Working Paper Series

#2023-007

Predicting social assistance beneficiaries: On the social welfare damage of data biases

Stephan Dietrich, Daniele Malerba and Franziska Gassmann

Published 27 March 2023

Maastricht Economic and social Research institute on Innovation and Technology (UNU-MERIT)
email: info@merit.unu.edu | website: http://www.merit.unu.edu

Boschstraat 24, 6211 AX Maastricht, The Netherlands
Tel: (31) (43) 388 44 00

UNU-MERIT Working Papers
ISSN 1871-9872

Maastricht Economic and social Research Institute on Innovation and Technology
UNU-MERIT | Maastricht University

UNU-MERIT Working Papers intend to disseminate preliminary results of research carried out at UNU-MERIT to stimulate discussion on the issues raised.

Predicting social assistance beneficiaries: on the social welfare damage of data biases

Stephan Dietrich[a]* , Daniele Malerba[b], Franziska Gassmann[a]

[a]UNU-MERIT, Boschstraat 24, 6211AX, Maastricht, Netherlands; [b]IDOS, Tulpenfeld 6, 53113 Bonn, Germany

Abstract

Targeting error assessments for social transfers commonly rely on accuracy as a performance metric. This process is typically insensitive to the distributional position of incorrectly classified households. In this paper we develop an extended targeting assessment framework for proxy means tests that accounts for societal sensitivity to targeting errors. We use a social welfare framework to weight targeting errors depending on their position in the welfare distribution and for different levels of societal inequality aversion. While this provides a more comprehensive assessment of targeting performance, we show with two case studies that bias in the data, here in the form of label bias and unstable proxy means testing weights, leads to substantial underestimation of welfare losses that disadvantage some groups more than others.

Keywords: Proxy Means Test; Targeting; Cash Transfers; Social Protection; Fair Machine Learning

JEL: I32; I38; H53; O12; C53

We are grateful for the invaluable assistance provided by Alex Hunns throughout the different stages of this project. We would also like to express our appreciation to Niclas Becker and Hannah Oettrich for their valuable contributions in reviewing and mapping cash transfer programs in the research region.

Cash transfer programs, the most common anti-poverty tool in low-and middle-income countries, have expanded massively throughout the last decade (Gentilini et al. 2022). In the context of limited budgets targeting of programs to the poor is often essential, with many programs relying on data driven systems to identify eligible households. Given that household living standards are difficult to measure and verify, beneficiary selection is often based on more or less precise methods aimed at ranking households and individuals from poorest to richest. Proxy means testing (PMT) is a popular tool to identify eligible households based on predicted income or wealth. In this context, policy designers aim to enhance prediction accuracy and the correct identification of beneficiaries to maximize benefit efficiency and impact. This process inherently contains errors; exclusion errors refer to the percentage of intended beneficiaries not reached by the program, and inclusion errors indicate the share of beneficiaries that should not benefit from the program. Recent studies have demonstrated how more flexible machine learning models and novel data sources can reduce targeting errors of such screening systems (Aiken et al. 2022; McBride and Nichols 2018).

However, if societies hold preferences for redistribution, the comparison of targeting accuracy gives an incomplete picture. Hence, performance assessments should give greater weight to targeting errors among poor households than among non-poor households. From a social welfare perspective, giving the transfer to a very poor household has more value than giving the same transfer to a richer household, in which case it matters who is erroneously targeted or excluded. From this perspective, an increase in prediction accuracy can even cause welfare losses if poorer households are falsely classified. Only few papers evaluate and assess targeting errors (Hanna and Olken 2018) and cash transfer programs more generally (Alderman, Behrman, and Tasneem 2019; Barrientos et al. 2022) through a social welfare lens. However, these welfare estimates implicitly assume that predictions are unbiased and targeting errors stochastic.

There is growing evidence of and discussions around biases in algorithmic decision making in the public policy domain that can result in discrimination (Obermeyer et al. 2019; Rambachan et al. 2020). In this regard discrimination implies that members of some groups in society are less likely to benefit from algorithmic decisions than others for reasons unrelated to the targeting criteria, which has led to the discussion of fairness considerations in prediction models (Corbett-Davies and Goel 2018; Kleinberg et al. 2015; Obermeyer et al. 2019). The origin of biases and algorithmic discrimination often lie in the data used to train models and not in the estimators themselves. This can for example be related to measurement errors in the form of biased proxy indicators of the true outcome of interest or non-generalizable data (Mehrabi et al. 2021). Applied to targeting error assessments, these data biases are difficult to observe, and would imply that poor households of algorithmically disadvantaged groups are less likely to be classified as such and instead benefits are allocated to other, comparatively better-off households. This has redistributive consequences and causes welfare losses that usually are unobservable as they are hidden in the data. For PMTs, this would imply welfare losses of targeting errors are underestimated.

In this paper, we formalise the welfare implications of targeting errors through a social welfare weight framework and illustrate how increasing prediction accuracy can cause welfare losses. In a second step, we follow the work of Gazeaud (2020) and McBride and Nichols (2018) building our own PMTs with data from Tanzania and Malawi. We use these case studies to exemplify that actual welfare losses can be driven by systematic measurement errors leading to biased targeting errors. Lastly, we use household size as an illustrative example of the underlying mechanisms that lead to an unequal distribution of welfare losses. We show that reporting bias and the instability of PMT weights is to the disadvantage of smaller households that end up disproportionately more likely to be falsely classified as non-poor and, hence, would not be identified as eligible by the PMT. Our results suggest that welfare assessments

tend to be significantly underestimated and that the analysed data biases account for up to half of targeting error related welfare losses in the two case studies.

The application of PMT in targeting anti-poverty programs has not been without scrutiny, and it faces criticism on multiple fronts. Recent discussions on PMT can be roughly divided into three categories. The first centres on the relation of the targeting process and the trade-off between equity and efficiency in beneficiary selection (Premand and Schnitzer 2021; Hanna and Olken 2018; Brown, Ravallion, and Van de Walle 2018). The second analyzes the problem of the algorithmic process itself and examines the efficacy of the prediction process using new estimators and data sources (Aiken et al. 2022; Aiken et al. 2023; Brown, Ravallion, and Van de Walle 2018; McBride and Nichols 2018). A third strand examines the presence of errors in the measurement of the dependent variable (Gazeaud 2020) or misreporting of PMT variables (Banerjee et al. 2020) - an issue that goes beyond PMT itself but remains a problematic inherent component. Yet, the welfare implications of the compound effects of data biases and the understanding of which groups are disadvantaged by design, remains unexplored. We see three key contributions of our paper to academic and policy discussions surrounding PMTs.

First, the paper seeks to contribute to the discussions about the use of data-driven decision-making systems in the public policy domain. The ever-increasing availability of data is unleashing new opportunities to target policies efficiently to those that need public support the most. Examples include poverty screening tools based on satellite imagery, cell phone or social media data (Aiken et al. 2022; Ayush et al. 2020; Blumenstock 2016; Ledesma et al. 2020). While inarguably a promising development, increasing complexity in systems may also increase the risk that problematic biases in benefit allocation remain unobserved because of black-box procedures. PMT weights are usually deliberately and for good reasons not disclosed (even though people may infer or form beliefs about weights (Banerjee et al. 2020;

Camacho and Conover 2011)). However, it seems legitimate for citizens to demand information on the general targeting procedures. In preparation of this project and to obtain an overview of the scale of opacity in PMT, we reviewed the available information on public cash transfer programs and targeting mechanisms in East Africa, the region of our case studies. In total we identified 10 public cash transfer programs with PMTs (see Annex A). Only half of the programs provide information on the targeting methodology, and we found just one instance of full PMT weights published. In this paper, we illustrate how, usually unobservable biases cause welfare losses that are unequally distributed and harm members of some groups more than others. These results call for a closer scrutiny and more transparency in targeting procedures ("fairness through awareness" (Dwork et al. 2012)), but it also raises the question whether in certain contexts targeting should be regarded as a prediction problem in the first place, or whether other targeting approaches that do not rely on predictions perform better in terms of social welfare if data biases and legitimacy considerations are taken into consideration. This paper does not provide a definite answer to this question as it depends on contextual factors and societal preferences, but the results call for a wider discussion of the application and implications of using PMT systems.

Second and relatedly, the paper touches on the growing fair machine learning literature that has gained momentum in recent years and received inputs from different disciplines. Several papers have found important biases in data used in the public sphere (e.g. policing, law, health) that resulted in and may have even self-enforced discriminatory practices (Obermeyer et al. 2019; Lum and Isaac 2016; Barocas and Selbst 2016). In this paper we regard deviations in the data from the unobserved reality, the ground truth, as bias. Underreporting of assets by households to appear less wealthy would be an example. We consider discrimination as the problematic cases of bias that systematically harm members of some groups more than others. While household size, the illustrative example used here, is not a protected class in the same

manner as race or gender, viewed in the abstract, the problem remains pertinent. A child born into a household has no control over its size (think of orphans). Several papers review sources of discrimination and discuss indicators for algorithmic unfairness (Mehrabi et al. 2021; Gajane and Pechenizkiy 2017; Corbett-Davies and Goel 2018; Rambachan et al. 2020; Ferrer et al. 2021). If biases are hardcoded in the data, statistical indicators derived from this data fail to detect true imbalances. In this paper, we show that welfare losses due to targeting errors are substantially underestimated because of biases in the data. For us these results indicate that opacity in procedures and a purely data driven examination bears important risks possibly obfuscating discriminatory practices. This paper seeks to contribute to the discussion of fairness considerations in social protection systems.

Third, the paper links the use of social welfare weights with targeting issues in cash transfer programs in low and middle-income countries. When assessing social public policies, a welfarist approach is usually implemented where the social planner (government) aims to maximise a social welfare function (Saez and Stantcheva 2016; Sen 1977). The social planner uses welfare weights reflecting the fact that the welfare increases of the worse off are more important in terms of social welfare than those of the better off. Nonetheless, cash transfer programs are still routinely evaluated by looking at the change in the outcomes of interest, such as poverty or mean consumption; this means in practice using a utilitarian approach, as it does not explicitly attach higher importance to improvements in the outcomes and productive capacity among low-income groups (Barrientos et al. 2022). In addition, current studies do not analyse issues of biases and discrimination, despite the aim of these policies that is usually to reach the ultra-poor and marginalized (Coady, D'Angelo, and Evans 2020; Creedy 2006).

The remainder of the paper is organized as follows: we first outline the social welfare weight framework. Thereafter, we introduce the data for the two case studies, Malawi and Tanzania,

that we were previously also used by McBride and Nichols (2018) and Gazeaud (2020). This is followed by a description of prediction models and prediction results. Thereafter, we apply the social welfare weight framework to assess welfare losses due to targeting errors and explore two case studies to discuss how measurement error induced biases can cause welfare losses and unequal distribution of these losses. In the last section, we discuss our findings.

