11. Correlations are quite high, with an average correlation coefficient of 0.66 in our experimental sample. Strikingly, the average correlation coefficient in the Alatas et al. data set is extremely similar, at 0.67, and the averages of the two samples are not statistically different. Moreover, looking at the full distribution of Spearman correlations in both samples in Figure 7, we see that the majority of individuals' rankings correlate with the community rankings with a coefficient of 0.5 or higher. Both distributions have a long left tail; the Alatas et al. data distribution is slightly smoother given the larger sample. This suggests that community members do share broadly similar notions of who is poorest, and, in the case of our experiment, who should receive extra money. There does not seem to be any radical shift in the targeting information/objectives used when going from the individual to the community level.

However, where we do see a notable difference between the two samples is when we consider the subsets of i) those participants who had at least one family/household member actually attend the community meeting and ii) those cases in which the survey respondent himself or herself attended the community meeting. Notably, in our experimental sample, the mean correlations for these subsets are statistically and economically similar to the full sample. However, in the Alatas et al. sample, both subsets have statistically larger Spearman correlations between individual and community ranking orders, at 0.706 and 0.708 for subsets i and ii, respectively. Notably, because in our experiment we encouraged attendance, both by providing incentives and by visiting households that had not yet sent a member to the meeting, we ended up

[54] For more details about the sample and experimental procedures, see Alatas et al. [2012].

getting a much higher overall attendance rate of survey respondents than in the Alatas et al. study. Ninety percent of families had a representative, and our strong encouragement of all to attend perhaps induced less selection among attendees. The much larger scale Alatas et al. experiment had a lower participation rate among surveyed households (about 66%), and there may have been more selection regarding which households were willing and able to attend the meeting. Hence the difference in correlation may reflect the stronger influence of these meeting participants' individual targeting opinions in the community targeting decision as compared to those that did not attend the meeting. Yet this possible selection effect is still relatively small: only about a 5% increase off of the mean correlation for the entire sample. In general, individuals provide fairly similar targeting rankings among themselves, and these ranking are quite similar to those created during a CBT meeting exercise.

7.3 Meeting Attendees' Stated Definitions of Welfare

In community meetings for both samples, there was a welfare definition exercise to frame the community's decision-making process. Looking at the community's ideas about the meaning of welfare can give us some additional insight into their decision-making processes. As discussed above, in the Alatas et al. study, the framing exercise consisted of asking meeting participants to list ways in which welfare status could be differentiated between households, and then asking the community to choose which of those factors were most important to the community's definition of welfare. The most commonly mentioned words, both in the initial brainstorm round and in the narrowed down list of important terms, are displayed in Table A20. Notably, we see that "house" comes up most frequently in both rounds. Houses are an economically important and very visible asset. Many of the other words that are mentioned commonly, like vehicles, land, livestock, wealth, assets, and furniture, also reflect longer-term wealth. Hence it seems that, similarly to individuals' explanations given during the individual targeting task in our experimental exercises, communities consider assets in differentiating relative welfare status between households. Land comes up less frequently in the Alatas et al. sample than in ours, but their sample is less rural than ours. Other more static characteristics, such as dependents/children, education, job, and business, also are mentioned frequently. Income, which could be considered a more dynamic measure, is also deemed quite important; however, measures of current welfare, like the ability to meet daily needs and expenditures, are mentioned much less frequently. Out of all of 214 meetings in the Alatas et al. study, "needs" are mentioned as important only six times, and expenditures only five. As a reference, a house is deemed one of the most important welfare characteristics in 143 cases.

The general pattern of responses given during the framing exercises at our community exercises are

similar to those in the Alatas et al. study. The most common words and trigrams can be seen in Tables A21 and A22 respectively.[55] As in the individual ranking task explanations, land was the most commonly used word in descriptions of both the rich and the poor. Interestingly, the more dynamic notion of "needs" also comes up relatively frequently, which is distinct from participant responses in the individual ranking task explanations, as well as from the Alatas et al. community meetings. This may be because participants had been recently asked to rank households on neediness during our household visit. However, we do not see a similar effect with expenditures or "assets" more generally being mentioned frequently, hence this is likely not solely a consequence of our experiment. It's possible that the notion of neediness did not come up more in the Alatas et al. experiments because participants were specifically asked to provide ways to differentiate welfare status between households. If neediness is hard to observe, than it may not actually be useful in differentiating welfare status. Perhaps in our experiments, where the prompt was a bit more general, the community did identify neediness as an important characteristic of welfare, regardless of the fact that they do not observe it particularly well.

7.4 Comparison of Individually-assessed and Community-assessed Welfare to Survey-assessed Welfare

While in both samples many participants' rankings broadly agree with those created by the community, most individuals' rankings aren't perfectly concordant; individuals do report some information that does not perfectly match the information reported by others. The goal of a community meeting exercise is to try to get more accurate information than a single individual would report, both by allowing people to share knowledge and by discouraging individual non-welfare related preferences from impacting the results. Hence, we next ask whether the rankings provided at the community meeting are more correlated than the average individual ranking with any of our three survey-measured welfare benchmark rankings. Notably, to perform a similar exercise in the Alatas et al. data set, we need to make sure we can construct relatively comparable survey metrics. We are able to construct analogous per capita consumption and neediness measures using this data, but their consumption panel differs from ours (asking about many fewer categories), and hence slightly different consumption good categories were used to calculate the MUE in the Alatas et al. data set. Additionally, instead of calculating an asset index, we use the PMT scores constructed by the authors of that study, which use similar asset and long-term welfare measures to predict total consumption.

The results of this exercise for both data sets can be seen in Table 12. In the top panel (A) are the rank correlations between individual targeting task rankings and the survey-assessed benchmark rankings, and in

[55] We did not provide trigrams for the Alatas et al. sample, as the responses provided were generally less than three words each.

the bottom panel (B) are the rank correlations between the community meeting ranks and survey-assessed benchmark rankings (also at the level of the ranking participant by ranked family/household for comparability). For the sample in our experiment, while the average correlation with the survey-assessed rankings is marginally higher for the community rankings than the individually formed ones, this relationship is not significant for any of the benchmarks. Hence it seems that allowing the community to collaborate and share information does not lead to significantly better rankings, as assessed by their correlation with the survey-assessed welfare metric. The results for the Alatas et al. sample (on the right side of the same Table 12) are qualitatively similar, except we do see the community meeting rankings being closer to per capita expenditures than the individually assessed ones in this data set. The average rank correlation increases by about 0.04. Perhaps there is some gain to having a community meeting rather than simply aggregating individual targeting preferences if we want to target individuals with lower per capita consumption. But, reassuringly, the general magnitudes of the average correlations in this table are also very qualitatively similar between the two data sets, despite various differences in construction of the measures and implementation of targeting procedures. Key exceptions are that the MUE measure is significantly more correlated with both individual and community rankings in our experiment than in the Alatas et al. data set, and that the correlation between expenditure rankings and community meeting ranks is significantly higher in the Alatas et al. study than in our experiment. In general, we do not see significant gains from allowing individuals to discuss and aggregate their preferences, relative to the same welfare information captured with a survey.

8 Conclusion

While it may be tempting to romanticize the "tight-knit rural community" in developing countries as a setting where individuals can easily observe others' welfare status, our study suggests this is not always the case. We find that community members, much like a centralized policymaker, can observe fairly visible, static welfare proxies like assets owned and family structure. Community members are *not* particularly accurate in reporting dynamic welfare information on consumption, neediness, shocks, and receipt of benefits, at least if we consider the survey measures as the "truth." Community members do not seem to consider dynamic information much when performing targeting tasks, which runs contrary to claims in previous literature.

Importantly, the reasons that communities do not incorporate dynamic welfare in targeting do not seem to be either that they do not hold notions of dynamic welfare or that they do not update their beliefs based on new information received. After all, individuals' rankings of "neediness" are quite highly correlated with their individual targeting task rankings, and the notion of neediness was discussed at our community meetings as an attribute of being poor and rich. Additionally, when a respondent knows that a family has recently

faced a shock, there is some evidence this is incorporated in their targeting decision. Yet, when tasked with stating ways to distinguish between the welfare of different households (both individually expressed by our participants and during the community meeting poverty definition exercises in Alatas et al.), participants rely on easily observable attributes, which are more static welfare metrics. Moreover, there is some evidence that community members incorporated information they learned at the community meeting in the follow-up round of our experimental tasks. This suggests that new welfare information may be incorporated in community assessments; it's simply the case that new information is not often readily available.

This work also provides support for the use of the Ligon [2020] MUE metric as a perhaps more consistent way (both theoretically and practically) of capturing dynamic welfare. Despite using the same data as needed to calculate per capita expenditures, in both rounds of data collection, MUE information seems to line up much more closely than per capita expenditures with the notion of welfare held by the community. This is despite the fact that correlation between survey rounds of the two metrics is similar, and that per capita expenditures and the MUE are also highly correlated with each other.

In terms of policy implications, our study suggests the need for careful consideration of when CBT methods are appropriate to use in targeting social program benefits. More specifically, if policymakers aim to identify individuals who have recently faced a negative shock or are in immediate distress, CBT may not be appropriate, nor as useful as previous literature suggests. On the other hand, CBT does perform relatively well at identifying individuals' longer-term welfare status. While proxy-based methods do have a slight predictive advantage for all of our survey benchmarks, CBT is often much cheaper than collecting data on proxies from entire communities, and gives community members more agency in local decision-making processes. Moreover, differences between CBT and proxy-based targeting outcomes in our experiment mainly stem from differences in the underlying information held, which may not be an issue, given that survey measures are also imperfect. At least within our experiment, we do not find deviations in outcomes due to differences between policymakers' and community preferences over transfer recipients (i.e., we do not find elite capture or nepotism) nor do we find differences between their conceptions of poverty, two types of deviations which a policymaker may find undesirable.

An important caveat is that we need to further test the external validity of these findings. Notably, Indonesian culture places a high value on "reputation" and community risk-sharing is not common, which might imply less community information sharing than in other places throughout the developing world. A critical avenue for future research would be to replicate this exercise in places with different cultural norms and social structures. We also acknowledge that our sample size is somewhat small, and we hope to replicate these findings at a larger scale in future work. However, the ability to replicate some results with the Alatas et al. [2012] data provides some confidence that our findings may translate to Indonesia more broadly.

References

A. M. Adams, T. G. Evans, R. Mohammed, and J. Farnsworth. Socioeconomic stratification by wealth ranking: Is it valid? *World Development*, 25(7):1165–1172, 1997.

V. Alatas, A. Banerjee, R. Hanna, B. A. Olken, and J. Tobias. Targeting the poor: Evidence from a field experiment in Indonesia. *American Economic Review*, 102(4):1206–40, 2012.

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

V. Alatas, R. Purnamasari, M. Wai-Poi, A. Banerjee, B. A. Olken, and R. Hanna. Self-targeting: Evidence from a field experiment in Indonesia. *Journal of Political Economy*, 124(2):371–427, 2016b.

V. Alatas, A. Banerjee, R. Hanna, B. A. Olken, R. Purnamasari, and M. Wai-Poi. Does elite capture matter? Local elites and targeted welfare programs in Indonesia. 109:334–39, 2019.

H. Alderman. Do local officials know something we don't? Decentralization of targeted transfers in Albania. *Journal of Public Economics*, 83(3):375–404, 2002.

J. M. Alix-Garcia, K. R. Sims, and L. Costica. Better to be indirect? Testing the accuracy and cost-savings of indirect surveys. *World Development*, 142:105419, 2021.

A. Bah, S. Bazzi, S. Sumarto, and J. Tobias. Finding the poor vs. measuring their poverty: Exploring the drivers of targeting effectiveness in Indonesia. *The World Bank Economic Review*, 33(3):573–597, 2019.

A. Banerjee, E. Duflo, R. Chattopadhyay, and J. Shapiro. Targeting efficiency: How well can we identify the poorest of the poor? *Institute for Financial Management and Research Centre for Micro Finance Working Paper*, 21, 2009.

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

J. Banks, R. Blundell, and A. Lewbel. Quadratic Engel curves and consumer demand. *Review of Economics and Statistics*, 79(4):527–539, 1997.

P. K. Bardhan and D. Mookherjee. Capture and governance at local and national levels. *American Economic Review*, 90(2):135–139, 2000.

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

L. Beaman, N. Keleher, J. Magruder, and C. Trachtman. Urban networks and targeting: Evidence from liberia. In *AEA Papers and Proceedings*, volume 111, pages 572–76, 2021.

G. Bergeron, S. S. Morris, and J. M. M. Banegas. How reliable are group informant ratings? A test of food security ratings in Honduras. *World Development*, 26(10):1893–1902, 1998.

F. Bloch and M. Olckers. Friend-based ranking in practice. In *AEA Papers and Proceedings*, volume 111, pages 567–71, 2021.

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

CEIC. Indonesia household expenditure per capita, 2021.

R. Chambers. The origins and practice of participatory rural appraisal. *World Development*, 22(7):953–969, 1994.

S. Chen, M. Ravallion, and Y. Wang. *Di bao: a guaranteed minimum income in China's cities?*, volume 3805. World Bank Publications, 2006.

D. Coady, M. Grosh, and J. Hoddinott. *Targeting of transfers in developing countries: Review of lessons and experience*. The World Bank, 2004.

R. C. Crook. Decentralisation and poverty reduction in Africa: The politics of local–central relations. *Public Administration and Development: The International Journal of Management Research and Practice*, 23 (1):77–88, 2003.

A. Deaton. *The analysis of household surveys: A microeconometric approach to development policy*. The World Bank, 1997.

E. Dervisevic, S. Garz, A. Mannava, and E. Perova. In light of what they know: How do local leaders make targeting decisions? *The World Bank Policy Research Working Paper*, (9465), 2020.

P. Dupas, M. Fafchamps, and D. Houeix. Measuring relative poverty through peer rankings: Evidence from Côte d'Ivoire. *Unpublished Manuscript*, 2021.

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

E. Galasso and M. Ravallion. Decentralized targeting of an antipoverty program. *Journal of Public Economics*, 89(4):705–727, 2005.

M. Garcia, C. G. Moore, and C. M. Moore. *The cash dividend: The rise of cash transfer programs in Sub-Saharan Africa*. World Bank Publications, 2012.

M. Grosh and J. L. Baker. Proxy means tests for targeting social programs. *Living Standards Measurement Study working paper*, 118:1–49, 1995.

J. R. Hargreaves, L. A. Morison, J. S. Gear, M. B. Makhubele, J. D. Porter, J. Busza, C. Watts, J. C. Kim, and P. M. Pronyk. "Hearing the voices of the poor": Assigning poverty lines on the basis of local perceptions of poverty. A quantitative analysis of qualitative data from participatory wealth ranking in rural South Africa. *World Development*, 35(2):212–229, 2007.

J. V. Henderson, A. Storeygard, and D. N. Weil. Measuring economic growth from outer space. *American Economic Review*, 102(2):994–1028, 2012.

R. Hussam, N. Rigol, and B. Roth. Targeting high ability entrepreneurs using community information: Mechanism design in the field. *Unpublished Manuscript*, 2017.

Indonesian Central Statistical Body. Hasil Pendataan Potensi Desa (PODES) 2020, 2021.

IPA. PPI Scorecards and look-up tables, 2020.

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

D. Karlan and B. Thuysbaert. Targeting ultra-poor households in Honduras and Peru. *The World Bank Economic Review*, 33(1):63–94, 2019.

E. Ligon. Estimating household welfare from disaggregate expenditures. *Unpublished Manuscript*, 2020.

K. Macours. Comparing a direct with an indirect approach to collecting household level data: Who tells the truth about what? *Washington, DC, United States: Johns Hopkins University. Mimeographed document*, 2003.

OECD. Purchasing power parities (PPP), 2021.

J.-P. Platteau. Monitoring elite capture in community-driven development. *Development and Change*, 35 (2):223–246, 2004.

P. Premand and P. Schnitzer. Efficiency, legitimacy and impacts of targeting methods: Evidence from an experiment in Niger. *World Bank Policy Research Working Paper*, (8412), 2018.

E. Schüring. Preferences for community-based targeting-Field experimental evidence from Zambia. *World Development*, 54:360–373, 2014.

Scorocs. Scorocs-brand Simple Poverty Scorecard Poverty-assessment Tool, 2019.

Q. Stoeffler, B. Mills, and C. Del Ninno. Reaching the poor: Cash transfer program targeting in Cameroon. *World Development*, 83:244–263, 2016.

J. Strauss, F. Witoelar, and B. Sikoki. *Household Survey Questionnaire for the Indonesia Family Life Survey, Wave 5.* RAND, 2016.

Y. Takasaki, B. L. Barham, and O. T. Coomes. Rapid rural appraisal in humid tropical forests: an asset possession-based approach and validation methods for wealth assessment among forest peasant households. *World Development*, 28(11):1961–1977, 2000.

A. E. Temu and J. M. Due. Participatory appraisal approaches versus sample survey data collection: A case of smallholder farmers well-being ranking in Njombe district, Tanzania. *Journal of African Economies*, 9 (1):44–62, 2000.

A. Ubaidillah, K. A. Notodiputro, A. Kurnia, and I. W. Mangku. Multivariate Fay-Herriot models for small area estimation with application to household consumption per capita expenditure in Indonesia. *Journal of Applied Statistics*, 2019.

9 Tables

Table 1: Baseline Sample Summary Statistics

Statistic	Mean	St. Dev.	Median	Min	Max	N
Household Size	3.30	1.44	3.00	1.00	7.00	300
Household Head: Male	0.849	0.358	1.000	0.000	1.000	299
Household Head Age	56.6	13.6	57.0	22.0	91.0	299
Household Head: Disabled	0.017	0.128	0.00	0.00	1.00	299
Household Head: Any Education	0.863	0.345	1.00	0.00	1.00	299
Household Head: Sr. High School	0.281	0.450	0.00	0.00	1.00	299
Household Head: Married	0.783	0.413	1.00	0.00	1.00	299
Household Head: Employed	0.833	0.374	1.00	0.00	1.00	299
Household Head: Employed (Ag.)	0.572	0.496	1.00	0.00	1.00	299
Household Head: Born in Village	0.847	0.361	1.00	0.00	1.00	300
Household Head: Years in Community	44.1	22.4	47.0	1.00	90.0	300
Household Member: Any Education	0.917	0.277	1.00	0.00	1.00	300
Household Member: Sr. High School	0.527	0.500	1.00	0.00	1.00	300
Household Member: Disabled	0.047	0.211	0.00	0.00	1.00	300
Javanese (Ethnicity)	0.990	0.100	1.00	0.00	1.00	300
Muslim	0.980	0.140	1.00	0.00	1.00	300
Speak Javanese	0.973	0.161	1.00	0.00	1.00	300
Local Official	0.030	0.171	0.00	0.00	1.00	300
Know Local Official	0.977	0.151	1.00	0.00	1.00	300
Participate in Community Org.	0.967	0.180	1.00	0.00	1.00	300
Weekly PC Food Expenditures ($PPP)	22.98	34.26	15.57	0.85	370.90	300
Weekly PC Expenditures ($PPP)	37.76	40.83	26.74	2.49	393.75	300
Asset Index Score	−0.00	2.78	−0.325	−560	8.72	266
Land Area Owned (hectares)	0.136	0.190	0.080	0.00	1.19	294
Land Value ($)	26732	34761	21277	0.00	236170	293
MUE	0.085	1.00	0.133	−3.82	6.07	300
Ladder Step	2.74	1.04	3.00	1.00	6.00	298
Had Shock	0.300	0.459	0.00	0.00	1.00	300
Receives Govt. Benefits	0.917	0.277	1.00	0.00	1.00	300
Receives COVID Benefits	0.353	0.479	0.00	0.00	1.00	300

Notes: Statistics are generally at the family level, for each of the 300 families surveyed at baseline. Variables specified as "Household Head" are referring to the characteristics of the head of the household, whereas those marked "Household Member" specify whether any household member meets the listed condition. The education variables capture the percentage that have achieved a given education level (and may have additional education as well). "Employed (Ag.)" applies when the individual reports primarily working in the agricultural sector. "Disabled" encompasses both physical and mental disabilities. Local Official indicates whether a household member is an official for the village(*desa*) or any sub-village division, whereas "Know Official" indicates whether the participant knows one of these officials personally. Expenditures are converted from Indonesian Rupiah to US Dollars using a PPP exchange rate of Rp.4,700≈$1. "Receives COVID Benefits" refers to whether a family member is receiving any additional social support from the government associated with the COVID-19 pandemic.

Table 2: Correlation Between Survey-assessed and Participant-assessed Benchmark-specific Welfare Rankings

	Exp (1)	MUE (2)	Asset (3)	Land (4)
All	0.160	0.272^{1}	$0.445^{1,2}$	$0.398^{1,2}$
	(0.022)	(0.021)	(0.020)	(0.018)
N	300	298	297	297
Known	0.138	0.264^{1}	$0.412^{1,2}$	$0.400^{1,2}$
	(0.026)	(0.026)	(0.026)	(0.022)
N	294	292	289	287

Notes: The values presented are the average Spearman (ranking) correlations between 300 participants' rankings of 10 households, and the ranking suggested by our survey for each of the welfare metrics considered (described in detail in the text). Standard errors are in parentheses, and N denotes the number of observations used to construct the means. Standard errors were created by bootstrapping from the sample of estimated individual-level Spearman coefficients 5000 times. The first panel presents the rank correlations using all 10 ranked families, whereas the bottom panel presents the rank correlations only for families that were indicated as well known by the participant. (Participants with no known families with complete data account for drops in observation numbers.) The superscripts indicate that a given average correlation is statistically greater than the average correlation of the column with the corresponding number (within that row), using a bootstrapped (5000 times) two-tailed test of means ($\alpha = 0.05$). None of the top panel means are statistically different from the means in the same column in the bottom panel using a similar test. All averages are both statistically different than zero and one.

Table 3: Correlation Structure of Survey-assessed versus Participant-assessed Rankings

	Survey-Based Rankings				Participant-Based Rankings			
	PC Exp	MUE	Asset Ind.	Land	Exp	MUE	Asset Ind	Land
Exp	1	0.777^{*}	0.420	0.218	1	0.430	0.522^{*}	0.522^{*}
MUE		1	0.438	0.290		1	0.766^{*}	0.766^{*}
Asset Ind.			1	0.397			1	1
Land				1				1

Notes: This table compares the average correlations among the rankings of survey-based welfare metrics considered to the average correlations among the participants' benchmark specific rankings. Note that participants were not asked specifically about land, so we are using participants' asset-based rankings to compare to the asset index and land value rankings. (Due to this, no statistical testing is done on the bottom two rows.) A star indicates that a given average correlation is significantly greater than its survey-assessed or participant-assessed counterpart at $\alpha = 0.05$ (e.g., the star next to 0.777 denotes that it is significantly greater than 0.430), using a bootstrapped (5000 times) two-tailed test of means.

Table 4: Comparison of Correlations with Participant Rankings with Survey-assessed Benchmark versus with Survey-assessed Assets

	Survey Benchmark	Survey Asset Index
PC. Exp	0.160	0.375*
	(0.022)	(0.022)
Neediness	0.272	0.365*
	(0.021)	(0.023)

Notes: This table compares the average correlations between participant rankings of per capita food expenditures and neediness with their survey-assessed counterparts, and compares it to the average correlations between participant rankings of per capita food expenditures and neediness with our survey-assessed asset index. Standard errors are in parentheses. Standard errors were created by bootstrapping from the sample of estimated individual-level Spearman coefficients 5000 times. A star indicates that a given average correlation is significantly greater than the other in the same row at $\alpha = 0.05$, using a bootstrapped (5000 times) two-tailed test of means.

Table 5: Survey-reported versus Community-reported Shocks and COVID benefits

Variable	Survey	Participant	%Match (Survey==1)	%Match (All)
Any Shock	30.5	13.4	12.7	62.9
Family Head Death	0.0	0.0	NA	100
Other Family Member Death	0.40	0.20	0.0	99.4
Family Head Grave Illness	3.0	2.7	5.6	94.1
Other Family Member Grave Illness	0.80	1.9	0	97.0
Employment Loss	5.9	2.3	10.3	92.9
Fire/Earthquake	0.0	0.0	NA	100
Harvest Failure	3.8	0.0	0.0	95.9
Small Harvest	16.0	5.1	1.3	80.1
Other Shock	2.6	1.8	11.8	95.5
COVID Benefits	36.8	48.0	68.2	62.0

Notes: The top rows of this table present results for the 493 participant/other family pairs for which the participant responded as to whether or not the other family had experienced a negative shock over the past year. Each participant was asked about 3 families for a possible total of 900 observations. However, we did not end up interviewing a few of the families we enquired about in a few cases, and in 386 cases the participant would not provide information as to whether or not the other family had experienced a shock (bringing the total number of observations to 493). The "Survey" column describes the percentage of cases where the respondent of the survey interview denoted experiencing a given type of shock. The 'Participant' column describes the percentage of cases where the experimental participant indicated that a family that was asked about experienced a given type of shock. The "% Match (Survey==1)" column shows the percentage of cases participants reported that family facing a shock, out of cases in which survey respondents had also reported their family faced a shock. The "% Match (All)" column denotes the percentage of all cases in which the survey interview and participant's answer about whether a given family faced a shock matches. The bottom row presents similar information regarding others' receipt of COVID-related benefits. In this case, we similarly didn't have the relevant survey data for a few observations (due to failure to survey those families), and in 471 cases, participants refused to provide an assessment, bringing our total number of observations down to 429. All variables are defined similarly to above rows, except corresponding to whether a surveyed respondent or neighbor reported a family was receiving benefits directly associated with COVID-19.