Targeting Errors and Social Welfare

The demand for anti-poverty programs may be indefinite and exceed the available government resources, which implies a need for targeting. From a purely theoretical perspective and given a limited budget, targeting the poor is the most efficient option to reduce poverty. Yet, the trade-offs between targeting costs and efficiency are well-known in the literature. The policy maker needs to choose a method to identify beneficiaries while having imperfect information on household living standards, which limits her ability to correctly rank individuals from poorest to richest.

Both vertical and horizontal inefficiencies can reduce the impact of public spending (Atkinson, 2005). Vertical efficiency is concerned with the targeting accuracy (only the target group is treated) and horizontal efficiency reflects the program comprehensiveness (all of the target group is treated). In the context of anti-poverty programs, notions of efficiency depend on the way poverty is measured, or how the policy objectives are set. Assuming that the primary objective of a transfer is to alleviate poverty, i.e. bring everybody up to a specified poverty line, then program efficiency is the extent to which the poverty gap is reduced given the available budget. If more weight is attached to those farthest from the poverty line, then targeting the poorest first is more efficient. This is reflected in the parameter α of the standard Foster, Greer, and Thorbecke (1984) class of poverty measures $P_\alpha = (1/n)\sum_i^q [z - y_i)/z]^\alpha$, where values of $\alpha > 1$ assign more weight to larger poverty gaps. If α approaches infinity,

only the poverty gap of the poorest person matters. For values of $0 \geq \alpha > 1$, the most efficient programme reduces the poverty headcount rate and assigns transfers to those close to the poverty line. The sharpness of the poverty reduction objective also affects the assessment of targeting efficiency. With a sharp objective, the marginal value of a transfer assigned to a non-poor is equal to zero. Yet, wider objectives might be concerned with the near-poor as well and still assign some efficiency to the transfer. Within a certain range of the poverty line, the marginal value would still be positive, but lower than one.

A PMT is a common way of ranking and identifying households in need. It predicts household wealth based on a set of easily verifiable household characteristics. As the scores are only a proximate measure of actual living standards, these predictions result in targeting errors, hence reducing both vertical and horizontal efficiency of the allocated budget.

Social Welfare Weights

Social welfare functions provide a framework for the evaluation of the benefits and costs of social programs and policies (Adler 2019). Within this framework, welfare weights link the preferences for redistribution of a society to social welfare through an inequality aversion parameter; in this sense, the inequality aversion parameter shows how strongly the population (represented by a social planner) prefers a more equal society compared to a (on average) richer one. One commonly used social welfare function is the Atkinson (1970) constant elasticity social welfare function of the following form:

$$SWF = \begin{cases} \frac{\sum_{i=1}^{n} Y^{1-p}}{1-p} \\ \sum_{i=1}^{n} \log(Y) \quad if \;\; p = 1 \end{cases} \tag{1}$$

where Y is household i's per-capita welfare, and ρ is the inequality aversion parameter, where higher values of ρ put higher weights on the welfare of the very poor.[1]

This welfare function is individualistic and additive. It also satisfies the 'transfers principle', meaning that a welfare transfer from a richer to a poorer person, which does not affect their relative positions, represents an improvement in social welfare (Sen 1976).

Welfare weights can be derived from equation (1). In fact, if we take the derivative of equation (1) for two individuals, individuals a and b, we have that a change in social welfare (w) arising from a transfer to individual b compared to the change in social welfare derived from the same transfer to individual a is:

$$-\frac{dy_a}{dy_b}|_W = \left(\frac{y_b}{y_a}\right)^p = \beta_b \qquad (2)$$

The weight β_b represents therefore the increase of social welfare arising from a transfer to individual b, relative to the situation of giving the same transfer to another individual (in this case individual a). The use of a reference individual means that we are calculating normalized welfare weights. In our setting, the reference point is the poverty line so that a household above this line is weighted with a lower weight than a household below the line.[2] In addition, it also follows that a change in social welfare is given not only by the welfare weight but also by the size of the transfer. In fact, social welfare can increase by the same amount in the

[1] Atkinson measured inequality in terms of the proportional difference between two income values. These are the arithmetic mean income, and the income level, called the 'equally distributed equivalent' income, which, if obtained by everyone, produces the same value of 'social welfare' as the actual distribution.
The utilitarian welfare function is parameterized with one parameter that controls for intratemporal inequality aversion but also risk aversion (Cooke et al. 2009). A welfare function of this kind forces one to use the same value for both concepts. The inequality aversion parameter is similar to a risk-aversion parameter in an expected-utility framework capturing the trade-off between higher expected payoffs and the uncertainty of those payoffs.

[2] The literature usually uses the median consumption or mean consumption as reference; (Kind, Wouter Botzen, and Aerts 2017; Van der Pol, Bos, and Romijn 2017), but in this setting the poverty lines is a more suitable benchmark. As we are looking at relative social welfare changes compared to a perfect targeting benchmark, the choice of the benchmark has no implications for the results in this paper.

following two cases: 1) if a small transfer is given to a household with high welfare weight; 2) if a big transfer is given to a household with a small welfare weight.[3]

An important factor is to correctly estimate the inequality aversion parameter. This parameter originates from the equality-efficiency trade-off that was initiated by Okun (1976). A parameter equal to zero means that there is no inequality aversion, and societies prefer to be richer. There are many ways in which to estimate the inequality aversion parameter (Del Campo, Anthoff, and Kornek 2021). Most studies try to reveal the inequality aversion parameter through hypothetical (using experiments) or actual data (using tax data). In this paper we use a range of parameters that have been estimated for lower income countries (Barrientos et al., 2022). Once the social welfare weight is calculated, we can measure the impact of a transfer on social welfare and compute welfare losses due to targeting errors.

Targeting Errors, Bias, and Welfare Loss

PMT targeting is based on predictions and by default targeting assessments implicitly assume unbiased data (an exception is Gazeaud (2020)). However, there is growing evidence of and discussions around biases in algorithmic decision making in the public policy domain that can result in discrimination (Obermeyer et al. 2019; Rambachan et al. 2020). In this paper, we assess the extent to which -usually unobservable- biases cause welfare losses. Thereby we regard biases as deviations of the observed data from the unobservable truth that is systematically related to group affiliations. This leads to an unequal distribution of welfare losses between group, which increase social welfare losses the greater societal preferences for redistribution.

[3] Alternatively, this can be represented by putting the benefits (b is the benefits per capita for household i) directly in the welfare function (Hanna and Olken, 2018): $SWF = \frac{\sum_{i=1}^{n}(y+b)^{1-p}}{1-p}$

To formalize this, we apply the framework described in Rambachan et al. (2020) to predicting consumption poverty $\tilde{Y}$ in time period t=1 with parameters trained with data collected in time period t=0:

$$\tilde{Y}_{t=1} = Y^*_{t=0} + \Delta y + \Delta\vartheta + \varepsilon \quad (3)$$

where consumption poverty is approximated with survey data on reported consumption per capita Y^* in time period t=0, which differs by Δy from ground truth consumption poverty $\tilde{Y}$ in t=0 and by $\Delta\vartheta$ which denotes the change in consumption poverty $\tilde{Y}$ between t=0 and t=1, and the estimation error ε. In our framework, differences in predicted consumption poverty $\hat{E}[Y^*_{t=0}]$ between two groups $G \in [1,2]$ may originate from four sources:

- **Base rate difference**; refers to a different prevalence of consumption poverty between groups, and are thus reflecting true discrepancies in the outcome of interest:
  $E[Y^*_{t=0}|G=1] - E[Y^*_{t=0}|G=2]$
- **Label bias**; systematic error in proxy for consumption poverty:
  $E[\Delta y|G=1] - E[\Delta y|G=2]$
- **Stability**; systematic difference in prediction errors related to the timing of prediction:
  $E[\Delta\vartheta|G=1] - E[\Delta\vartheta|G=2]$
- **Estimation error**; bias introduced by algorithms putting more weight on predictors favoring one group over the other:
  $\hat{E}[\varepsilon|G=1] - \hat{E}[\varepsilon|G=2]$

If the distributions of Δy and $\Delta\vartheta$ respectively are identical between both groups, measurement errors are captured by the estimation error. If this is not the case, measurement errors distort predictions to the disadvantage of one group, $E[\tilde{Y} - Y^*|G=1] \neq E[\tilde{Y} - Y^*|G=2]$. As a result, the welfare ranking using $\tilde{Y}$ can differ from Y^*, where the disadvantaged group receives on average a higher ranking than it should according to Y^*.

Let's assume the before mentioned individuals a and b are part of group 1 and 2 and $Y_a^* = Y_b^*$, but predicted welfare levels are different ($E\left[\tilde{Y}_{t=0}\middle|G_a = 1\right] < E\left[\tilde{Y}_{t=0}\middle|G_b = 2\right]$) because of measurement errors. As measurement errors are unobserved, the predicted social welfare change of a transfer to b instead of a would be $\left(\frac{Y^* + \Delta y_b + \Delta\vartheta_b}{Y^* + \Delta y_a + \Delta\vartheta_{ba}}\right)^p$ if $\rho \neq 1$ even though ground truth social welfare changes are the same. Taking the derivatives with respect to Y^*, $\Delta y_b, \text{and } \Delta\vartheta_b$ suggests that social welfare losses increase with the size of the relative difference in measurement errors between both groups. This is amplified the lower Y^* and the higher the aversion for inequality ρ is. From this, we derive three propositions regrading PMT assessments that we want to highlight in this paper:

1. For a given inequality aversion parameter, the social welfare loss depends on the transfer size and exclusion errors.
2. A reduction in estimation errors of $\tilde{Y}_{t=0}$ is not sufficient to improve w. In fact, following equation (3), if $\Delta y = 0$ there could be still large $\Delta\vartheta$ and such systematic measurement error can cause unobserved social welfare losses.
3. Welfare loss inequality increases the stronger the bias and the poorer the disadvantaged group.

Data

We examine PMTs that we build with experimental data from Tanzania and Malawi. The data have been used in previous work on PMTs (Gazeaud 2020; McBride and Nichols 2018), which has the advantage that we can benchmark our results against their findings and that we can rely on a predefined set of PMT variables. In addition, both hypothetical case studies provide interesting facets that allow us to examine label bias and the instability of PMT weights.

Malawi

The 2004/5 Second Integrated Household Survey consists of 11,280 households. The survey was conducted over the course of 12 months, where enumerators in randomly selected areas interviewed one enumeration area per month.[4] We make use of the staggered data collection to examine how the timing of the data collection influences screening outcomes. Data collection was carried out during both the lean (October-March) and harvest season (April-September). The share of interviews conducted during both periods is balanced, and we observe no significant differences in time-invariant household characteristics between households surveyed in both periods. We refer to Annex A for summary statistics of all PMT variables, the same as applied in McBride and Nichols (2018).