Table 6: Changes in Participant-assessed Benchmark-Specific Rankings vs. Changes in Survey-assessed Benchmark Specific Rankings

	Dependent variable: Changes in Participant Ranking of:					
	Exp.	Need	Assets	Exp.	Need	Assets
	(1)	(2)	(3)	(4)	(5)	(6)
Δ Survey Exp. Rank	−0.080**					
	(0.033)					
Δ Survey Need Rank		0.016				
		(0.044)				
Δ Survey Asset Rank			−0.026			
			(0.056)			
Δ Survey Exp. Value				−0.000		
				(0.000)		
Δ Survey Need Value					0.250*	
					(0.103)	
Δ Survey Asset Value						0.038
						(0.052)
Constant	0.075***	−0.144***	0.053	0.081**	−0.167**	0.042
	(0.026)	(0.038)	(0.077)	(0.026)	(0.041)	(0.058)
Observations	829	810	611	829	810	611
R^2	0.005	0.0002	0.0003	0.003	0.008	0.001
Adjusted R^2	0.004	−0.001	−0.001	0.002	0.007	−0.001
F-statistic	4.395	0.2001	0.1901	2.433	6.702	0.4342

Notes: The above table presents regression analysis considering whether changes between surveys in participant-assessed welfare and survey-assessed move in tandem. Observations are at the ranking participant by ranked family level. Coefficients are presented with standard errors in parentheses below. Standard errors are cluster (by ranking participant) bootstrapped with 5000 bootstrap samples. One star indicates significance at $\alpha = 0.05$, and two stars indicate significance at $\alpha = 0.01$. The dependent variables are changes in participant-assessed welfare rankings between the baseline and follow-up surveys (subtracting the baseline rank position of a given family from the follow-up rank position of that same family) for our three welfare benchmarks: per capita expenditures, neediness/MUE, and assets. The independent variables in columns 1-3 are the changes in welfare rankings between the baseline and follow-up surveys of survey-assessed ranking positions, calculated analogously. In columns 4-6 we use the value of the change in the welfare benchmarks themselves, rather than the ranking position. Note that a positive coefficient indicates that changes in participants' ranking of a household between rounds moves in tandem with survey measured changes in the same welfare-benchmark domain. Numbers of observations less than 890 reflect cases in which data to construct one of the relevant variables is missing.

Table 7: Rank Correlations of Proxy Means Scores and Survey-based Metrics vs. Rank Correlations of Participant Reports and Survey-based Metrics

	Proxy Means Scores		Participant Ranks			
	PPI	SPS	Targeting	Exp.	Need	Assets
	(1)	(2)	(3)	(4)	(5)	(6)
Survey Exp.	0.364^{P}	0.380^{P}	0.293	0.160	0.253	0.277
	(0.017)	(0.018)	(0.020)	(0.022)	(0.021)	(0.019)
N	300	300	300	300	298	297
Survey MUE (Need)	0.304^{4}	$0.343^{1,4,5}$	0.315	0.186	0.272	0.306
	(0.016)	(0.017)	(0.019)	(0.022)	(0.021)	(0.019)
N	300	300	300	300	298	297
Survey Asset Index	$0.598^{2,P}$	0.531^{P}	0.462	0.375	0.365	0.445
	(0.014)	(0.013)	(0.020)	(0.021)	(0.023)	(0.020)
N	300	300	300	300	298	297

Notes: The values presented are the average Spearman (ranking) correlations between 89 follow-up survey proxy means test (PMT) score-based rankings of participants' 10 ranked households (Columns 1-2) and participants' rankings in the targeting/information elicitation tasks (Columns 3-6), and the ranking suggested by our survey for each of the welfare metrics considered (described in detail in the text). Standard errors are in parentheses, and number of observations is listed below. Standard errors were created by bootstrapping from the sample of estimated individual-level Spearman coefficients 5000 times. The numerical superscripts indicate that a given average correlation is statistically greater than the average correlation of the column with the corresponding number (within that row), using a bootstrapped (5000 times) two-tailed test of means ($\alpha = 0.05$). The "P" superscript indicates that a given average correlation is significantly greater than all participant rankings (columns 3-6) using the same tests, for visual simplicity. We only test here whether the proxy means test scores (columns 1-2) are significantly different from all other columns; we do not test for statistical differences among columns 3-6. All averages are both statistically different than zero and one.

Table 8: Correlation Between Individual Targeting Task Rankings and Participant-assessed Benchmark-specific Welfare Rankings

	Exp (1)	IMUE (2)	Asset (3)
All	0.486	0.764^{1}	$0.887^{1,2}$
	(0.029)	(0.023)	(0.016)
N	300	298	297
Known	0.465	0.771^{1}	$0.892^{1,2}$
	(0.032)	(0.024)	(0.017)
N	294	292	291

Notes: The values presented are the average Spearman (ranking) correlations between 300 participants' rankings of 10 households in the individual targeting task, and in the specified welfare metric-specific ranking task. Standard errors are in parentheses, and N denotes the number of observations used to construct the means. Standard errors were created by bootstrapping from the sample of estimated individual-level Spearman coefficients 5000 times. The first panel presents the rank correlations using all 10 ranked families, whereas the bottom panel presents the rank correlations only for families that were indicated as well known by the participant. (Participants with no known families with complete data account for drops in observation numbers.) The superscripts indicate that a given average correlation is statistically greater than the average correlation of the column with the corresponding number (within that row), using a bootstrapped (5000 times) two-tailed test of means. All averages are statistically different than both zero and one.

Table 9: Regression of Individual Targeting Task Rankings on Participant-assessed Welfare Information

	Dependent variable:			
	Ind. Targeting Rank	Residuals (Net Asset Rank)	Ind. Targeting Rank	Residuals (Net Asset Rank)
	(1)	(2)	(3)	(4)
PC Expenditure Rank	0.026*	−0.013		
	(0.013)	(0.015)		
Neediness Rank	0.202**	0.090**		
	(0.037)	(0.016)		
Asset Rank	0.719**		0.914**	
	(0.039)		(0.027)	
Negative Shock			−0.622**	−0.697**
			(0.217)	(0.156)
Constant	0.290**	−0.423**	0.675**	0.210*
	(0.086)	(0.087)	(0.220)	(0.088)
Observations	2,970	2,970	147	147
R^2	0.805	0.034	0.899	0.057
Adjusted R^2	0.805	0.034	0.898	0.051
F Statistic	4084	52.64	644.145	8.845

Notes: The above table presents regression analysis regarding how welfare-benchmark specific information may be used to make targeting decisions. Observations are at the ranking participant by ranked family pair level. Coefficients are presented with standard errors in parentheses below. Standard errors in the first two columns are cluster (by ranking participant) bootstrapped with 5000 bootstrap samples, and standard errors in the latter two columns are bootstrapped (due to lack of multiple observations by ranking participant). One star indicates significance at $\alpha = 0.05$, and two stars indicate significance at $\alpha = 0.01$. The dependent variable in columns 1 and 3 is the position assigned to a family by a participant in the individual targeting task. The dependent variable in columns 2 and 4 is the residual from regressing the individual targeting task ranking position on the asset-specific benchmark ranking position. The independent variables are participant-assessed benchmark-specific ranking positions and indicators of whether the participant reported that an individual experienced a shock in the past year. The shock variable is only available when the participant provided a definitive yes/no answer about whether a given family faced a shock, and also happened to rank this family in the later ranking tasks. Note that for all rankings, lower scores mean that a ranked family is worse off and higher scores mean that they are better off.

Table 10: Correlations Between Preference and Second Order Belief Rankings with Targeting/Information Benchmark Rankings

Category	Targeting	Exp	Need	Assets
Preferences	0.919	0.532	0.803	0.885
	(0.015)	(0.028)	(0.019)	(0.014)
N	300	300	298	297
Second Order Beliefs	0.938	0.514	0.797	0.918
	(0.013)	(0.028)	(0.021)	(0.012)
N	298	298	297	297

Notes: Participants were asked to rank 10 families based on who they would most to least prefer receive additional money ("Preferences") as well as to rank them most similarly to how they thought other households would do so in the individual targeting task ("Second Order Beliefs"). Average Spearman correlations between these rankings (Preference rankings in the top panel and second order belief rankings in the bottom panel) and the individual targeting task rankings and welfare benchmark-specific rankings (specified in the column header) are presented with standard errors in parentheses below. Standard errors were created by bootstrapping from the sample of estimated individual-level Spearman coefficients 5000 times. The number of observations of each row is also displayed, and missing values reflect participants who failed to complete the two rankings used to calculate a given correlation. All averages are statistically different than both zero and one at $\alpha = 0.05$.

Table 11: Correlations between Individually-assessed and Community-assessed Targeting Rankings

	All (1)	Any Member Attend (2)	Participant Attend (3)
A) Experiment	0.660	0.656	0.660
	(0.015)	(0.017)	(0.020)
N	300	270	190
B) Alatas et. al.	0.670	$0.706^{1,A}$	$0.708^{1,A}$
	(0.008)	(0.009)	(0.011)
N	1,910	1,252	791

Notes: These tables present the average Spearman correlations between individual participant-assessed rankings and the relative rankings of the same set of families/households at the community meetings. The top panel (A) shows the average of the Spearman correlations between the rankings of 10 families on the individual targeting task and the community targeting task for all participants, those who sent any family member to the meeting, and those in which the survey respondent specifically attended the meeting. The bottom panel (B) shows the correlations between the rankings of up to eight households ("Don't know" responses were allowed) in an individual unincentivized targeting task for all participants, those who sent any household member to the meeting, and those in which the survey respondent specifically attended the meeting in the Alatas et al. (2012) data set. Note that in Panel B, cases where a member who was not the household head or spouse filled out the survey and a member who was not the household head or spouse also attended the meeting, we counted this as the participant attending the meeting even though we are unable to guarantee that this is the same person. (There are only 10.) In the 4 cases with multiple conflicting attendance reports (that affect the attendance variables' values), we err on the side of assuming more attendance. Standard errors are indicated in parentheses below in both panels. Standard errors were created by bootstrapping from the sample of estimated individual-level Spearman coefficients 5000 times. The numerical superscripts indicate that a given average correlation is statistically greater than the average correlation of the column with the corresponding number (within that row), using a bootstrapped (5000 times) two-tailed test of means ($\alpha = 0.05$). The alphabetical superscripts indicate that a given average correlation is statistically greater than the average correlation of the panel row with the corresponding letter (within that column), using a bootstrapped (5000 times) two-tailed test of means ($\alpha = 0.05$). All averages are statistically different than both zero and one.

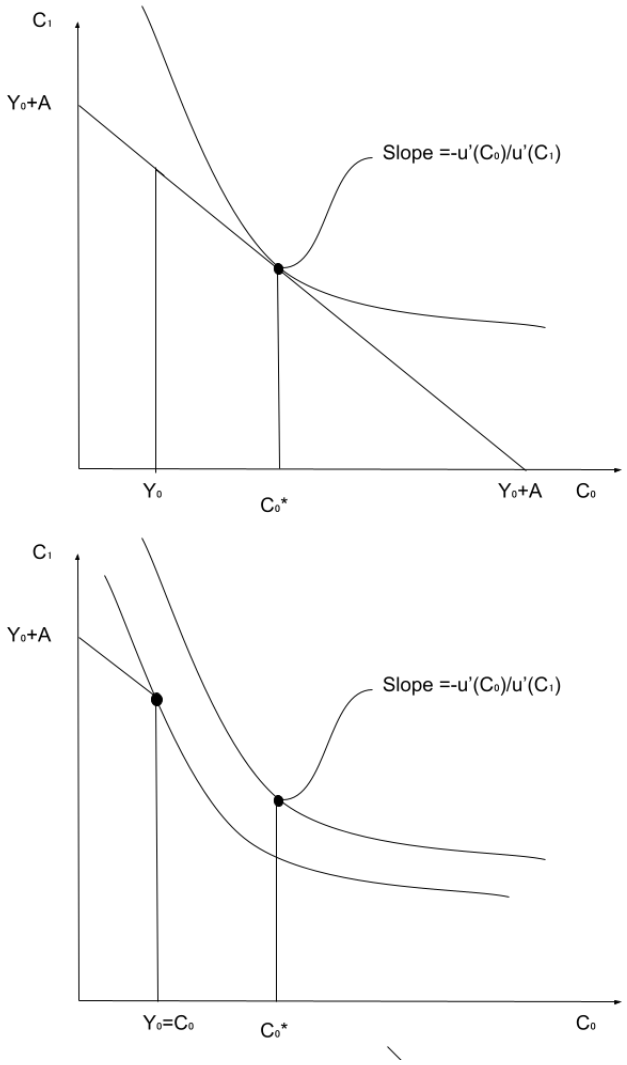
Table 12: Correlation of Individual Targeting Ranks and Community Meeting Rankings with Survey-measured Welfare Benchmarks

	Experiment			Alatas et al.		
	PC Exp (1)	MUE (2)	Asset Ind. (3)	PC Exp (4)	MUE (5)	PMT (6)
Individual Targeting Task (A)	0.293	0.315^{5}	0.462	0.321	0.220	0.454
	(0.020)	(0.019)	(0.020)	(0.009)	(0.009)	(0.008)
N	300	300	300	1,910	1,910	1,910
Community Meeting (B)	0.309	0.325^{5}	0.500	$0.363^{1,A}$	0.231	0.462
	(0.019)	(0.019)	(0.019)	(0.009)	(0.010)	(0.009)
N	300	300	300	1,910	1,910	1,910

Notes: This table presents the average Spearman correlations between individual and community participant-assessed rankings and the rankings of the same set of families/households based on the survey-measured welfare benchmarks indicated. The top panel (A) shows the average of the Spearman correlations between the rankings of families/households in on the individual targeting task and the rankings suggested by various survey benchmarks. The bottom panel (B) shows the correlations between the Community Meeting rankings of families/households and the rankings suggested by various survey benchmarks. Columns 1-3 display these statistics for our experimental sample and Columns 4-6 display them for the Alatas et al. 2012 sample, as described in the text. Standard errors are indicated in parentheses below in both panels, with number of observations listed below that. Standard errors were created by bootstrapping from the sample of estimated individual-level Spearman coefficients 5000 times. The numerical superscripts indicate that a given average correlation is statistically greater than the average correlation of the column with the corresponding number (within that row), using a bootstrapped (5000 times) two-tailed test of means ($\alpha = 0.05$). (We only test statistics constructed with the same welfare benchmark in this table, i.e. 1 against 3, 2 against 5 and 3 against 6.) The alphabetical superscripts indicate that a given average correlation is statistically greater than the average correlation of the panel row with the corresponding letter (within that column), using a bootstrapped (5000 times) two-tailed test of means ($\alpha = 0.05$). All averages are statistically different than both zero and one.