In Table 1 we summarize poverty headcount by smaller and larger households and data collection season. As a first result, 65% reported consumption per capita below the consumption poverty line, and poverty increases from 61% in the harvest period to 69% in the lean period. Secondly, we define smaller households with the median household size, which reaches 4 members in the data (mean is 4.5). The prevalence of poverty is higher among larger households (80%) compared to smaller households (52%). For smaller households, poverty increases by about 11 percentage points (pp) between harvest and the lean season while for larger households this change is only 7pp which implies that the relative change between seasons is more important for smaller households.

Tanzania

Data were drawn from the Survey of Household Welfare and Labour in Tanzania (SHWALITA) project. The project experimentally tested and compared the consistency of

[4] https://microdata.worldbank.org/index.php/catalog/2307/related-materials

consumption reports using different household survey modules. The project delivered a survey to all 4,029 households with a consumption experiment that took the form of eight different consumption questionnaire treatments that were randomly assigned to approximately 500 households each. The eight treatments varied the approach (recall or diary) and the duration of recall (between 7 days and 12 months). These modules included: (i) long list of items and 14 day recall (ii) long list of items and 7 day recall (iii) subset list 7 day recall (iv) collapsed list 7 day recall (v) long list of items monthly recall (vi) 14 day household diary frequent visits (vii) 14 day household diary infrequent visits (viii) 14 day personal diary frequent visits. For ease of presentation, we group treatments into diary and recall modules but also ran all results separately by treatment. The experiment was delivered to seven districts across Tanzania from September 2007 to August 2008. The multi-stage sampling strategy saw villages selected using a probability-proportional-to-size, with sub-villages selected at random. Within sub-villages, three households per sub-village were randomly assigned to one of the eight modules. The results of the experiment permit comment on the severity of the problem posed by non-random error in measurement of consumption introduced by – inter alia – recall error, telescoping, rule of thumb measures, and personal leave-out error (Beegle et al. 2012; Caeyers, Chalmers, and De Weerdt 2012).

In the analysis, we use the same PMT input variables as Gazeaud (2020) and refer to Annex B for a list with summary statistics. In Table 1 we summarize poverty headcount by smaller and larger households and whether recall or diary modules were used to collect consumption data. Overall, 41% of households reported per capita consumption below the $1.25/day poverty line. But when considering diary consumption, the figure is 37% in the diary data compared to 44% in the recall data; this follows the fact that consumption from diaries is higher as recall values tend to underestimate true consumption. This discrepancy is larger

among larger households (more than 5 household member; 8pp) compared with smaller households (5 or less household member; 3pp).

Among smaller households, only 28% fall below the poverty line which increases to 59% among larger households. This changes markedly depending which consumption module is used: recall-based modules lead to higher levels of reported poverty with a gap of 7pp compared with diary-based modules. Interestingly this is mainly driven by larger households where the gap between recall and diary modules increases to 12pp. This is a substantial difference presumably driven by the fact that the responding household members are less aware of all consumption activities in larger households leading to underreporting.

Table 1 Poverty headcount in Malawi and Tanzania data

Poverty	**Malawi**		**Tanzania**	
All	65% (0.45)		41% (0.78)	
Smaller HH	52% (0.55)		28% (0.94)	
Larger HH	80% (0.64)		59% (1.18)	
	lean	**harvest**	**recall**	**diary**
All	69% (0.62)	61% (0.65)	37% (1.24)	44% (0.99)
Smaller HH	57% (0.88)	46% (0.92)	26% (1.5)	29% (1.2)
Larger HH	84% (0.74)	77% (0.8)	64% (1.47)	52% (1.94)

Note: Standard errors in parentheses. *Lean* and *harvest* refer to period of the year data was collected. *Recall* and *diary* refer to the consumption data collection module. (n=11280 in Malawi; n=4032 in Tanzania)

Prediction model and targeting errors

In line with current PMT practices, we first build models to predict household consumption and then classify poor households based on predicted consumption. We use this approach to

mimic current PMT procedures, even though training models directly to classify poor households would be a more straight-forward approach to predict the outcome of interest (poverty). We tested a wide range of specifications of different model classes but focus the discussion on a simple linear model as benchmark and a gradient boosting model that performed best in this application. In the linear regression model, we use all standardized PMT variables as inputs. Thereafter, we use the more flexible and efficient xgboost library to train a gradient boosting model. A description of the parameter tuning process is provided in Annex C.

In practice PMT scores are often still estimated and validated with the same data, which bears the risk of overfitting the model to the data at hand resulting in poor out-of-sample predictions. To reduce the risk of overfitting, we randomly draw training data ($N*0.8$) to train the models and a test data ($N*0.2$) that we hold back to compare model predictions. We select our preferred model specifications by comparing how much of the variation in consumption (using R2 as performance metric) is explained by the model using 10-fold cross-validation. Based on the predictions of the best performing model, we classify households in the test data as poor if their predicted consumption is below the poverty line.

Figure 1 displays the prediction results for the test data. For ease of presentation, we show the prediction results graphically using multidimensional scaling to present the similarity of households in PMT variables in a two-dimensional space. That means, households with similar values in the PMT variables are close to each other in the scatter plot where identical households in PMT variables would be overlapping. The colour of the markers shows whether a household is classified poor (orange) or non-poor (blue). In the first row of Figure 1 we show actually poor households and rows 2 and 3 show predicted poor households according

to the linear and xgboost models. The first column refers to results for Malawi and the second column to Tanzania.

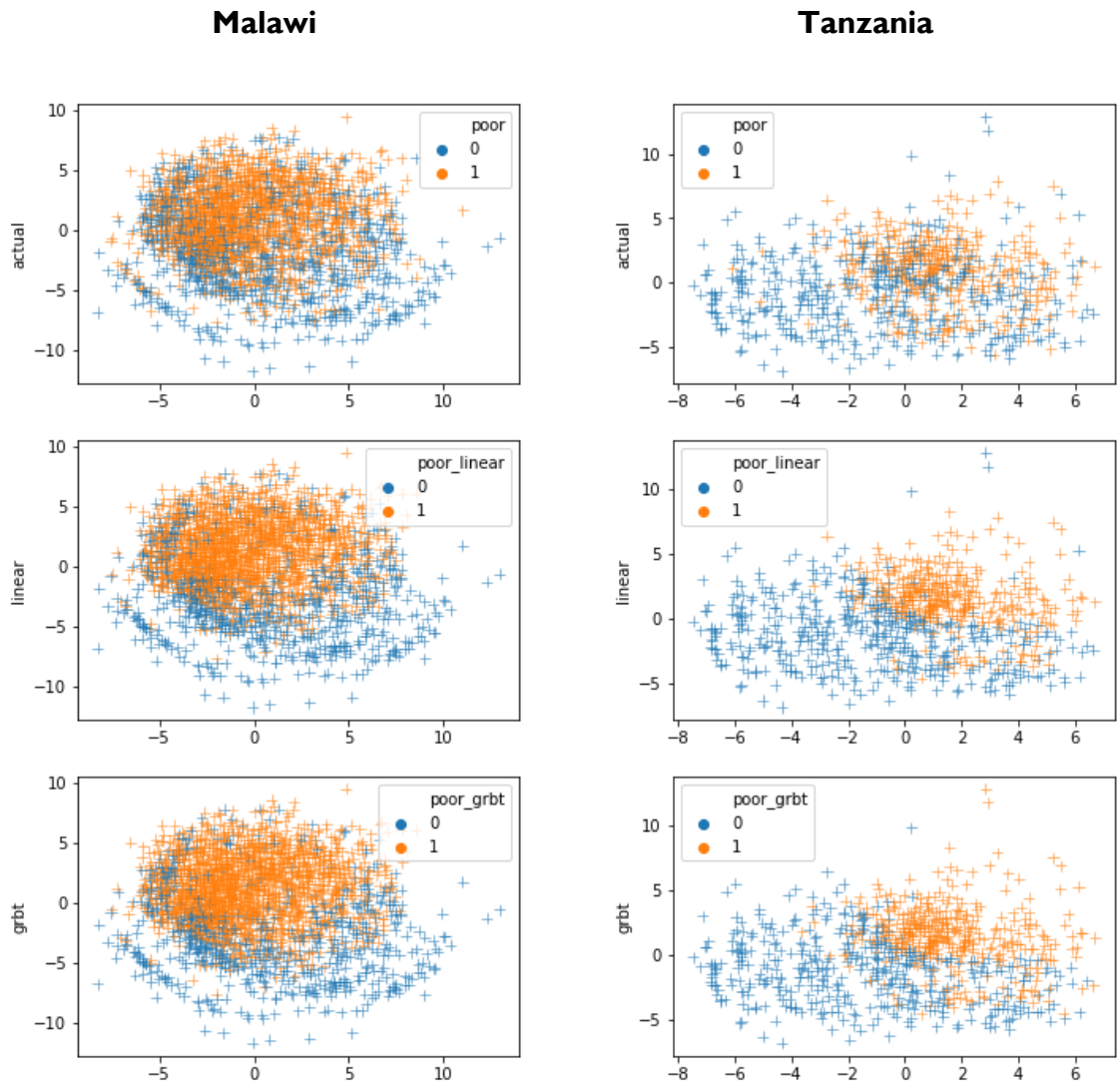
The results suggest that poverty is quite dispersed with respect to PMT variables and there are many households that differ in their poverty status, but that are very similar in PMT variable values (see first row). This illustrates the difficulty to distinguish poor from non-poor households with the available PMT variables as input. As expected, the xgboost model performs better than the linear model in explaining variation in consumption reaching an R2 of 0.62 in the test data compared to 0.58 of the linear model in Malawi. With regards to the classification of poor households, about 81% and 80% of households in Malawi are correctly classified by the xgboost and linear model. In Tanzania, the R2 is lower and reaches 0.56 and 0.54 with xgboost and linear models resulting in 75% correct classifications in both cases.

Despite a seemingly high prediction accuracy, about 7% of poor households are not classified as such (exclusion error) and 13% of non-poor households are classified as poor (inclusion error) for the Malawi data and 15% inclusion and 11% exclusion error for Tanzania, which is similar for both model classes (an overview of classification results for the xgboost model is provided in Tables 3 and 4). Even though the aggregate performance measures for the xgboost and linear model are quite similar, the classification of 2.2% and 3.6% of households in Tanzania and Malawi, respectively, changes depending on whether the xgboost or linear model is used.

To summarize, the xgboost model performs slightly better than the linear model in predicting consumption. The relatively small performance difference in classifying poor households could be related to the list of pre-selected input variables that perform well with linear models and because the model was not specifically trained to classify poor households but to predict consumption. In both cases, predictions are more clustered than the distribution of ground-truth poverty and imply that targeting errors are more likely to occur for certain PMT variable

combinations. In the following we compute welfare losses and examine to what extent they are driven by data biases.

Figure 1 Actual (1st row) and predicted (2nd and 3rd row) poor households



Notes: x and y axis show two-dimensional test data of PMT variables: first PMT variables were standardized and thereafter re-scaled using multidimensional scaling. Actual refers to true (consumption) poverty status and linear and grbt refer to predicted poverty status with linear and xgboost model.

Social Welfare Loss

For the social welfare loss assessments, we use the test data to simulate an anti-poverty policy for which we allocate a fixed budget to households that are predicted poor according to the

PMTs.[5] Following common practice, the transfer size is the same for all beneficiaries and transfer amounts distributed by each PMT (linear, xgboost or perfect targeting) are adjusted to the fixed budget. That means, if a model overpredicts poverty, the transfer size per beneficiary will be smaller compared to a well calibrated model. This is important as usually targeting assessments do not consider the budget implications of overpredictions or the cost at which a certain prediction accuracy is achieved.

Welfare losses due to targeting errors

Figure 2 shows the marginal welfare loss of a transfer allocated through either the linear or xgboost model compared to a perfect targeting benchmark. In the simulations the fixed transfer budget is distributed through our algorithms to (predicted) poor households and compared to the benchmark scenario in which transfers are allocated to all (actually) poor households without targeting errors. In the perfect targeting benchmark case, each poor households receives a 1-unit transfer.[6]

We compute the welfare loss assuming different levels of inequality aversion. If societies hold no preference for redistribution (ρ =0), there is no welfare loss of targeting errors according to our framework. In fact, whether a rich or a poor person receives a transfer does not impact social welfare in the absence of inequality aversion. With increasing inequality aversion, the welfare loss rises because societies weight the exclusion of poorer households more strongly than inclusion errors. As a consequence, the same prediction accuracy can result in different welfare losses depending on where in the welfare distribution errors occurred.

[5] The fixed budget is defined as a unit transfer to all actually poor households in the analysis.
[6] The poverty lines used here are 1910 Kwacha in Malawi and 208147 Schilling in Tanzania.

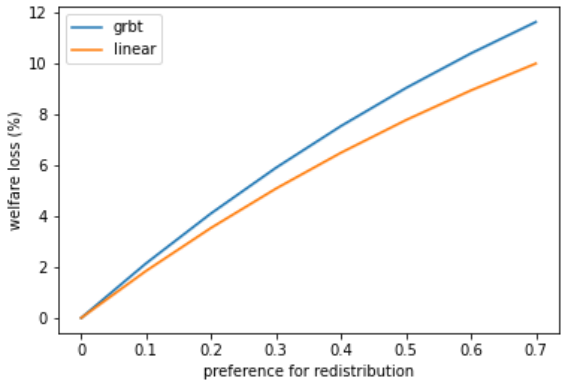
Not surprisingly, the results suggest that welfare losses are larger in Tanzania than in Malawi because of prediction accuracy differences. Yet surprisingly, in Malawi the welfare loss with the xgboost model is larger than with the linear model despite a slightly higher classification accuracy of the former model. That is, the xgboost model is better in predicting consumption and more accurate in classifying poor households but welfare losses are higher than with the linear model. This means that the linear model leads to lower welfare losses and outperforms the xgboost model in this setting if positive welfare weights are used. Thereby welfare losses are a function of the of the distortionary biases in the allocation of benefits and the transfer size that is defined by the fixed budget and the number of predicted beneficiaries. The role of transfer size is usually not regarded in targeting performance assessments, but it has strong effects on the levels of redistribution and therefore welfare losses. If no budget constrains was imposed i.e. if always the same 1-unit transfer would be allocated in all scenarios, the xgboost model would actually perform better that the linear model in social welfare terms. By imposing the budget constraint we consider the cost of obtaining a certain level of targeting accuracy.

A classification model that is trained to reduce exclusion errors (for example using area under the precision-recall curve as performance metric) leads to lower welfare losses particularly at higher levels of inequality aversion and in principle the social welfare function could also be directly coded into the algorithm and training process, but this would imply that it needs to be optimized for a specific level of inequality aversion that is pertinent to the ex ante societal rating of the ex post model outcome.

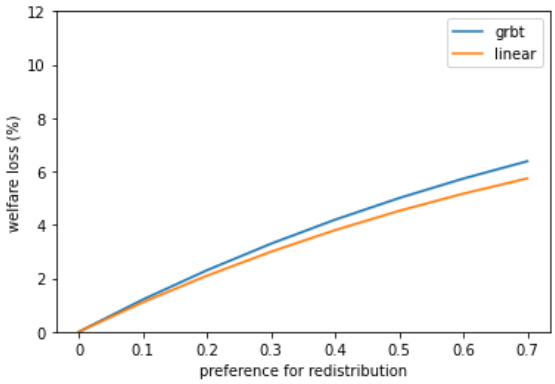
These findings suggest that the evaluation of prediction in welfare terms matters and, as in the case of Malawi, changes the selection of the preferred model. Next, we turn our attention to the distribution of welfare losses and particularly examine how label bias and unstable predictors lead to systematic underestimation of welfare losses.

Figure 2 Marginal welfare loss of unit transfer with linear and xgboost model

Tanzania



Malawi



Notes: Welfare loss computed as percentage change in welfare in comparison to perfect targeting assuming different levels of inequality aversion.

Welfare loss due to measurement errors

As outlined in the conceptual framework, we regard measurement error and temporal instability of PMT weights as drivers of welfare losses. In this section we assess the extent to which actual welfare losses can be explained by these factors. With the experimental data from Tanzania, we consider welfare losses due to label bias that we examine with consumption data collected with recall versus diary consumption measurement methods. Therefore, we train two separate PMT models each using only recall or only diary consumption data. Thereafter, we validate both models using the same test data i.e. we classify the same households using PMTs built with diary or recall data. Differences in classification between PMTs can thus be attributed to differences in PMT weights. Similarly, with the staggered data collection in Malawi, we consider welfare losses due to PMT instability by applying two algorithms either trained with lean or harvest season data to the same test data. In addition, and to explain the underlying mechanisms, we show that our PMTs disadvantage smaller households (see Annex C for an overview of feature importance and coefficients in the models).

Table 2 shows the different training data used to build the PMT pairs and the test data to validate these models. What is important to highlight is that we always use the same households to validate and compare the PMT pairs. For consumption measurement we rely on diary test data as benchmark and for the seasonal stability we use harvest season data as benchmark. We subsequently focus on the xgboost model only, but results hold qualitatively when using the linear model.

Table 2 PMT and test data used for assessment comparisons

	PMT dependent variable		**Test data**
	PMT 1	***PMT 2***	
Tanzania	*Recall*		*Diary*
		Diary	
Malawi	*Lean*		*Harvest*
		Harvest	

Note: lean and harvest refer to period of the year data was collected. Recall and diary refer to consumption data collection module. Test data refer to 20% of randomly selected data used for validation. Diary and Harvest only use the subset of the test data in in which consumption diaries were used or data were collected during the harvest period.

Label Bias

Household consumption reports are often treated as unbiased proxy for true consumption. However, recall bias is known to distort consumption reports and has been found to be more pronounced in larger households where single respondents may less accurately report consumption of other household members (Beegle et al. 2012; Gibson and Kim 2007). To assess the extent to which this can distort poverty screening decisions, we explore the experimental data collected in Tanzania similar to Gazeaud (2020). In the main analysis we distinguish between data collected using consumption diaries and recall methods. We consider consumption diaries as a more accurate measurement approach where individual

consumption diaries with a high supervision frequency have been regarded the gold standard approach (Beegle et al. 2012; Caeyers, Chalmers, and De Weerdt 2012; Gazeaud 2020). We build two PMT models each trained exclusively with recall and diary consumption data and thereafter rely on diary consumption test data to validate and compare both PMTs. The difference in welfare losses between both PMTs is an indicator of the welfare loss caused by consumption measurement error.

Table 3 shows the confusion matrix for the full model and the results using separate PMTs for recall and diary data. Households are more likely to be predicted poor if the PMT is estimated with recall than diary data. The predicted poverty rate in the (same) test data is 53% versus 39% using the recall and diary PMT respectively, which presumably is related to underreporting in recall modules (Beegle et al. 2012; Caeyers, Chalmers, and De Weerdt 2012; Gazeaud 2020). As a result, many non-poor households were classified as poor by the recall PMT or in other words, there are many inclusion errors. Exclusion errors are also lower with the recall PMT compared to the diary PMT. Overall, the accuracy of the diary and recall PMTs are quite similar with 72% and 70% respectively, however, these indicators do not consider the distributive consequences and resulting welfare losses.

Table 3 Confusion Matrix Tanzania

		Predicted poor	Predicted non-poor
All (diary+recall)	Poor	261	89
(n=805)	Non-poor	120	335
Diary PMT Model	Poor	75	43
(n=300)	Non-poor	42	140
Recall PMT Model	Poor	93	25
(n=300)	Non-poor	66	116

Note: All predictions based on xgboost model. *All* refers to model trained and validated with mix of recall and diary consumption data. *Diary* and *Recall Model* refer to models trained exclusively with diary

and recall data, both evaluated with diary test data. *Smaller* and *Larger* refer to subset of the latter distinguished by the median household size. n refers to sample size of test data.

The left side of Figure 3 displays simulated welfare losses for both PMT estimates. It shows that welfare losses are considerably higher for the model trained with recall data. The difference reaches 5pp which accounts for almost 30% of the welfare loss. That means, with a PMT trained with recall data, we underestimate true welfare losses by about one third. If we only use data of the diary treatment with a higher supervision frequency, the gold-standard in the literature, the gap further increases accounting for almost half of the welfare loss.[7]

Why is the welfare loss so heavily underestimated? The accuracy of both models is similar, but the recall PMT overpredicts poverty meaning that the transfer size is smaller leading to lower levels of redistribution. As a result, the selected beneficiaries under the diary PMT receive larger transfers and as they tend to be more likely to be allocated to the poorest, welfare losses with the diary PMT are lower with increasing preference for redistribution.

In the right panel of Figure 3, we break the results further down into smaller and larger households (defined by the median household size). As already seen, welfare losses are larger for the PMT trained with recall data. Yet while this difference reaches about 8pp for smaller households, with less than 3pp the effect is smaller for larger households. That means that consumption measurement error leads to significant underestimation of about 40% of welfare losses among smaller households. This share of underestimations is more than double compared with larger households, which implies that the welfare losses induced by

[7] Note that there are only 105 observations in the test data that were collected with the gold-standard approach. For simplicity, we group all diary and recall treatments and do not consider differences between those treatments in the main analysis. For more details about the effects of the treatments we refer to the original articles (Beegle et al. 2012; Caeyers, Chalmers, and De Weerdt 2012).