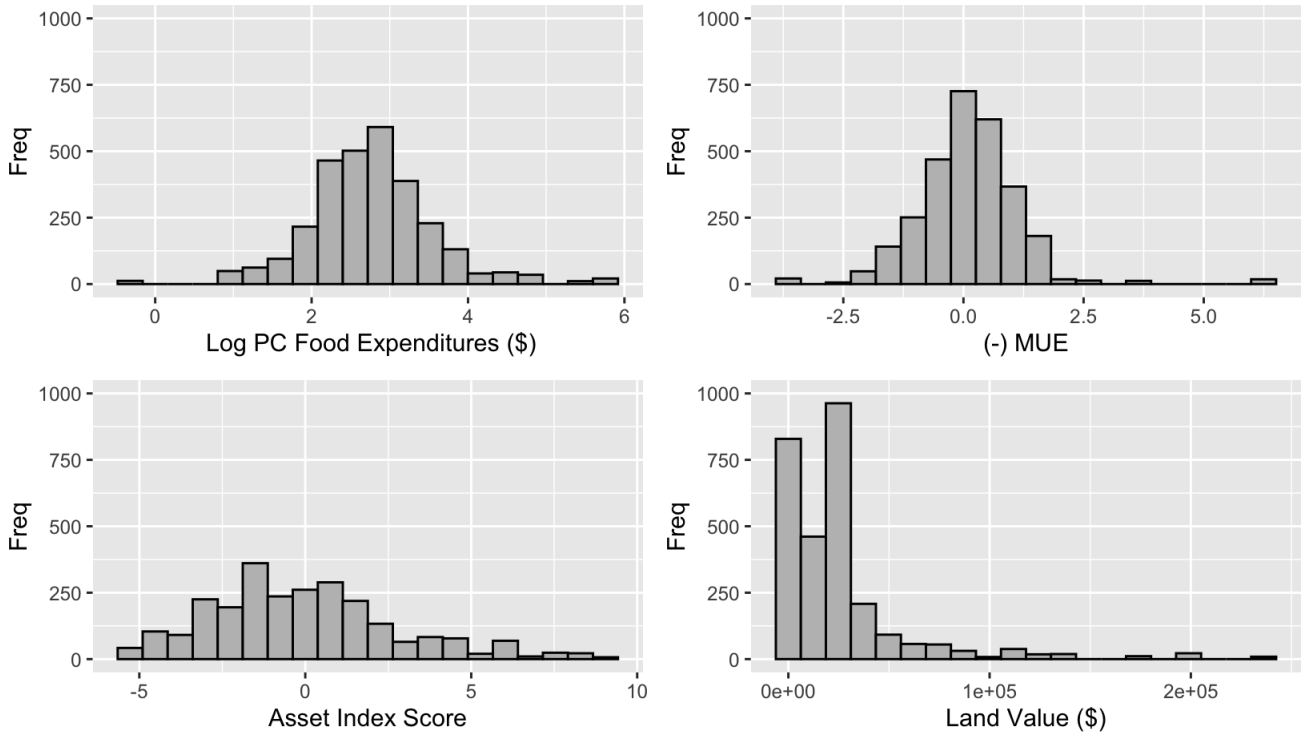
10 Figures

Figure 1: Response of Welfare Metrics to Income Shocks with and without Complete Asset Markets



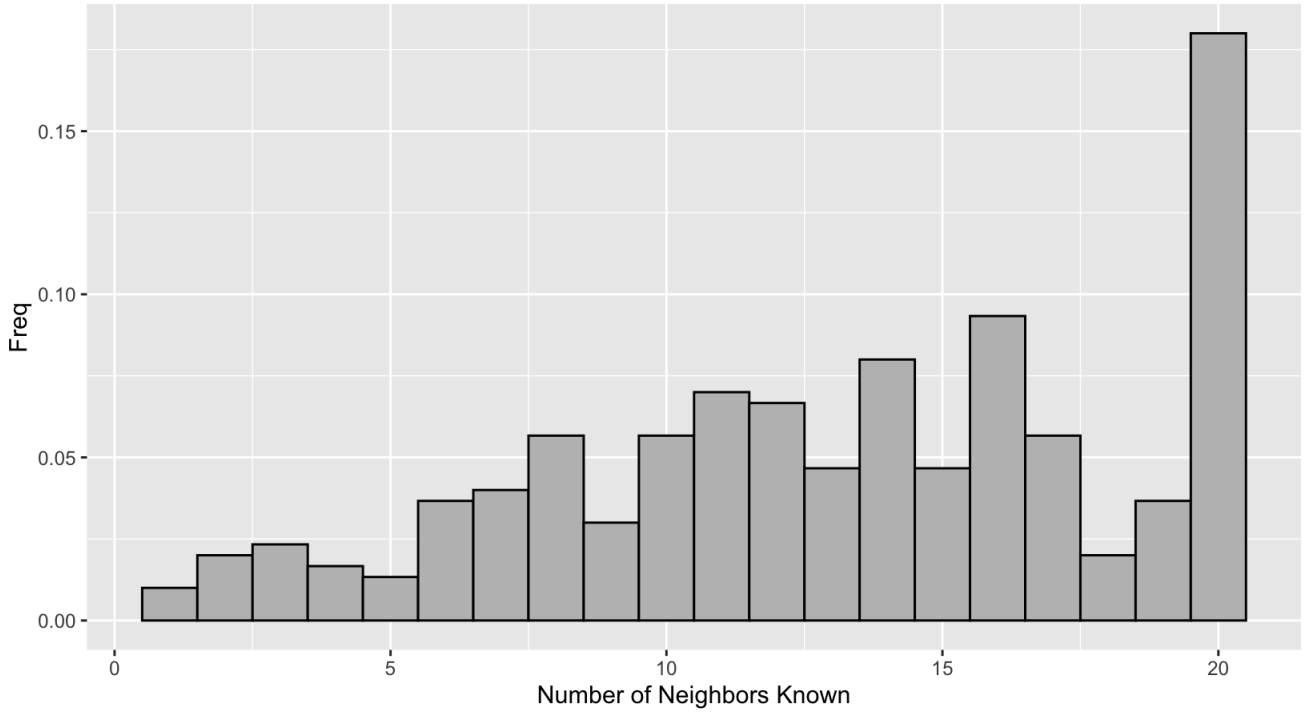
Top Figure: When individuals can borrow from their assets in the current period, they can smooth consumption over time. A decrease in assets would potentially indicate a negative income shock.
Bottom Figure: When individuals cannot borrow from their assets in the current period, they may not be able to smooth their consumption. A decrease in current consumption and increase in current marginal utility of consumption would potentially indicate a negative income shock.

Figure 2: Distribution Welfare Metrics for Ranked Population Sample



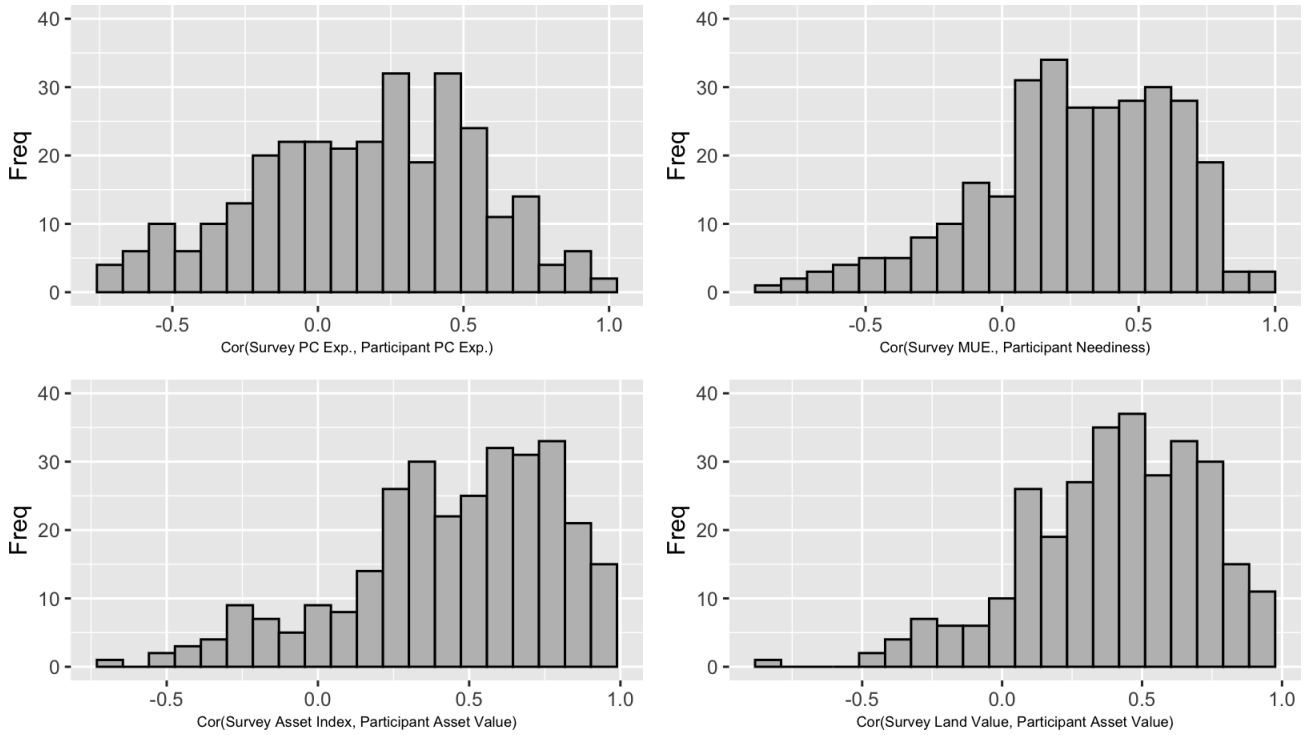
Notes: 300 participants ranked 10 households each, for up to 3,000 members of the ranked population sample. These graphs show the distribution of various welfare metrics within this sample. Due to some missing observations (either because a household has incomplete survey data or was not interviewed), the sample sizes vary slightly between plots: log PC Food Expenditures (N=2891), MUE (N=2891), Asset Index Score(N=2534), Land Value (N=2821). For the MUE, we present the distribution of the negated measures given that higher MUE values indicate worse welfare, for easy comparisons across metrics. The frequencies represent numbers of observations falling in each bin plotted.

Figure 3: Distribution of Number of Households Known Well by Participants



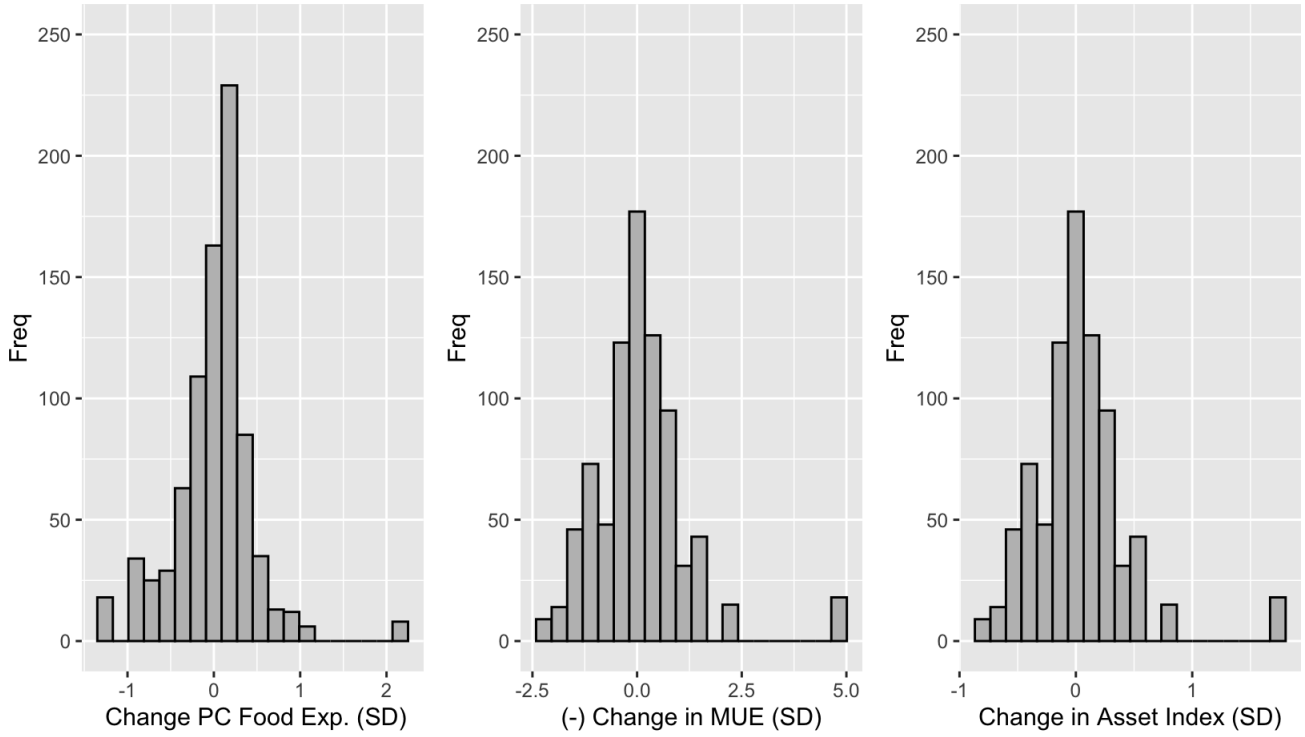
Notes: Each participant was asked how well they knew 20 randomly selected families in their community. We consider the participant knows a family well if they have 1) talked with a member of the family before frequently, 2) are a close friend/family member or 3) they are close colleagues at their job. The average household knows 13.2 (median: 14) of the presented families.