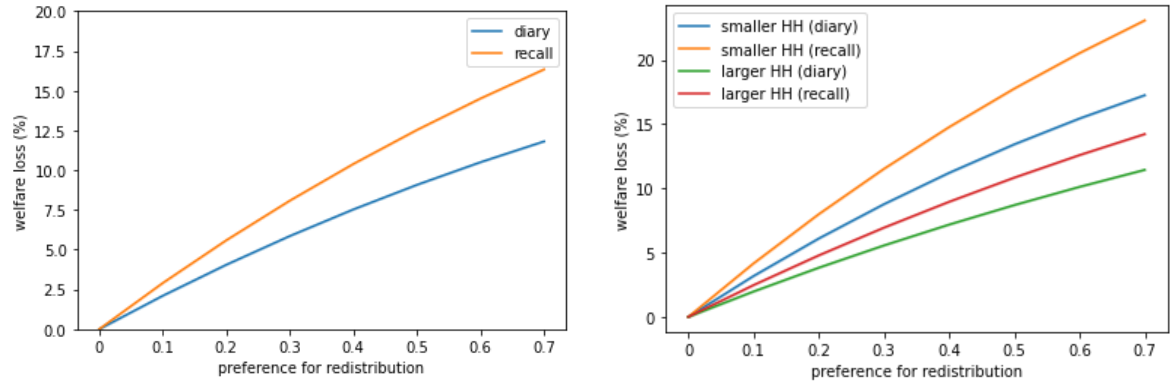
consumption measurement error are predominantly driven by distortive effects to the disadvantage of smaller households.

Why are smaller households more sensitive to consumption measurement error? Following the original results of the data experiment, recall bias is more pronounced in larger households leading to an underestimation of true consumption. This bias is transmitted to the PMT, meaning that predictors associated with household size are underestimated which leads to social welfare losses as smaller households with the same (ground truth) consumption level are less likely to be selected by the PMT.

On a higher level, this relates to the problem of using household level information (household consumption) as proxy for individual welfare (poverty). Implicitly it is assumed that all household members are either poor or not, regardless of the actual distribution of resources within households and possible economies of scale resulting from sharing goods. This debate is long-standing, and in practice using consumption per capita has established as the standard metric for poverty assessments which allows for comparability of results. However, alternatives that account for economies of scales in larger households exist and can potentially have important distributional implications (Jolliffe and Tetteh-Baah 2022). To illustrate this point, we examine the sensitivity of our result to using constant-elasticity scale adjustment of household consumption as exemplified in Jolliffe and Tetteh-Baah (2022). That is, we divide total consumption by the square root of household size instead of only using household size (i.e. per capita). To simulate the resulting welfare losses, we adjust the poverty line, meaning that the poverty rate and the budget of hypothetical policy remains the same, but the weights to allocate the benefits change. The results suggest that using this alternative approach even further increases welfare losses due to the recall bias among smaller households accounting for 80% of the welfare loss. For more details, we refer to Annex C.

It is beyond the scope of this paper to discuss whether per capita consumption is a better or worse approximation of welfare. This is essentially an empirical question as the magnitudes of economies of scale are highly context-dependent. However, our analysis illustrates that the a priori not evident choice has important implications for allocation distortions and resulting welfare losses that usually remain unobserved in targeting assessments and the comparisons of targeting approaches.

Figure 3 Welfare loss depending on consumption measurement module and by household size



Notes: Welfare loss computed as percentage change in welfare in comparison to perfect targeting assuming different levels of inequality aversion. Evaluation based on the same test data, but models were trained with data only inluding recall or diary data.

Data collection season

Stable predictors ensure that screening outcomes (and resulting errors) are robust to the exact timing of the screening. Often PMT weights are applied to data collected in a different period of the year than the training data (or different years see Brown, Ravallion, and Van de Walle (2018)). We also know that in many settings in which PMTs are applied, household consumption varies substantially over the course of a year (Hopper 2020). Using relatively stable household characteristics to predict a volatile target variable leads to variance in errors. To understand the welfare implications of that, we explore the staggered data collection of the Malawi data by using a similar strategy as in the previous section. We build two separate

PMTs with lean and harvest period data and validate both PMTs using the same harvest season test data. We use harvest season test data to have a "pure" validation set but despite the case of consumption modules in the previous section, there is no clear reason as to why lean season data would be superior to harvest season test data. Regardless of which test data is used, the narrative remains the same.

Table 4 provides an overview of the prediction accuracy, inclusion, and exclusion errors of the different PMTs. The model trained with lean season data overpredicts true poverty by 16pp. As a result, the simulated transfer size per predicted poor household is 4% larger with the harvest PMT than with the lean season PMT. The level of accuracy is slightly higher with the harvest season PMT (80%) compared with the harvest season PMT (77%) mainly because inclusion errors are lower with the harvest season PMT.

Table 4 Malawi Confusion Matrix

		Predicted poor	Predicted non-poor
All (n=2255)	Poor	1279	157
	Non-poor	289	530
Harvest Model (n=1125)	Poor	586	90
	Non-poor	133	316
Lean Model (n=1125)	Poor	633	43
	Non-poor	222	227

Note: All predictions based on xgboost model. *All* refers to model trained and validated with mix of lean and harvest season consumption data. *Harvest* and *Lean Model* refer to models trained exclusively with harvest and lean season data, both evaluated with harvest test data.

To understand the resulting welfare implications, the left side of Figure 4 displays simulated welfare losses for both PMTs. In contrast to the accuracy of classifications, it shows that welfare losses are significantly higher for the lean season PMT and would suggest that the harvest PMT is preferable in this setting. That means, assuming a preference for redistribution

of 0.7, the lean season PMT underestimates the welfare loss if applied in the harvest period by almost 5pp which accounts for about half of the welfare loss.

The main driver of this results is the transfer size. Predicted poverty is 12pp higher with the leans season PMT than with the harvest season PMT. That means that transfers size is substantially lower (15% less). Even though there are less exclusion errors, the harvest season PMT allocates larger amounts to extremely poor households (more than under perfect targeting) which leads to more redistribution than with the lean season PMT. The stronger these transfers are weighted, the larger the welfare loss difference between both PMTs.

The seasonal dynamics of poverty are more pronounced among smaller households. Therefore, the right column of Figure 4 shows that the largest relative welfare loss is carried by smaller households when using the lean season PMT. With the harvest PMT there is no substantial difference in welfare losses between smaller and larger households. If the lean season PMT is applied in the harvest period, it would underestimate 'true' welfare losses by up to 5pp at a preference of redistribution of 0.7. Thereby smaller households bear more than 2/3 of that burden (the lean-harvest PMT gap is more than twice as large for smaller households).

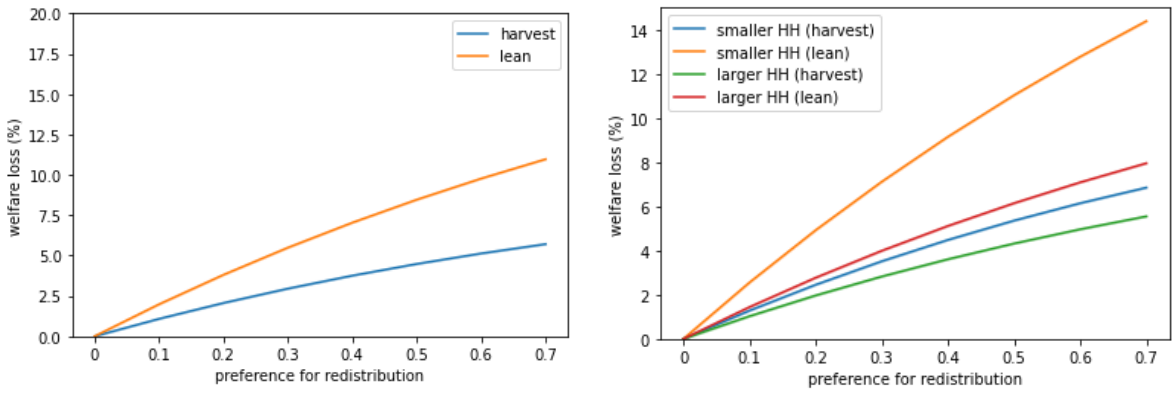
Why is the change in welfare losses among smaller households almost double of larger households? The difference in both PMTs is more pronounced for smaller households and in about 16% the two PMTs lead to different classifications compared to only 5% among larger households. That means that the PMT weights for smaller households are less stable and prediction errors arise more strongly if the PMT is applied to data that was collected in a different season than the training data.

Why are smaller households more sensitive to the timing of the data collection? Household size is a weighty predictor and adding or subtracting one household member changes

classifications of smaller households substantially. However, the weight of household size as a predictor as well as household size itself are not stable. Problems measuring household size are well documented (Beaman and Dillon 2012), and in many contexts household size can fluctuate markedly even in the short run. For example, monthly World Bank High Frequency Phone data in Malawi conducted during the COVID-19 pandemic suggest that household size is quite volatile. Feeding the month-on-month variation in household size from the nine waves of the monthly survey into a simulation with our prediction model suggests that classifications of smaller households are twice as sensitive to such short-term month-on-month variation than larger households i.e. the standard deviation of changes in classification after resampling is twice as high for smaller households (see Annex E for more information).

Our findings illustrate the difficulty of predicting the temporal dynamics of poverty if the input data do not cover these dynamics. This can lead to substantial underestimation of welfare losses, that are unevenly distributed. Household size is highly correlated with poverty, which is why aspects related to the number of people in a household has an important weight in predictions. However, these weights should not be static as also household size can vary over the course of a year (labour migration, seasonal work etc) and the measurement of household size is error prone, which particularly matters for smaller households.

Figure 4 Welfare loss distribution by household size



Notes: Welfare loss computed as percentage change in welfare in comparison to perfect targeting assuming different levels of inequality aversion. Evaluation based on the same test data, but models were trained with data only inluding harvest or lean season data.

Discussion

The COVID-19 pandemic has accelerated the introduction of social protection programs (Gentilini et al. 2022), continuing a trend of social protection roll-out witnessed over the past decade as a result of the extensive evidence on social protection's effectiveness (Bastagli et al. 2019). In the context of limited budgets, targeting of programs is often essential, with many programs relying on PMT to identify eligible households. Given the importance of social protection it is paramount to ensure that targeting is effective, transparent, and fair. However, in practice targeting procedures are quite opaque and it is often difficult for citizens to understand allocation procedures. This is problematic because a black-box decision making environment makes it complicated to monitor procedures and to appeal to unfair practices. It also jeopardizes the extension of cash transfer programs as opaque selection methods may reduce political support for the allocation of budgets.[8]

[8] For example, Uganda's Vulnerable Family Grant programme was discontinued in 2015 because the beneficiary selection was "contentious and not well accepted by the community" (https://socialprotection.org/discover/blog/social-assistance-grants-empowerment-sage-programme-uganda).

In this paper, we first argue that targeting error assessments should not solely focus on accuracy as performance metric because the measure implicitly assumes that societies are indifferent about who is being incorrectly classified and the measure does not reflect the social costs for achieving a given level of accuracy. Instead or in addition, we use a social welfare framework that weights targeting errors depending on the position in the welfare distribution and for different levels of societal inequality aversion. This extended framework helps to illustrate our point that increasing accuracy may even cause welfare losses in the case of fixed budgets. While this provides a more comprehensive assessment of targeting performance, we show that bias in the data, here in the form of label bias and unstable PMT weights, leads to substantial underestimation of welfare losses. The magnitude of the usually unobserved welfare loss components is concerningly high, which raises more general questions about the reliability of targeting assessments. Lastly, we show that these unobserved welfare losses are unequally distributed and disproportionately carried by smaller households.