Figure 4: Distribution of the Correlations Between Survey-assessed and Participant-assessed Benchmark-specific Welfare Rankings



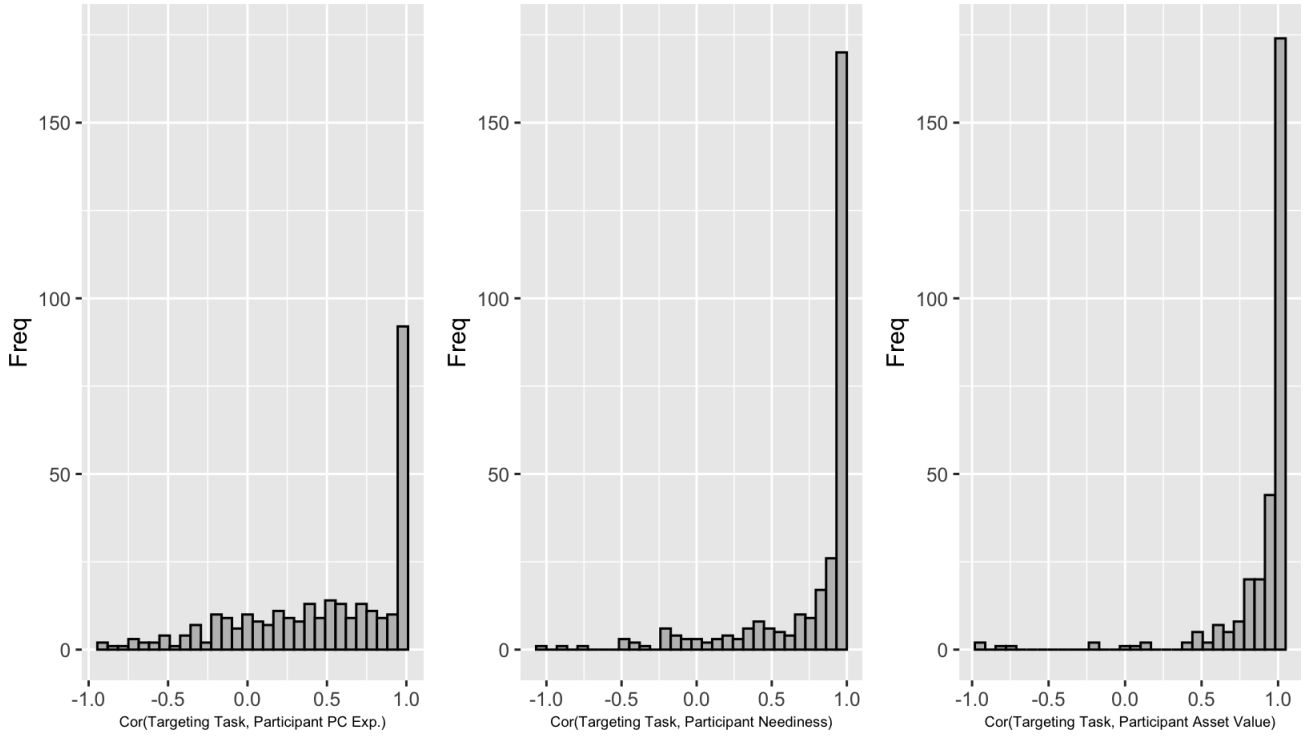
Notes: These plots show the distribution of Spearman rank correlations between each participant's benchmark specific welfare rankings and the welfare rankings suggested by our survey-calculated metrics. The frequencies represent numbers of observations falling in each bin plotted. Note that participants were not asked explicitly about land value, and hence we use participants' asset-based rankings for this measure.

Figure 5: Distribution of Changes in Survey-measured Welfare Metrics Between Baseline and Follow-up Survey.



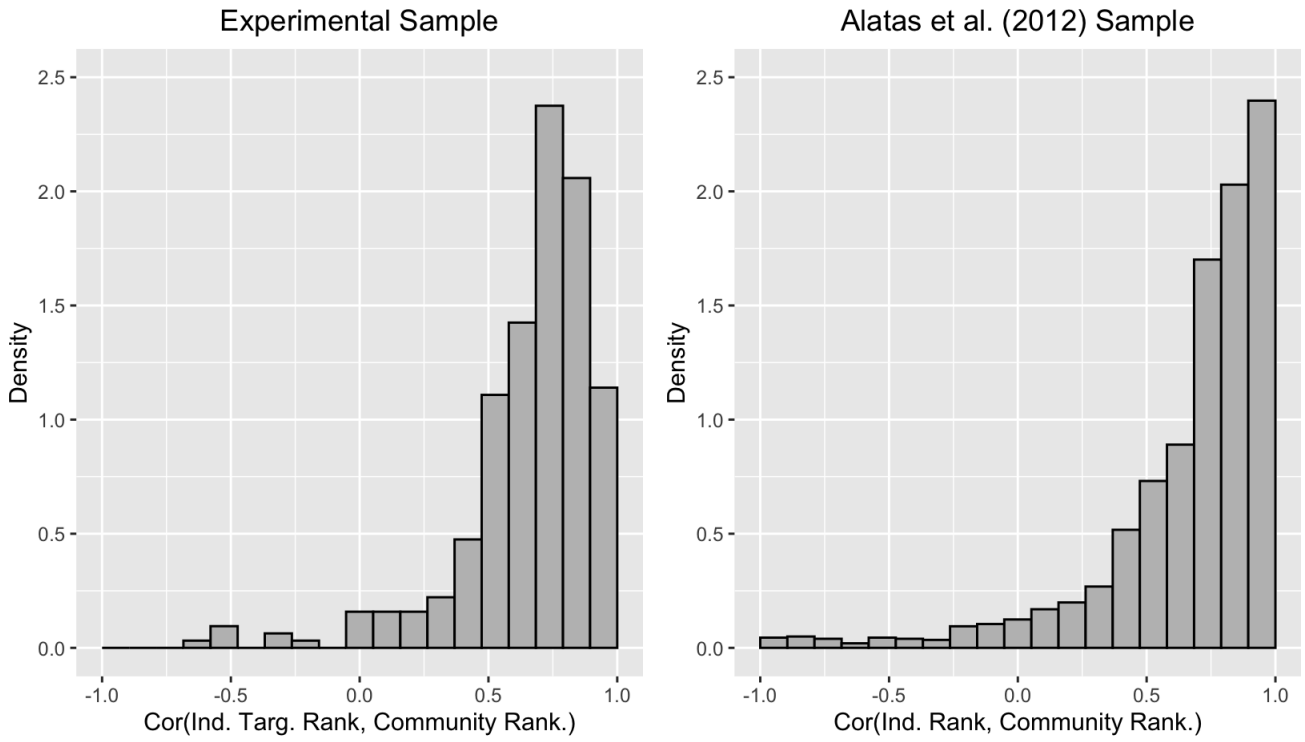
Notes: These plots show the distribution of change in survey-measured welfare for the 89 families re-interviewed in the follow-up survey, in terms of standard deviations (of the baseline sample; see Table 1). The frequencies represent numbers of observations falling in each bin plotted.

Figure 6: Distribution of the Correlations Between Participant Targeting Task Orderings and Participant-assessed Benchmark-specific Welfare Rankings



Notes: These plots show the distribution of Spearman rank correlations between each participant's benchmark specific welfare rankings and the ranking they first created in the individual targeting task. The frequencies represent numbers of observations falling in each bin plotted.

Figure 7: Distribution of the Correlations Between Participant Targeting Task Orderings and Community Targeting Task Orderings



Notes: These plots show Spearman correlations between individual participant-assessed rankings and the relative rankings of the same set of families/households at the community meetings. The graphed values represent the density of observations falling in each bin plotted, for more direct comparison between samples. The left panel shows the correlations between the rankings of 10 families in the individual targeting task and the community targeting task for the 300 participants in our experiment at 10 community meetings. The right panel shows the correlations between the rankings of up to eight households ("Don't know" responses were allowed) in an individual unincentivized targeting task, and a community targeting task for 1,910 participants in 214 community meetings (approximately 9 individuals per meeting, save some with incomplete data/who ranked two few households with complete data) in the Alatas et al. (2012) data set.

A Appendix

A.1 Appendix Tables

Table A1: Estimated Expenditure Elasticities from MUE Estimation Procedure

Good	Expenditure Elasticity
Meats	0.819
Fruits	0.785
Prepared Foods	0.746
Spices/Condiments	0.710
Beverages	0.621
Dairy Products	0.516
Tobacco Products	0.503
Starches	0.493
Other Vegetables	0.489
Fish Products	0.474
Oils and Fats	0.424
Soy Products	0.350
Rice	0.011

Notes: These estimates of relative "total expendure" elasticity were generated within the process of estimating the IMUE. Taking the ratio of any two of these can be interpreted as a ratio of total expenditure elasticities between two categories. Meats here consist of both fresh and preserved meat products, coming from beef, chicken, goats, etc. Fruits include all fresh and preserved fruits. Prepared Foods include various packaged food products (like instant noodles) and other foods consumed outside the household. Spices/Condiments include various spice varieties as well as sauces and condiments commonly used in cooking. Beverages include drinks like tea and coffee, as well as pre-packaged beverages. Dairy products include milk, cheese, and other similar items. Tobacco products consist of cigarettes and cigars. Starches include staples like potatoes, cassava, wheat, and flour. Other Vegetables includes any vegetables that were not considered as starches. Fish Products includes any type of fresh or preserved fish or seafood. Oils and Fats includes cooking oils and margarine. Soy Products includes tofu, tempe, and other similar products. Rice includes all varieties of rice. The very low relative elasticity for rice likely reflects the fact that own consumption of rice is also very prevalent in this setting, and is not contained in these expenditure elasticities.

Table A2: Asset Index Components Summary Statistics

Statistic	Mean	St. Dev.	Median	Min	Max	N
# Cupboards	2.41	1.36	2.0	0.0	9.0	266
# Tables	2.01	1.31	2.0	0.0	7.0	266
# Chairs	4.13	2.83	4.0	0.0	18.0	266
# Sofas	0.34	0.83	0.0	0.0	7.0	266
# Beds	1.90	1.21	2.0	0.0	5.0	266
# Mattresses	2.55	1.08	2.0	0.0	7.0	266
# Gas Stoves	1.02	0.50	1.0	0.0	4.0	266
# Refrigerators	0.53	0.54	1.0	0.0	2.0	266
# Rice Cookers	0.78	0.48	1.0	0.0	2.0	266
# Mixers/Blenders	0.36	0.62	0.0	0.0	2.0	266
# Fans	0.83	0.82	1.0	0.0	4.0	266
# AC Units	0.02	0.15	0.0	0.0	2.0	266
# Radios/Tape Recorders	0.20	0.41	0.0	0.0	2.0	266
# TVs	0.89	0.45	1.0	0.0	3.0	266
# DVD/VCD Players	0.06	0.24	0.0	0.0	1.0	266
# Parabolic Antenna	0.25	0.43	0.0	0.0	1.0	266
# Laptop/PC	0.15	0.43	0.0	0.0	2.0	266
# Cell Phones	1.55	1.22	1.0	0.0	6.0	266
# Bicycles	1.31	0.94	1.0	0.0	5.0	266
# Motorcycles	1.20	0.92	1.0	0.0	5.0	266
# Outboard Motors	0.01	0.12	0.0	0.0	2.0	266
# Cars/Trucks	0.07	0.33	0.0	0.0	4.0	266
# Chickens/Ducks	6.41	11.6	3.0	0.0	100.0	266
# Goats	0.61	2.01	0.0	0.0	17.0	266
# Cows/Buffaloes	0.08	0.44	0.0	0.0	4.0	266
# Horses	0.004	0.06	0.0	0.0	1.0	266
# Husker Machine	0.05	0.26	0.0	0.0	3.0	266
# Sewing Machines	0.10	0.31	0.0	0.0	2.0	266
# Electric Water Pumps	0.72	0.48	1.0	0.0	2.0	266
# Water Dispensers	0.10	0.30	0.0	0.0	1.0	266
# Ovens	0.12	0.32	0.0	0.0	1.0	266
# Washing Machines	0.252	0.468	0	0	3	266
Household owns gold/jewelry	0.17	0.38	0.0	0.0	1.0	266
Area of House (m^2)	72.1	28.3	63.0	4.0	170	266
# of Rooms	2.74	0.85	3.0	1.0	6.0	266
House has Kitchen Space	0.95	0.22	1.0	0.0	1.0	266
Per Capita Land Area (m^2)	543.8	1,093	210.0	0.0	11,900	266
Floor Material:Ceramic/Marble	0.47	0.50	0.0	0.0	1.0	266
Floor Material:Tiles/Terazzo	0.32	0.47	0.0	0.0	1.0	266
Floor Material:Cement/Red Brick	0.18	0.38	0.0	0.0	1.0	266
Roof Material: Tiles	0.91	0.29	1.0	0.0	1.0	266
Roof Material: Palm Fiber	0.06	0.24	0.0	0.0	1.0	266
Wall Material: Plaster	0.96	0.19	1.0	0.0	1.0	266
Own House	0.87	0.34	1.0	0.0	1.0	266
Main Water Source: Pumped Well	0.34	0.49	0.0	0.0	1.0	266
Main Water Source: Protected Well	0.26	0.44	0.0	0.0	1.0	266
Main Water Source: Mineral/Bottled	0.11	0.31	0.0	0.0	1.0	266
Main Water Source: Piped	0.07	0.26	0.0	0.0	1.0	266
Main Water Source: Unprotected Well	0.06	0.23	0.0	0.0	1.0	266
Main Water Source: Unprotected Well	0.06	0.23	0.0	0.0	1.0	266
Main Water Source Outside Home	0.52	0.50	1.0	0.0	1.0	266
Household has Own Latrine	0.91	0.28	1.0	0.0	1.0	266
Latrine Type: Gooseneck	0.99	0.11	1.0	0.0	1.0	266

Notes: The above variables were used in the asset index constructed using factor analysis of mixed data. The number of observations is less than 300 because observations with incomplete data in any of these categories was omitted. Note that for some of the categorical variables, categories comprising less than 5% of total responses are omitted from this table for brevity. For the per capita land ownership measure, if households responded that they did own one of five types of land asked about but were unable to provide the total land area, we imputed this value to be the median value of the type of land in question. If the household did not know whether or not they owned a given type of land, that was kept as missing (this applies to only 6 observations).

Table A3: Comparison of Participants in Baseline Survey, Follow-up Survey, and Community Meeting

Variable	Baseline	Follow-Up	Meeting	P-Value (1) v. (2)	P-Value (1) v. (3)
Respondent/Participant Male	0.394	0.348	0.473	0.434	0.059
	(0.028)	(0.051)	(0.031)		
Respondent/Participant Age	53.6	54.1	54.7	0.780	0.350
	(0.82)	(1.58)	(0.89)		
Household Size	3.30	3.20	3.30	0.576	1.00
	(0.083)	(0.153)	(0.088)		
Household Head: Male	0.849	0.854	0.840	0.918	0.759
	(0.021)	(0.038)	(0.022)		
Household Head: Age	56.6	58.2	57.2	0.340	0.606
	(0.786)	(1.50)	(0.826)		
Household Head: Disabled	0.017	0.011	0.019	0.684	0.867
	(0.007)	(0.011)	(0.008)		
Household Head: Any Education	0.863	0.831	0.851	0.482	0.695
	(0.020)	(0.040)	(0.022)		
Household Head: Sr. High School	0.281	0.281	0.260	0.999	0.580
	(0.026)	(0.048)	(0.027)		
Household Head: Married	0.783	0.775	0.773	0.885	0.789
	(0.024)	(0.044)	(0.026)		
Household Head: Employed	0.833	0.787	0.825	0.344	0.813
	(0.022)	(0.044)	(0.023)		
Household Head: Employed (Ag.)	0.572	0.584	0.572	0.837	0.989
	(0.029)	(0.053)	(0.030)		
Household Head: Born in Village	0.847	0.876	0.852	0.467	0.863
	(0.021)	(0.035)	(0.022)		
Household Head: Years in Community	44.1	48.3	44.9	0.126	0.661
	(1.29)	(2.43)	(1.36)		
Household Member: Any Education	0.917	0.865	0.911	0.198	0.814
	(0.016)	(0.036)	(0.017)		
Household Member: Sr. High School	0.527	0.483	0.522	0.473	0.916
	(0.029)	(0.053)	(0.030)		
Household Member: Disabled	0.047	0.056	0.052	0.729	0.776
	(0.012)	(0.025)	(0.014)		
Javanese (Ethnicity)	0.990	0.978	0.989	0.460	0.897
	(0.006)	(0.016)	(0.006)		
Muslim	0.980	1.00	0.978	0.014*	0.854
	(0.008)	(0.00)	(0.009)		
Speak Java	0.973	0.966	0.974	0.742	0.956
	(0.009)	(0.019)	(0.010)		

Local Official	0.0300	0.0337	0.0259	0.864	0.768
	(0.010)	(0.019)	(0.010)		
Know Local Official	0.977	1.00	0.974	0.008**	0.843
	(0.009)	(0.00)	(0.010)		
Participate in Community Org.	0.967	0.978	0.963	0.566	0.811
	(0.010)	(0.016)	(0.012)		
Weekly PC Food Expenditures ($PPP)	22.98	20.98	22.94	0.434	0.990
	(1.98)	(1.61)	(2.17)		
Weekly PC Expenditures ($PPP)	37.76	35.48	37.31	0.523	0.898
	(2.36)	(2.67)	(2.55)		
Asset Index Score	-0.00	0.57	-0.12	0.871	0.632
	(0.186)	(0.307)	(0.193)		
Land Area Owned (hectares)	0.136	0.142	0.129	0.790	0.669
	(0.011)	(0.021)	(0.011)		
MUE	0.0846	0.0115	0.121	0.564	0.667
	(0.058)	(0.112)	(0.061)		
Ladder Step	2.74	2.91	2.72	0.181	0.838
	(0.060)	(0.113)	(0.065)		
Had Shock	0.300	0.281	0.296	0.728	0.923
	(0.027)	(0.048)	(0.028)		
Receives Govt. Benefits	0.917	0.888	0.926	0.437	0.682
	(0.016)	(0.034)	(0.016)		
Receives COVID Benefits	0.353	0.281	0.367	0.192	0.741
	(0.028)	(0.048)	(0.029)		

Notes: This table compares baseline characteristics of participants who (1) answered the baseline survey (N=300), (2) answered the follow-up survey in 3 communities (N=89), and (3) participated in the community meeting in their community (N=270). Means of each variable are presented, with standard errors in parentheses below. The p-values correspond to a t-test of means between the indicated columns. One star indicates significance at $\alpha = 0.05$, and two stars indicate significance at $\alpha = 0.01$.

Table A4: Overall Familiarity with Other Households

Response	Frequency
Not familiar at all	5.7%
Know of/ can recognize someone in the family but have never interacted	13.9%
Have talked to a member of the family before infrequently	14.6%
Have talked with a member of the family before frequently	58.8%
Close friend or family member	6.7%
Close colleague at job/position	0.3%

Notes: Each participant (300 total) was asked how well they knew 20 households. Out of the 6,000 reports (300x20), these are the percentage of cases in which each of these relationships between the participant and the respondent was reported. Note that in 190 cases (3.2% of overall cases), participants were asked about their own family. In these cases, participants mostly answered "Close friend or family member" (about 80% of the cases–about 2.6% of the overall sample cases), or "Have talked before frequently" (about 17% of the cases–about 0.5% of the overall sample cases).

Table A5: Experimental Task Descriptions

Task	Order	Prompt	Incentive
Individual Targeting Task	1	"Rank the families from poorest to richest."	A random number is drawn between 0 and 10.That number of households ranked poorest will receive an additional Rp. 5,000 ($US 1.06 PPP).[56]
Preference Elicitation Task	2	"Rank the families from that which you would most like to receive an additional cash payment of Rp. 5,000 to that which you would least like to receive an additional cash payment of Rp. 5,000. You can use any criteria that you like to determine the ranking."	A random number is drawn between 0 and 10.That number of households ranked most preferable will receive an additional Rp. 5,000 ($US 1.06 PPP).
Expenditure Benchmark Information Task	3-6	"Like we are doing today, we have surveyed families in your community to gain information on how much money they spend on food in one week. Please rank the households from spending the least on food per person a week' to spending the most on food per person a week'."	Participants are paid more based on the closer the rankings are to what we calculate in our data (Per capita food expenditures).[57]
Neediness Benchmark Information Task	3-6	"Like we are doing today, we have surveyed families in your community to gain information on the value of their asset wealth. This includes the durable, valuable things they own, like their house, appliances, vehicles, livestock, and other items that retain value over time. Please rank the households from least asset wealth' to most asset wealth'."	Participants are paid more based on the closer the rankings are to what we calculate in our data (Asset index).
Asset Benchmark Information Task	3-6	"Like we are doing today, we have surveyed families in your community and calculated a measure of which households are most needy of more money and would benefit most from additional money.˜ Please rank the households from least needy of additional money' to most needy of additional money'."	Participants are paid more based on the closer the rankings are to what we calculate in our data (MUE).
Second-Order Beliefs Elicitation Task	3-6	"We asked other members in your community also to complete the same exercises that you are doing now, including the one you did first where you ranked households from poorest' to richest' with no further instructions. Rank the households from poorest' to richest' in a way that is most similar to how other respondents ranked them in the first round."	Participants are paid more based on the closer the rankings were to others' (by calculating how many of their pairwise family comparisons matched family comparisons made by other participants).

[56] Number chosen is not revealed to the participant in this task or the following Preference Elicitation Task. Source of payment is not revealed to the recipient.

[57] Participants are not told how "correct" they are at any point in this task or any other of the 3rd through 6th tasks. Participants are not explicitly told how we will calculate the metric in parentheses, just that they want to match our collected data.

Table A6: Robustness of Participant-based/Survey-based Rank Correlations to Alternate Survey-based Metrics: Expenditures

PC Exp.	AE Exp.	PC Cons.	PC Predicted Exp.
0.160	0.180	0.104	0.146
(0.022)	(0.023)	(0.023)	(0.021)

Notes: The values presented are the average Spearman (ranking) correlations between 300 participants' rankings of 10 households, and the ranking suggested by various alternative survey-based metrics. Standard errors are in parentheses below. Standard errors were created by bootstrapping from the sample of estimated individual-level Spearman coefficients 5000 times. The first column shows the correlation with our preferred measure (per capita food consumption expenditures) ranking with participant consumption expenditure rankings. The following columns instead utilize: adult equivalent per capita expenditures (women =0.8 men, children=0.5 men), value of per capita food consumption (including self-production/gifts), and per capita food expenditures as predicted by the MUE estimation procedure. A star indicates that a given average correlation is statistically different than the correlation using our preferred measure in the first column, under a bootstrapped (5000 times) two-tailed test of means ($\alpha = 0.05$).

Table A7: Robustness of Participant-based/Survey-based Rank Correlations to Alternate Survey-based Metrics: Neediness

MUE (Exp. Only)	MUE (All Cons.)
0.272	0.266
(0.021)	(0.020)

Notes: The values presented are the average Spearman (ranking) correlations between 300 participants' rankings of 10 households, and the ranking suggested by an alternative survey-based metric. Standard errors are in parentheses below. Standard errors were created by bootstrapping from the sample of estimated individual-level Spearman coefficients 5000 times. The first column shows the correlation with our preferred measure (MUE calculated using only consumption expenditures) ranking with participant neediness rankings. The following column instead utilizes: the MUE calculated using both expenditures and the value of any self-production/gifts. A star indicates that a given average correlation is statistically different than the correlation using our preferred measure in the first column under a bootstrapped (5000 times) two-tailed test of means ($\alpha = 0.05$).

Table A8: Robustness of Participant-based/Survey-based Rank Correlations to Alternate Survey-based Metrics: Assets

Asset Index	Land Value	PC Land Value	Land Area	Asset Count	House Index
0.445	0.398	0.345*	0.330*	0.304*	0.359*
(0.020)	(0.018)	(0.020)	(0.020)	(0.023)	(0.020)

Notes: The values presented are the average Spearman (ranking) correlations between 300 participants' rankings of 10 households, and the ranking suggested by various alternative survey-based metrics. Standard errors are in parentheses below. Standard errors were created by bootstrapping from the sample of estimated individual-level Spearman coefficients 5000 times. The first column shows the correlation with our preferred measure (asset index) ranking with participant asset rankings. The following columns instead utilize: total land value, per capita land value, land area, an unweighted sum of the number of durables/assets owned, and an index of only housing-related variables. A star indicates that a given average correlation is statistically different than the correlation using our preferred measure in the first column under a bootstrapped (5000 times) two-tailed test of means ($\alpha = 0.05$).

Table A9: Ranking Accuracy vs. Familiarity with other Participants (Ranker and Ranked Family)

	Dependent variable:		
	Absolute Difference: Survey/Participant Rank Position		
	Exp.	Need	Asset
	(1)	(2)	(3)
% Times Known (Ranker)	0.558*	−0.145	0.163
	(0.259)	(0.229)	(0.237)
% Times Known (Ranked)	0.199	−0.135	0.386
	(0.251)	(0.249)	(0.232)
# Times Asked About (Ranker)	−0.012	−0.004	0.013
	(0.009)	(0.009)	(0.010)
# Times Asked About (Ranked)	0.001	−0.019	−0.008
	(0.013)	(0.012)	(0.011)
Constant	2.637***	3.359***	1.822***
	(0.333)	(0.340)	(0.318)
Observations	2,891	2,872	2,507
R^2	0.005	0.001	0.004
Adjusted R^2	0.004	0.00004	0.002
F-Statistic	2.74	1.08	1.79

Notes: The above table presents regression analysis considering whether the absolute value of the difference between a family's survey-assessed and participant-assessed rank position (a measure of concordance) can be explained by how well known the ranking participant or ranked family is. Observations are at the ranking participant by ranked family level. Coefficients are presented with standard errors in parentheses below. Standard errors are cluster (by ranking participant) bootstrapped with 5000 bootstrap samples. One star indicates significance at $\alpha = 0.05$, and two stars indicate significance at $\alpha = 0.01$. The dependent variables are the absolute values of the difference between a family's survey-assessed and participant-assessed rank position for our three welfare benchmarks: per capita expenditures, neediness/MUE, and assets. The independent variables are the percentage of all times a ranking participant was asked about in which they were reported to be in one of the "well-known" indication categories, the percentage of all times a ranked family was asked about in which they were reported to be in one of the "well-known" indication categories, the number of times another participant was asked about the ranking participant (since family lists were randomly generated for each participant), and the number of times another participant was asked about the ranked family (since family lists were randomly generated for each participant). Note that a negative coefficient on the first independent variable indicates that more well-known individuals rank in a way that is more concordant with the survey rankings in the given welfare-benchmark domain, and a negative coefficient on the second independent variable indicates that more well-known families are ranked in a way that is more concordant with the survey rankings in the given welfare-benchmark domain. Numbers of observations less than 890 reflect cases in which data to construct one of the relevant variables is missing.