We focus only on two sources of bias related to the measurement of target variable and the stability of weights. Other sources of bias could be at play too and in other contexts, other factors may play a more important role. However, our results indicate that even focusing on single biases can lead to significant underestimation of welfare losses. This becomes particularly evident if assessed through a social welfare lens that explicitly weights in distortionary effects. The accuracy of the predictions alone does not show these discrepancies clearly, as it does not discriminate between inclusion and exclusion errors and fails to reflect at which costs the accuracy is achieved (over versus under prediction of poverty). The downside of the approach is that arbitrary assumptions about the social welfare function and societal inequality aversion need to be made. Besides that, other societal preferences for instance for fairness and risk aversion could further affect welfare losses and render our estimates incomplete and partial at best. For example, disadvantaging smaller households is

likely to be regarded as unfair by many people and could thus cause welfare losses by itself. In this paper, we focus on the distributional implications, but future work could regard fairness ratings further emphasising the trade-off between equity and efficiency along the lines of Premand and Schnitzer (2021) giving room for legitimacy perceptions of citizens.

Our findings and subsequent conclusions lead us to call for a broader discussion, removing layers of opacity in decision-making and bringing accountability and evaluation to all stages of the lifecycle of a social protection program. Does increasing awareness lead to fairness? Unfortunately not, and most of the welfare losses we found are coming from data biases that tend to be hidden in the data. Some of this could probably be mitigated but the question remains whether predictions should be used in the first place. Fairness is subjective and measuring statistical indicators of unfairness of prediction outcomes ex ante is by design incomplete. Another route, proposed in discussions concerning prediction fairness, could be to focus more on causal mechanisms that lead to the need to predict the outcome of interests in the first place.[9] An example of this in the social protection domain could be Kenya's Hunger Safety Net Program automatic emergency scale up component that expands the program if a remotely sensed drought indicator falls below a critical threshold. This insurance component aims to prevent households from dropping into poverty due to natural disasters instead of predicting ex post who is poor.

Finally, it is also important to address significant data issues. Many countries use household surveys to build PMT coefficients to target beneficiaries; but these coefficients are applied to other data (from registries or census) to actually decide if a household can receive or not a program. Our analysis has shown how using PMT estimates with a different set of data can

[9] https://fairmlbook.org/index.html

increase welfare losses. Therefore, a discussion on how to harmonize household surveys and administrative data is crucial.

References

Adler, Matthew D. 2019. "Cost-Benefit Analysis and Social Welfare Functions." *In: White, M. D. (Ed.). (2019). The Oxford Handbook of Ethics and Economics. Oxford University Press, USA.*

Aiken, Emily, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua E Blumenstock. 2022. "Machine Learning and Phone Data Can Improve Targeting of Humanitarian Aid." *Nature* 603 (7903): 864–70.

Aiken, Emily L, Guadalupe Bedoya, Joshua E Blumenstock, and Aidan Coville. 2023. "Program Targeting with Machine Learning and Mobile Phone Data: Evidence from an Anti-Poverty Intervention in Afghanistan." *Journal of Development Economics* 161: 103016.

Alderman, Harold, Jere R Behrman, and Afia Tasneem. 2019. "The Contribution of Increased Equity to the Estimated Social Benefits from a Transfer Program: An Illustration from PROGRESA/Oportunidades." *The World Bank Economic Review* 33 (3): 535–50.

Atkinson, Anthony B. 1970. "On the Measurement of Inequality." *Journal of Economic Theory* 2 (3): 244–63.

———. 2005. "On Targeting Social Security: Theory and Western Experience with Family Benefits." *In: Van de Walle, D., Nead, K. (Eds.). Public Spending and the Poor. Theory and Evidence. Washington DC: The World Bank.*

Ayush, Kumar, Burak Uzkent, Marshall Burke, David Lobell, and Stefano Ermon. 2020. "Generating Interpretable Poverty Maps Using Object Detection in Satellite Images." *ArXiv Preprint ArXiv:2002.01612*.

Banerjee, Abhijit, Rema Hanna, Benjamin A Olken, and Sudarno Sumarto. 2020. "The (Lack of) Distortionary Effects of Proxy-Means Tests: Results from a Nationwide Experiment in Indonesia." *Journal of Public Economics Plus* 1: 100001.

Barocas, Solon, and Andrew D Selbst. 2016. "Big Data's Disparate Impact." *California Law Review*, 671–732.

Barrientos, Armando, Stephan Dietrich, Franziska Gassmann, and Daniele Malerba. 2022. "Prioritarian Rates of Return to Antipoverty Transfers." *Journal of International Development* 34 (3): 550–63.

Bastagli, Francesca, Jessica Hagen-Zanker, Luke Harman, Valentina Barca, Georgina Sturge, and Tanja Schmidt. 2019. "The Impact of Cash Transfers: A Review of the Evidence from Low-and Middle-Income Countries." *Journal of Social Policy* 48 (3): 569–94.

Beaman, Lori, and Andrew Dillon. 2012. "Do Household Definitions Matter in Survey Design? Results from a Randomized Survey Experiment in Mali." *Journal of Development*

Economics 98 (1): 124–35.

Beegle, Kathleen, Joachim De Weerdt, Jed Friedman, and John Gibson. 2012. "Methods of Household Consumption Measurement through Surveys: Experimental Results from Tanzania." *Journal of Development Economics* 98 (1): 3–18.

Blumenstock, Joshua Evan. 2016. "Fighting Poverty with Data." *Science* 353 (6301): 753–54.

Brown, Caitlin, Martin Ravallion, and Dominique Van de Walle. 2018. "A Poor Means Test? Econometric Targeting in Africa." *Journal of Development Economics* 134: 109–24.

Caeyers, Bet, Neil Chalmers, and Joachim De Weerdt. 2012. "Improving Consumption Measurement and Other Survey Data through CAPI: Evidence from a Randomized Experiment." *Journal of Development Economics* 98 (1): 19–33.

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

Campo, Stellio Del, David Anthoff, and Ulrike Kornek. 2021. "Inequality Aversion for Climate Policy."

Coady, David P, Devin D'Angelo, and Brooks Evans. 2020. "Fiscal Redistribution and Social Welfare: Doing More or More to Do?" EUROMOD Working Paper.

Cooke, Ira R, Simon A Queenborough, Elizabeth H A Mattison, Alison P Bailey, Daniel L Sandars, A R Graves, J Morris, Philip W Atkinson, Paul Trawick, and Robert P Freckleton. 2009. "Integrating Socio-economics and Ecology: A Taxonomy of Quantitative Methods and a Review of Their Use in Agro-ecology." *Journal of Applied Ecology* 46 (2): 269–77.

Corbett-Davies, Sam, and Sharad Goel. 2018. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." *ArXiv Preprint ArXiv:1808.00023*.

Creedy, John. 2006. "Evaluating Policy: Welfare Weights and Value Judgements." *University of Melbourne, Department of Economics: Research Paper*, no. 971.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. "Fairness through Awareness." In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–26.

Ferrer, Xavier, Tom van Nuenen, Jose M Such, Mark Coté, and Natalia Criado. 2021. "Bias and Discrimination in AI: A Cross-Disciplinary Perspective." *IEEE Technology and Society Magazine* 40 (2): 72–80.

Foster, James, Joel Greer, and Erik Thorbecke. 1984. "A Class of Decomposable Poverty Measures." *Econometrica: Journal of the Econometric Society*, 761–66.

Gajane, Pratik, and Mykola Pechenizkiy. 2017. "On Formalizing Fairness in Prediction with Machine Learning." *ArXiv Preprint ArXiv:1710.03184*.

Gazeaud, Jules. 2020. "Proxy Means Testing Vulnerability to Measurement Errors?" *The Journal of Development Studies* 56 (11): 2113–33.

Gentilini, Ugo, Mohamed Bubaker Alsafi Almenfi, T M M Iyengar, Yuko Okamura, John Austin Downes, Pamela Dale, Michael Weber, David Locke Newhouse, Claudia P Rodriguez Alas, and Mareeha Kamran. 2022. "Social Protection and Jobs Responses to

COVID-19."

Gibson, John, and Bonggeun Kim. 2007. "Measurement Error in Recall Surveys and the Relationship between Household Size and Food Demand." *American Journal of Agricultural Economics* 89 (2): 473–89.

Hanna, Rema, and Benjamin A Olken. 2018. "Universal Basic Incomes versus Targeted Transfers: Anti-Poverty Programs in Developing Countries." *Journal of Economic Perspectives* 32 (4): 201–26.

Hopper, Robert. 2020. "The Dynamics of Deprivation in Malawi: The Multi-Dimensional Effects of the Lean Season on Children."

Jolliffe, Dean, and Samuel Tetteh-Baah. 2022. "Identifying the Poor Accounting for Household Economies of Scale in Global Poverty Estimates."

Kind, Jarl, W J Wouter Botzen, and Jeroen C J H Aerts. 2017. "Accounting for Risk Aversion, Income Distribution and Social Welfare in Cost-benefit Analysis for Flood Risk Management." *Wiley Interdisciplinary Reviews: Climate Change* 8 (2): e446.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5): 491–95.

Ledesma, Chiara, Oshean Lee Garonita, Lorenzo Jaime Flores, Isabelle Tingzon, and Danielle Dalisay. 2020. "Interpretable Poverty Mapping Using Social Media Data, Satellite Images, and Geospatial Information." *ArXiv Preprint ArXiv:2011.13563*.

Lum, Kristian, and William Isaac. 2016. "To Predict and Serve?" *Significance* 13 (5): 14–19.

McBride, Linden, and Austin Nichols. 2018. "Retooling Poverty Targeting Using Out-of-Sample Validation and Machine Learning." *The World Bank Economic Review* 32 (3): 531–50.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys (CSUR)* 54 (6): 1–35.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53.

Pol, Thomas Van der, Frits Bos, and Gerbert Romijn. 2017. "Distributionally Weighted Cost-Benefit Analysis: From Theory to Practice."

Premand, Patrick, and Pascale Schnitzer. 2021. "Efficiency, Legitimacy, and Impacts of Targeting Methods: Evidence from an Experiment in Niger." *The World Bank Economic Review* 35 (4): 892–920.

Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. 2020. "An Economic Perspective on Algorithmic Fairness." In *AEA Papers and Proceedings*, 110:91–95.

Saez, Emmanuel, and Stefanie Stantcheva. 2016. "Generalized Social Marginal Welfare Weights for Optimal Tax Theory." *American Economic Review* 106 (01): 24–45.

Sen, Amartya. 1976. "Poverty: An Ordinal Approach to Measurement." *Econometrica: Journal*

of the Econometric Society, 219–31.

———. 1977. "On Weights and Measures: Informational Constraints in Social Welfare Analysis." *Econometrica: Journal of the Econometric Society*, 1539–72.

Annex

D PMT review

A small case study review into transparency in PMT was conducted in 2021. East Africa was chosen as a case study due to the number of programmes which rely on PMT for targeting and the focus of this study. The review focused on areas of methodological rigour and transparency highlighted as important on the basis of the research conducted for this note. The review's scope was limited to official programme documentation published by the relevant government for the most part (the review of RCTs included university or consultancy published documents). Moreover, given the importance of social protection programmes to individual welfare and poverty eradication, the review included solely documents found in the public domain - those that could be found by individuals. In addition, an email soliciting further documents and information (all in the public domain) from individuals or departments which were responsible for the implementation of programs which might not be visible in our searches. The review found that while the targeting methodology had been published in approximately two-thirds of cases, fewer than half had used even a rudimentary estimation method. No programs in our review had trained models for out-of-sample predictions. While approximately half of the key programs in our review did have the PMT variables published, fewer than two-fifths had PMT variables weights available in the public domain. In light of the findings of this note, this is notable limitation in the accountability and transparency of programmes. The proportion of programs that had a published RCT conducted (36%) - not including non-RCT based evaluations - indicates that ex-post evaluation of impact appears to be of greater concern to policymakers than ex-ante targeting evaluations.

Table 4 Public Cash Transfer Programs with PMT in East-Africa

	Methodology published	*Estimation method*	*OoS prediction*	*Variables published*	*Weights published*	*RCT*
***Kenya** HSNP*	Yes	Standard OLS	No	No	No	Yes
***Kenya** OVC-CT*	Yes	Standard OLS	No	No	No	Yes
***Malawi** SCTP*	No	No	No	Yes	No	Yes
***Zambia** SCT*	Partially	Principal Component Analysis	No	Yes	No	Yes
***Zimbabwe** HSCT*	Yes	No	No	Yes	No	No
***Mozambique** PSSB*	Yes	No	No	No	No	No
***Madagascar** Let us learn cash transfer*	No	No	No	No	No	No
***Djibouti** PNSF*	No	No	No	Yes	No	No

***Mauritius** Social Aid Benefits*	Yes	Quantile regression	No	Yes/No	Yes	No
***Ethiopia** Urban Productive Safety Net Project*	Yes	Standard OLS	No	Yes	Yes	No
***Ethiopia** PSNP*	Yes	No	No	No	No	No
Total	7/11 (64%)	5/11 (45%)	0/11 (0%)	6/11 (54%)	2/11 (18%)	4/11 (36%)

B Summary Statistics

Malawi

Table 5 Summary of PMT variables, Malawi

Variable	*non-poor*	*poor*	*smaller HH*	*larger HH*	*Variable*	*non-poor*	*poor*	*smaller HH*	*larger HH*
Household size	3,48	5,12	2,82	6,54	*Soap*	0,24	0,08	0,12	0,15
	(2,16)	(1,77)	(1,04)	(1,77)		(0,42)	(0,36)	(0,33)	(0,36)
Household size sq.	16,77	31,17	9,05	45,84	*Bed*	0,48	0,24	0,27	0,38
	(22,66)	(30,46)	(5,53)	(30,46)		(0,50)	(0,48)	(0,45)	(0,48)
Age head	40,15	43,70	40,96	44,20	*Bike*	0,40	0,34	0,28	0,45
	(16,56)	(13,37)	(18,44)	(13,37)		(0,49)	(0,50)	(0,45)	(0,50)
Age head sq.	1886,24	2169,89	2017,55	2131,86	*Music player*	0,28	0,10	0,13	0,20
	(1612,31	1343,62	1822,06	1343,62		0,45	0,40	0,34	0,40
North	0,15	0,15	0,14	0,16	*Coffee table*	0,24	0,05	0,10	0,14
	(0,36)	(0,37)	(0,34)	(0,37)		(0,43)	(0,35)	(0,29)	(0,35)
Central	0,44	0,35	0,36	0,41	*Iron roof*	0,34	0,13	0,16	0,26
	(0,50)	(0,49)	(0,48)	(0,49)		(0,47)	(0,44)	(0,37)	(0,44)
Rural	0,77	0,93	0,86	0,88	*Dimba garden*	0,29	0,34	0,27	0,37
	(0,42)	(0,32)	(0,34)	(0,32)		(0,45)	(0,48)	(0,45)	(0,48)
Household head never married	0,08	0,01	0,05	0,00	*Goats*	0,20	0,22	0,16	0,28
	(0,27)	(0,06)	(0,23)	(0,06)		(0,40)	(0,45)	(0,37)	(0,45)
Share no education	0,11	0,19	0,17	0,16	*Dependency ratio*	0,71	1,34	0,79	1,50
	(0,26)	(0,19)	(0,31)	(0,19)		(0,74)	(0,96)	(0,80)	(0,96)
Share can read	0,71	0,55	0,58	0,63	*hfem*	0,20	0,24	0,28	0,16
	(0,37)	(0,34)	(0,41)	(0,34)		(0,40)	(0,37)	(0,45)	(0,37)
Number of rooms	2,56	2,47	2,17	2,88	*Grass roof*	0,56	0,83	0,76	0,71
	(1,39)	(1,43)	(1,07)	(1,43)		(0,50)	(0,45)	(0,43)	(0,45)
Cement floor	0,36	0,11	0,18	0,22	*Mortar pestle*	0,45	0,53	0,40	0,61
	(0,48)	(0,41)	(0,39)	(0,41)		(0,50)	(0,49)	(0,49)	(0,49)
Electricity	0,15	0,01	0,05	0,07	*Table*	0,46	0,30	0,29	0,44
	(0,35)	(0,25)	(0,22)	(0,25)		(0,50)	(0,50)	(0,45)	(0,50)
Flushing toilet	0,07	0,01	0,02	0,03	*Clock*	0,34	0,12	0,17	0,24
	(0,25)	(0,18)	(0,15)	(0,18)		(0,47)	(0,42)	(0,37)	(0,42)

Note: Standard deviation in parentheses.

Lean vs harvest season

The survey was conducted over a period of 12 months in 2004/5 based on 30 strata, with 240 households to be sampled per strata. The enumeration of households was designed to be spread over the entire year to take into account differences in rural communities in the

harvest and lean seasons. Households in each Enumeration Area - progression from one to the next determined by the enumerator - were sampled on the basis of registers, with each Enumeration Area taking one month to sample. Given the random sampling design and the simultaneous nationwide roll-out of the survey, differences between lean and harvest seasons should be negligible. The following test of non-fungible household characteristics is further evidence.

Table 6 Lean season balance tests

	Non-lean season	*Lean season*	*Pr Chi2*
No cement floor	4527	4509	0.92
Cement floor	1127	1117	
No electricity	5321	5299	0.86
Electricity	333	327	
No flushing toilet	5503	5459	0.34
Flushing toilet	151	167	
No grass roof	1506	1447	0.27
Grass roof	4148	4179	

Note: chi2 test for differences in distribution of variables between lean and harvest season

Tanzania

Table 7 Summary of PMT variables, Tanzania

variable	***non-poor***	***poor***	***smaller HH***	***larger HH***	***variable***	***non-poor***	***poor***	***smaller HH***	***larger HH***
Urban	0,47	0,17	0,41	0,26	HH head widowed	0,13	0,14	0,16	0,09
	(0,50)	(0,44)	(0,49)	(0,44)		(0,33)	(0,29)	(0,37)	(0,29)
Age	45,46	48,30	45,45	48,19	Improved floor	0,44	0,75	0,52	0,64
	(16,46)	(13,58)	(18,01)	(13,58)		(0,50)	(0,48)	(0,50)	(0,48)
Age squared	2337,05	2586,53	2389,89	2506,47	Improved roof	0,27	0,43	0,36	0,30
	(1706,16)	(1468,07)	(1879,16)	(1468,07)		(0,44)	(0,46)	(0,48)	(0,46)
Household size	4,45	6,44	3,32	7,84	Improved wall	0,59	0,90	0,68	0,77
	(2,64)	(2,29)	(1,35)	(2,29)		(0,49)	(0,42)	(0,47)	(0,42)
Household size sq.	26,76	49,31	12,86	66,63	Number of rooms	3,35	3,88	2,89	4,47
	(35,33)	(48,49)	(8,53)	(48,49)		(1,82)	(1,83)	(1,44)	(1,83)
Children under 5	0,77	1,51	0,62	1,67	Water supply	0,38	0,12	0,32	0,20
	(0,95)	(1,16)	(0,78)	(1,16)		(0,48)	(0,40)	(0,47)	(0,40)
Elderly householder	0,30	0,37	0,33	0,33	Flushing toilet	0,17	0,01	0,12	0,09
	(0,57)	(0,59)	(0,58)	(0,59)		(0,38)	(0,28)	(0,32)	(0,28)
Primary education head	0,78	0,64	0,70	0,74	Type of stove	0,36	0,04	0,29	0,14
	(0,42)	(0,44)	(0,46)	(0,44)		(0,48)	(0,34)	(0,45)	(0,34)
Secondary education head	0,15	0,02	0,11	0,08	Electricity	0,20	0,05	0,16	0,11
	(0,35)	(0,27)	(0,31)	(0,27)		(0,40)	(0,31)	(0,37)	(0,31)

Household head married	0,70	0,78	0,64	0,86
	(0,46)	(0,35)	(0,48)	(0,35)

Note: Standard deviations in parentheses.