Table A10: Percentage of Survey-identified "Poorest" Families identified by Participants

	Exp	MUE	Assets
% identified of 2 poorest	29.2%	34.6%	46.3%
% identified of 3 poorest	39.7%	44.6%	52.3%
% identified of 4 poorest	47.1%	52.9%	58.0%

Notes: This table displays the percentage of cases where the set of the families ranked as poorest by each survey-assessed benchmark are also ranked poorest by the participants' benchmark-specific assessment. The numbers in each row denotes the "cut-off" for being considered one of the poorest families. For instance, in the first row marked "% identified of 2 poorest", we consider (the approximately 300x2=600) cases of families that would be ranked as 1 and 2 poorest by the survey, and look at the percentage that were also ranked as 1 or 2 poorest by the participant. Accordingly the corresponding samples increase in the following rows to the approximately 900 families ranked 1, 2, or 3 poorest, and the approximately 1200 families ranked 1, 2, 3, or 4 poorest. Percentages are all significantly greater than 20% (row 1), 30% (row 2), or 40% (row 3) which is what we would expect if choosing at random. All percentages are also significantly lower than 100%.

Table A11: Simulated Effects of Random Survey Measurement Error

	Survey Noise Level			
Community Noise	+/- 0.5SD	+/- 1SD	+/- 2SD	+/- 3SD
ρ (truth)	0.909	0.784	0.557	0.414
	[0.887, 0.928]	[0.740, 0.823]	[0.478, 0.629]	[0.320,0.503]
ρ (+/- 0.5SD)	0.857	0.747	0.534	0.398
	[0.825, 0.886]	[0.696, 0.793]	[0.453, 0.608]	[0.302,0.491]
ρ (+/- 1SD)	0.747	0.659	0.475	0.355
	[0.695, 0.794]	[0.593, 0.719]	[0.386, 0.558]	[0.255,0.453]
ρ (+/- 2SD)	0.533	0.475	0.348	0.261
	[0.449, 0.609]	[0.385, 0.557]	[0.247, 0.444]	[0.152,0.365]
ρ (+/- 3SD)	0.398	0.356	0.262	0.197
	[0.301, 0.490]	[0.256, 0.450]	[0.156, 0.364]	[0.085,0.303]

Notes: This table shows the results of a simulation exercise in which random noise was added to the survey and participant reports of the "true value" of a welfare metrics, and average Spearman coefficients (over participants) were estimated between the resulting "true" and "noisy" rankings for 300 simulated participants. The noise levels indicated are the parameters of a uniform distribution from which the additive noise was drawn (e.g. +/- 0.5SD indicates random noise was drawn uniformly from [-0.5SD, 0.5SD], where SD is the standard deviation of the normal distribution from which the data was drawn. Each simulation was repeated 5,000 times and the values displayed are the means of the participant average Spearman coefficient from each of the 5,000 iterations. 95% bootstrap confidence intervals are in brackets below.

Table A12: Correlation Between Baseline and Follow-up Participant-assessed and Survey-assessed Benchmark-specific Rankings

	Exp (1)	MUE (2)	Asset (3)
A) Participant	0.335	0.394	$0.667^{1,2}$
	(0.040)	(0.043)	(0.033)
N	89	89	87
B) Survey	0.405	0.402	$0.773^{1,2,A}$
	(0.032)	(0.032)	(0.022)
N	89	89	89

Notes: This table presents the average Spearman correlations between the baseline and follow-up rounds of benchmark-specific rankings. The top panel (A) shows the average of the Spearman correlations between participant-assessed rankings between rounds for 89 follow-up survey participants. (There are a few missing values where indicated for individuals who accidentally did not complete the asset ranking task in the first round.) The bottom panel (B) shows the average of the Spearman correlations between survey-assessed rankings between rounds for all 89 follow-up survey participants. Standard errors are indicated in parentheses below in both panels. Standard errors were created by bootstrapping from the sample of estimated individual-level Spearman coefficients 5000 times. The numerical superscripts indicate that a given average correlation is statistically greater than the average correlation of the column with the corresponding number (within that row), using a bootstrapped (5000 times) two-tailed test of means ($\alpha = 0.05$). The alphabetical superscripts indicate that a given average correlation is statistically greater than the average correlation of the panel row with the corresponding letter (within that column), using a bootstrapped (5000 times) two-tailed test of means ($\alpha = 0.05$). All averages are statistically different than both zero and one.

Table A13: Correlation Between Survey-assessed and Participant-assessed Benchmark-specific Welfare Rankings: Follow-up Survey

	Exp (1)	MUE (2)	Asset (3)
A) Baseline (3 communities)	0.150^{B}	0.301^{1}	0.399^{1}
	(0.041)	(0.034)	(0.052)
N	89	89	89
B) Follow-up	-0.201	0.249^{1}	$0.414^{1,2}$
	(0.037)	(0.037)	(0.051)
N	89	88	87

Notes: The values presented are the average Spearman (ranking) correlations between 89 follow-up survey participants' rankings of 10 households during the baseline survey (top panel) and follow-up survey (bottom panel), and the ranking suggested by our survey for each of the welfare metrics considered (described in detail in the text). Standard errors are indicated in parentheses below in both panels, as well as sample size. Standard errors were created by bootstrapping from the sample of estimated individual-level Spearman coefficients 5000 times. The superscripts indicate that a given average correlation is statistically greater than the average correlation of the column with the corresponding number (within that row), using a bootstrapped (5000 times) two-tailed test of means ($\alpha = 0.05$). The alphabetical superscripts indicate that a given average correlation is statistically greater than the average correlation of the panel row with the corresponding letter (within that column), using a bootstrapped (5000 times) two-tailed test of means ($\alpha = 0.05$). All averages are both statistically different than zero and one.

Table A14: Predictors of Participants' Benchmark-specific Rankings in the Follow-up Survey

	Dependent variable: Follow-up Participant-assessed Ranking of:					
	Exp	Need	Asset	Exp	Need	Asset
	(1)	(2)	(3)	(4)	(5)	(6)
Baseline Part. Rank	0.331**	0.298**	0.399**	0.322**	0.278**	0.402**
	(0.043)	(0.056)	(0.054)	(0.043)	(0.058)	(0.049)
Change in Survey-assessed Rank	−0.134**	0.026	−0.009			
	(0.029)	(0.034)	(0.043)			
Follow-up Survey-assessed Rank				−0.235**	0.111**	0.046
				(0.038)	(0.039)	(0.037)
Community Meeting Rank	−0.055	0.204**	0.378**	0.021	0.185**	0.371**
	(0.047)	(0.056)	(0.050)	(0.047)	(0.058)	(0.052)
Constant	4.071**	2.660**	1.310**	4.956**	2.304**	1.144**
	(0.287)	(0.250)	(0.232)	(0.357)	(0.257)	(0.221)
Observations	829	810	611	829	810	717
R^2	0.125	0.202	0.490	0.154	0.211	0.512
Adjusted R^2	0.121	0.199	0.487	0.151	0.208	0.510
F-statistic	39.12	67.91	194.1	49.9	71.78	249

Notes: The above table presents regression analysis showing predictors of participants-assessed benchmark-specific rankings in the follow-up survey. Observations are at the ranking participant by ranked family level. Coefficients are presented with standard errors in parentheses below. Standard errors are cluster (by ranking participant) bootstrapped with 5000 bootstrap samples. One star indicates significance at $\alpha = 0.05$, and two stars indicate significance at $\alpha = 0.01$. The dependent variables are participant-assessed welfare rankings in the follow-up survey for our three welfare benchmarks: per capita expenditures, neediness/MUE, and assets. The independent variables in columns 1-3 are participants' baseline ranking positions of the same families according to the same benchmark (in the column title), the change in survey-assessed rankings between the baseline and follow-up rounds of those families according to that same benchmark, and the relative ranking position assigned to the family during the community meeting exercise (omitting other families ranked during the meeting). In columns 4-6, instead of including the change in survey-assessed benchmark-specific rankings, we instead include the survey-assessed benchmark specific ranking in the follow-up survey. Numbers of observations less than 890 reflect cases in which data to construct one of the relevant variables is missing.

Table A15: Proxy Means Test Scores

Statistic	Mean	St. Dev.	Median	Min	Max	N
PPI	59.66	10.15	61.5	32.0	74.0	300
SPS	51.44	11.03	52.0	14.0	79.0	300

Notes: Scores are at the family level, for each of the 300 families surveyed at baseline. The PPI formula uses: province of residence, total household members, household members ages zero to five, household members ages six to ten, ownership of a refrigerator, ownership of a motorcycle, floor material, toilet type, cooking fuel source, and recent internet use. Practically, province is the same for all sampled families (and hence has little meaning) and recent internet use is omitted (as we did not collect this information during the main baseline survey). The SPS formula uses: regency of residence, total household members, total currently working household members, total members with "permanent" jobs, receipt of the *Rastra* program rice subsidy, ownership of a refrigerator, ownership of a motorcycle, toilet type, cooking fuel source, and whether the female head of household has a cell phone. Practically, there is no variation in the "regency" variable in our sample, and proxy the female head's ownership of a cell phone with whether the household owns more than one cell phone (since we did not collect this information).

Table A16: Percentage of Survey-identified Poorest also chosen by Proxy Means Test and Participant Ranks

Poverty Line		Proxy Means Scores		Participant Ranks			
		PPI	SPS	Targeting	Exp.	Need	Assets
		(1)	(2)	(3)	(4)	(5)	(6)
20%	Survey Exp.	0.398^{P}	0.383^{P}	0.322	0.292	0.331	0.328
	Survey MUE (Need)	0.352	0.368^{4}	0.355	0.313	0.346	0.367
	Survey Asset Index	$0.575^{2,P}$	$0.485^{4,5}$	0.463	0.397	0.423	0.463
30%	Survey Exp.	$0.471^{4,5,6}$	$0.462^{4,5}$	0.431	0.396	0.411	0.426
	Survey MUE (Need)	0.454^{4}	0.484^{4}	0.466	0.412	0.446	0.461
	Survey Asset Index	0.600^{P}	$0.563^{4,5}$	0.528	0.486	0.480	0.524
40%	Survey Exp.	0.572^{P}	0.575^{P}	0.517	0.471	0.502	0.519
	Survey MUE (Need)	0.527	$0.568^{1,4}$	0.543	0.490	0.529	0.543
	Survey Asset Index	0.628^{P}	0.624^{P}	0.580	0.555	0.545	0.580

Notes: This table displays the percentage of cases where the set of the families ranked as poorest by each survey-assessed benchmark are also ranked poorest by the proxy means test scores and various participants' assessments. The "Poverty Line" in each row denotes the "cut-off" for being considered one of the poorest families. For instance, in the first row marked "%20," we consider (the approximately 300x2=600) cases of families that would be ranked as 1 and 2 poorest by the survey, and look at the percentage that were also ranked as 1 or 2 poorest by the corresponding column metric. The numerical superscripts indicate that a given average "percentage poorest identified" is statistically greater than the average percentage in the column with the corresponding number (within that row), using a bootstrapped (5000 times) two-tailed test of means ($\alpha = 0.05$). The "P" superscript indicates that a given average percentage identified is significantly greater than all participant rankings (columns 3-6) using the same tests, for visual simplicity. We only test here whether the proxy means test scores (columns 1-2) are significantly different from all other columns; we do not test for statistical differences among columns 3-6. Percentages are all significantly greater than 20% (row 1), 30% (row 2), or 40% (row 3) which is what we would expect if choosing at random. All percentages are also significantly lower than 100%.

Table A17: Most Common Words in Participant Explanations of Their Individual Targeting Task Rankings

Order	Word	Occurrences
1	sawah (rice field)	430
2	luas (land area)	118
3	buruh (labor/laborer)	104
4	lahan (land)	102
5	anak (child)	88
6	janda (widow)	75
7	usaha (business)	62
8	penghasilan (income)	59
9	pensiunan (pension)	51
10	sendiri (alone)	48
11	tani (farmer/peasant)	43
12	bekerja (work/labor)	42
13	rumah (house)	38
14	kerja (work/job)	33
15	pekerjaan (job/profession)	33
16	orang (person)	31
17	pertanian (agriculture)	27
18	tanggungan (dependent)	27
19	tua (old)	27
20	petani (farmer/agriculturalist)	26

Notes: After completing the Individual Targeting Task, participants were asked to explain their reason behind the pairwise placement of 3 randomly selected pairs of families among the 10 ranked families in an open-ended response. (This was done successfully in 298/300 cases.) Here we list the 20 most commonly mentioned words throughout these responses (after extracting common stop words), as well as the number of times each was mentioned in any response. Indonesian words are listed with approximate English translations in parentheses. Cases where the term household "rumah tangga" is mentioned are not counted in the tally of the word "rumah" or house.

Table A18: Most Three-word Phrases in Participant Explanations of Their Individual Targeting Task Rankings

Order	Trigram Phrase	Occurrences
1	tidak punya sawah (Does not have rice field)	56
2	sawah lebih banyak (More rice fields)	34
3	hanya buruh tani (Only Farm Laborer)	23
4	punya sawah dan (Has rice field and)	13
5	janda tidak punya (Widow does not have)	13
6	punya sawah lebih (Have more fields)	11
7	punya sawah sendiri (Have your own farm)	10
8	sawahnya lebih sedikit (Less fields)	10
9	tidak punya tanggungan (Does not have dependents)	10

Notes: After completing the Individual Targeting Task, participants were asked to explain their reason behind the pairwise placement of 3 randomly selected pairs of families among the 10 ranked families in an open-ended response. (This was done successfully in 298/300 cases.) Here we list the 9 most common short 3 word-phrases throughout these responses (retaining common stop words), as well as the number of times each was mentioned in any response.(We list all phrases with at least 10 occurrences.) Indonesian phrases are listed with approximate English translations in parentheses.

Table A19: Evidence of the Lack of Differential Ranking Behavior of Close Family and Friends or other Well-Known Families

	Dependent variable:					
	Individual Targeting Rank					
	(1)	(2)	(3)	(4)	(5)	(6)
Close Friend/Family	0.298	0.073	0.142			
	(0.239)	(0.134)	(0.215)			
Known Well				0.210	0.096	0.255*
				(0.130)	(0.066)	(0.128)
Controls	None	Part.	Survey	None	Part.	Survey
Observations	3,000	2,970	2,534	3,000	2,970	2,534
R^2	0.001	0.805	0.221	0.001	0.805	0.222
Adjusted R^2	0.0002	0.805	0.220	0.001	0.805	0.221
F Statistic	1.749	3062	179.2	2.834	3066	180.6

Notes: The above table presents regression analysis regarding whether individuals systematically rank their close friends/family or other well-known contacts as poorer so that they receive benefits, controlling for their actual welfare status. Observations are at the ranking participant by ranked family pair level. Coefficients are presented with standard errors in parentheses below. Standard errors are cluster (by ranking participant) bootstrapped with 5000 bootstrap samples. One star indicates significance at $\alpha = 0.05$, and two stars indicate significance at $\alpha = 0.01$. The dependent variable is the position assigned to a family by a participant in the individual targeting task. The independent variable in columns 1-3 is a binary indicator of whether the ranked family was indicated (before the ranking tasks were introduced) to be a family member or close friend of the participant's family. The independent variable in columns 4-6 is a binary indicator of whether the ranked family was indicated (before the ranking tasks were introduced) to be known well (either speak frequently, close work colleague, or close family/friend). All specifications include a constant (not displayed). Specifications with "Part." or Participant controls, include the expenditure, neediness, and asset ranking positions provided by the participant for the family in question as controls. Specifications with "Survey" controls include the expenditure, neediness, and asset ranking positions suggested by our survey for the family in question as controls. Note that for all rankings, lower scores mean that a ranked family is worse off and higher scores mean that they are better off.

Table A20: Most Common Community Meeting Welfare Descriptors- Alatas et al. Data

	Word	Mentioned	Word	Top Importance
1	rumah (house/home)	202	rumah (house/home)	143
2	pekerjaan (job/profession)	144	penghasilan (income)	98
3	pendidikan (education)	121	pekerjaan (job/profession)	91
4	penghasilan (income)	109	pendidikan (education)	54
5	kendaraan (vehicle)	88	harta (property/wealth)	28
6	kesehatan (health)	62	tanggungan (dependents)	27
7	tanggungan (dependents)	52	pendapatan (income/revenue)	25
8	harta (property/wealth)	49	kesehatan (health)	24
9	tanah (land)	44	sawah (rice field)	24
10	ternak (livestock)	42	tanah (land)	24
11	sawah (rice field)	34	kendaraan (vehicle)	16
12	kepemilikan (belongings)	33	ternak (livestock)	11
13	lahan (land)	30	ekonomi (HH mgmt/econ.)	10
14	pendapatan (income/revenue)	30	kepemilikan (belongings)	10
15	anak (child)	21	lahan (land)	9
16	keluarga (family)	17	keluarga (family)	7
17	perabot (furniture)	17	aset (asset)	6
18	usaha (business)	17	kebutuhan (needs)	6
19	kekayaan (wealth)	13	kekayaan (wealth)	5
20	aset (asset)	12	pengeluaran (expenditures)	5
20 (tie)	ekonomi (HH mgmt./econ.)	12	sehari-hari (daily)	5

Notes: At the community meeting exercises in Alatas et al. (2012), communities were asked to first list terms representing ways in which one could differentiate the welfare status of different households. They were then asked which of those terms were most important to the community's view of welfare. These terms usually only consist of one or two words. Here we list the 20 most commonly mentioned words in each of these exercises (after extracting common stop words), as well as the number of times each was mentioned in any response. Accounting for ties, we list 21 words in each category. Indonesian words are listed with approximate English translations in parentheses. Cases where the term household "rumah tangga" is mentioned are not counted in the tally of the word "rumah" or house.

Table A21: Most Common Community Meeting Welfare Descriptor Words- Experimental Community Meetings

#	Word ("Poor")	Freq.	Word ("Rich")	Freq.
1	kebutuhan (needs)	10	sawah (rice field)	10
2	sawah (rice field)	9	kebutuhan (needs)	6
3	hutang (debt)	9	hutang (debt)	5
4	pendapatan (income)	7	anak (child)	5
5	rumah (house)	7	kecukupan (sufficiency)	4
6	pekerjaan (job/profession)	6	mobil (car)	4
7	tetap (permanent)	6	rumah (house)	4
8	uang (money)	5	pengeluaran (expenditures)	4
9	anak (child)	5	penghasilan (income)	4
10	sehari-hari (daily)	4	uang (money)	3
11	makan (eat)	4	sehari-hari (daily)	3
12	sulit (tough)	4	kendaraan (vehicle)	3
13	bekerja (work/labor)	3	harta (property/wealth)	3
14	kekurangan (deficiency)	3	usaha (business)	3
15	penghasilan (income)	3		
16	pengeluaran (expenditures)	3		

Notes: In the community meetings, attendees were asked to brainstorm characteristics associated with being poor, and then characteristics associated with being rich. Generally attendees shared short phrases. Here we list any words that were mentioned at least 3 times (after extracting common stop words), as well as the number of times each was mentioned in any response. Indonesian words are listed with approximate English translations in parentheses. Cases where the term household "rumah tangga" is mentioned are not counted in the tally of the word "rumah" or house.

Table A22: Most Common Community Meeting Welfare Descriptor Trigrams- Experimental Community Meetings

#	Trigram (Poor)	Freq.	Trigram (Rich)	Freq.
1	Tidak punya sawah (Doesn't have rice fields)	7	Tidak punya hutang (Doesn't have debt)	4
2	Tidak punya rumah (Doesn't have house)	4	Memenuhi kebutuhan sehari-hari (Meet daily needs)	3
3	Tidak punya pekerjaan (Doesn't have job)	4	Penghasilan lebih besar (Bigger income)	3
4	Tidak punya uang (Doesn't have money)	3		
5	Punya pekerjaan tetap (Has permanent job)	3		

Notes: In the community meetings, attendees were asked to brainstorm characteristics associated with being poor, and then characteristics associated with being rich. Generally attendees shared short phrases. Here we list any three-word trigrams that were mentioned at least 3 times, as well as the number of times each was mentioned in any response. Indonesian phrases are listed with approximate English translations in parentheses.

A.2 CRediT Roles

CRediT contributor roles describe how each author contributed to the research output. The CRediT roles for this project are as follows (listed for each in order of greatest to least contribution level, with "/" indicating equal contribution):

1. **Conceptualization:** Trachtman
2. **Data Curation:** Trachtman
3. **Formal Analysis:** Trachtman
4. **Investigation:** Permana/Sahadewo
5. **Methodology:** Trachtman, Permana, Sahadewo
6. **Project Administration:** Sahadewo, Trachtman, Permana
7. **Supervision:** Sahadewo, Permana, Trachtman
8. **Writing – original draft:** Trachtman
9. **Writing – review & editing:** Trachtman, Permana/Sahadewo

A.3 Model of Individual's Transfer Allocation Decision

We now introduce a simple model to help conceptualize the intuition behind why the MUE may be especially relevant in explaining community members' targeting decisions. The key insights are that individuals may indeed try to target transfers to the households that are currently most in need (allowing for the influences of other preferences/objectives that are uncorrelated with others' welfare status), which maps directly to the MUE estimate.

A.3.1 Preliminaries: The Task

Each individual is tasked with creating an individual targeting ranking, where they are asked to rank n community members from poorest to richest. N transfers of fixed amount $\bar{t}$ will be given to the community members ranked as poorest. $N \in \{0, \ldots, n\}$ and is chosen after the individual makes their ranking. Note that we can think of each value of N as a state of the world. The "true" discrete probability distribution $F(N)$, as determined by the program implementers, is unknown to the participant, but individual i has beliefs about the probability of each value of N: $p_i(N)$. Given this, we can think of the exercise of creating a complete ranking as individual i simply choosing which individuals receive transfers under each possible

realization of N. More specifically, given that transfer distribution may occur with some error or there may be uncertainty in the process, we can think of the individual selecting the probability that each individual gets a transfer under each possible realization of N, called $\chi_{ij}(N)$, for a ranking individual i and a ranked family j. For tractability, we suppose that in any given state of the world, all other community members must have some positive probability of receiving the transfer (perhaps by error).

A.3.2 The Individual's Problem

Individuals choose probabilities of others receiving the transfers for each N, such that they maximize the weighted sum of the utilities of expenditures of the households to be ranked, with preference weights $\{\theta_{ij}\}_j$ (for ranking individual i and ranked family j). These preference weights are unrelated to the welfare status of the household, and simply reflect preferring a given household for uncorrelated reasons (like a preference that the transfer be received by another family member). Additionally, without loss of generality, we impose $\sum_j \theta_{ij} = 1 \forall i$ and $\theta_{ij} > 0 \forall i, j$. Additionally, suppose household j has a continuous utility function described by V, which depends on consumption expenditures x_j, a vector of prices P (which is common to everyone in the same community), and a vector of other characteristics of household j, z_j. z_j may contain information about household structure and demographics, household asset holdings, or other information affecting consumption utility.

Given this, the individual's ranking problem can be stated as follows:

$$\max_{\{\chi_{ij}(N)\}_{N,j}} \sum_{j=1}^{n} \theta_{ij} \sum_{N=1}^{n} p_i(N)\{\chi_{ij}(N)V(x_j) + \overline{t}, P, z_j) + (1 - \chi_{ij}(N))V(x_j), P, z_j)\}$$

subject to:

$$\chi_{ij}(N) \geq 0 \ \ \forall j, N \ \ \text{(With Lagrange multipliers } p_i(N)\eta_{jN})$$

$$\sum_{j=1}^{n} \chi_{ij}(N) = N \ \ \forall N \ \ \text{(With Lagrange multipliers } p_i(N)\gamma_N)$$

This yields first order conditions:

$$\theta_{ij} p(N)[V(x_j) + \overline{t}, P, z_j) - V(x_j), P, z_j)] = p_i(N)[\gamma_N - \eta_{jN}]$$

Summing over states of the world N, we get:

$$\theta_{ij}[V(x_j) + \overline{t}, P, z_j) - V(x_j), P, z_j)] = \overline{\gamma} - \overline{\eta}_j$$

Note that if λ_j represents individual family j's marginal utility of the transfer (the parenthetical part of the left side of the above equation), then, as the transfer size approaches zero, we see that the respondent will want to allocate transfers such that $\theta_{ij}\lambda_j$ is equalized across households. Given that the transfers are of discrete size $\bar{t}$, individuals may not be able to do this perfectly, but will make their best attempt. Because the MUE is essentially an estimate of individuals' relative values of $log(\lambda)$, we may expect this to be directly relevant to the community's targeting decisions.

A.4 Special Protocols: COVID-19 Pandemic

In conducting field research during a global pandemic, we were very careful to ensure the safety of our participants and field staff. Notably, we chose not to do this research through a phone survey, as we felt that the exercises would not adapt well to this change in format. Given this, besides standard IRB approval from both UC Berkeley and Universitas Gadjah Mada, we also went through an additional clearance process specifically focused on the COVID safety of the study, through the Vice Chancellor for Research at UC Berkeley.

Some of the specific safety protocols we implemented were as follows:

- We employed a small, relatively local field team (all live within Purworejo Regency), so as not to encourage disease spread from other areas in the country.
- We closely monitored cases and deaths in the study area, and were fully prepared to cease activities should it become unsafe to continue. (Notably, cases were very low in our study area, espeically during the baseline study/community meeting period, despite there being many cases in the United States during this period.)
- Enumerators received PCR tests regularly.
- Participants and enumerators were provided with masks, hand sanitizer, and individually wrapped snacks.
- Households interviews were held on the respondent's outdoor porch when possible, and in a well-ventilated indoor space when not.
- Community meetings were all held in partially open-air spaces with hand-washing opportunities, and participants were encouraged to practice social distancing.