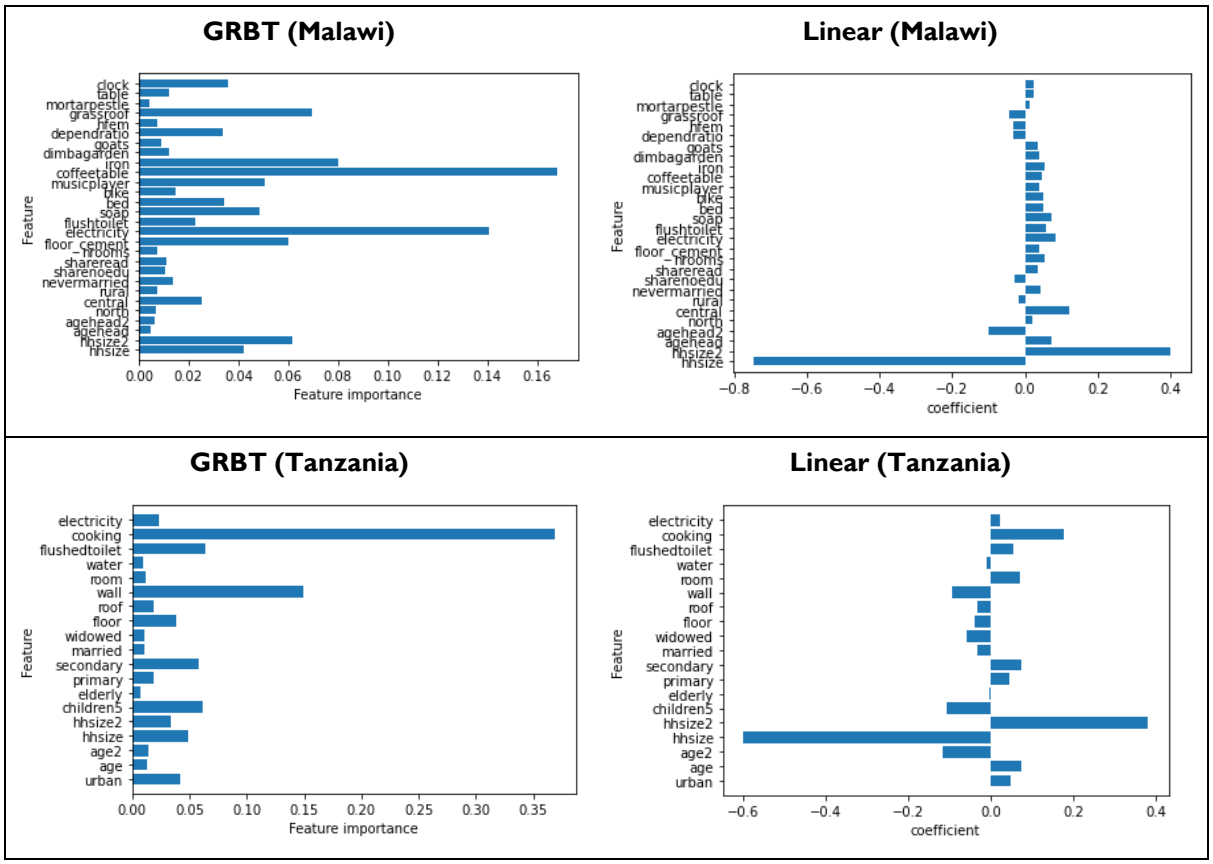
C Prediction model

We randomly draw training data ($N * 0.8$) to estimate the parameters of the models and test data ($N * 0.2$) that we hold back to examine classification errors. We search over a range of hyper-parameter values to select the best specification. As we are considering a large number of combinations of hyper-parameter values in the gradient boosting models, we randomly tested 10000 model specifications out of all possible combinations and thereafter fine-tuned the models. We measure the model performance based on the accuracy of predictions using tenfold cross-validation. The parameters of the preferred specifications, as presented in the main analysis, are depicted below. Below we also show the feature importance and coefficients resultsing from the xgboost and linear models.

Table 8 Hyper-parameter grid search gradient boosting model

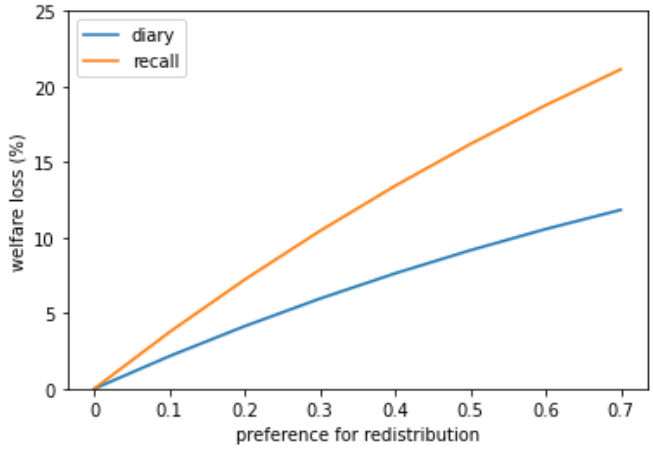
Parameters	**Malawi**	**Tanzania**
max_depth	4	2
min_samples_split	2	10
min_samples_leaf	76	66
max_features	14	1
sub_sample	0.48	0.86
learning_rate	0.055	0.13
n_estimators	220	310

Figure 2 Feature importance/coefficients



Note: Xgboost model shows feature importance by counting appearances of features in decision trees and logit shows the estimated coefficients size of standardized PMT variables.

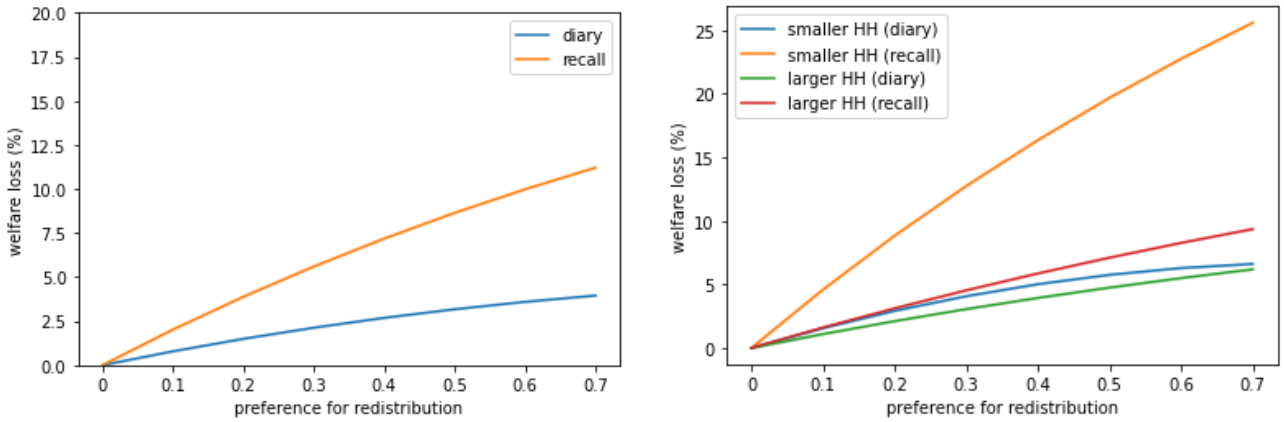
Figure 3 Marginal welfare loss of unit transfer with diary and recall PMT
(only using <u>personal diaries and high supervision frequency</u> treatment to train diary model and for model validation)



D Per Capita vs. Adult Equivalent scale

To account for economies of scale within households we follow () in dividing household by the square root of the number of household members instead of using per capita reports as robustness test. After converting consumption reports, we use the same approach as in the main analysis i.e. we train two separate PMTs with diary and recall data and validate those with diary test data. In the simulations we adjust the poverty line in a way such that the poverty rates remain the same and the budget we allocate remains the same as in the per capita case of the main text. The figure below show the resulting welfare losses of the two PMTs overall and separately by household size.

Figure 4 Welfare losses if consumption is converted to account for household economies of scale



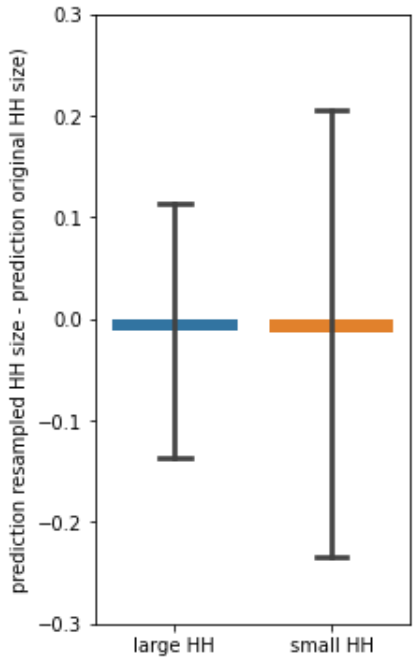
Notes: Welfare loss computed as percentage change in welfare in comparison to perfect targeting assuming different levels of inequality aversion. Evaluation based on the same test data, but models were trained with data only inluding recall or diary data.

E Household size variance (Malawi)

Household size variance was examined using High Frequency Phone Survey (HFPS) data collected by the National Statistical Office of Malawi (supported by the World Bank) monthly over a one-year period from May 2020 and June 2021. The sampling frame draws on the Integrated Household Panel Survey (IHPS) conducted in 2019. At the time of analysis, nine months of data were available. The probability of a household being size x in month m+1 dependent on their household size in month m is given in the table below. Given that the HFPS survey builds on the IHPS survey, the household roster was pre-filled, with respondents asked to confirm whether each member of the roster was still a member of the household, and asked whether there were members of the household at that time not included in the roster. A household member was defined as a person who normally sleep in the same dwelling and share their meals together.

Table 9 Month-on-month variation in household size, Malawi phone survey

		Household Size m								
		1	2	3	4	5	6	7	8	9
Household Size m+1	1	0,86	0,08	0,04	0,03	0,01	0,00	0,00	0,00	0,00
	2	0,09	0,78	0,06	0,04	0,01	0,01	0,00	0,00	0,00
	3	0,02	0,10	0,68	0,05	0,02	0,00	0,01	0,00	0,00
	4	0,01	0,03	0,12	0,65	0,07	0,02	0,01	0,00	0,00
	5	0,00	0,01	0,04	0,17	0,72	0,08	0,03	0,01	0,00
	6	0,00	0,01	0,01	0,03	0,14	0,59	0,08	0,02	0,01
	7	0,00	0,00	0,00	0,02	0,02	0,18	0,62	0,10	0,04
	8	0,01	0,00	0,01	0,00	0,02	0,04	0,14	0,60	0,09
	9	0,00	0,00	0,04	0,01	0,00	0,08	0,11	0,28	0,86
source: World Bank's high frequency phone surveys										



Note: Resampling of household size is based on monthly phone survey data. Marker show standard deviation of original prediction minus prediction with resampled household size. Results based on Monte Carlo simulation with 1000 iterations.

The UNU-MERIT WORKING Paper Series

2023-01 *Can international mobility shape students' attitudes toward inequality? The Brazilian case* by Cintia Denise Granja, Fabiana Visentin and Ana Maria Carneiro

2023-02 *Demand-led industrialisation policy in a dual-sector small open economy* by Önder Nomaler, Danilo Spinola and Bart Verspagen

2023-03 *Reshoring, nearshoring and developing countries: Readiness and implications for Latin America* by Carlo Pietrobelli and Cecilia Seri

2023-04 *The role of product digitization for productivity: Evidence from web-scraping European high-tech company websites* by Torben Schubert, Sajad Ashouri, Matthias Deschryvere, Angela Jäger, Fabiana Visentin, Scott Cunningham, Arash Hajikhani, Lukas Pukelis and Arho Suominen

2023-05 *More than a feeling: A global economic valuation of subjective wellbeing damages resulting from rising temperatures* by Stephan Dietrich and Stafford Nichols

2023-06 *Was Robert Gibrat right? A test based on the graphical model methodology* by Marco Guerzoni, Luigi Riso and Marco Vivarelli

2023-07 Predicting social assistance beneficiaries: On the social welfare damage of data biases *by Stephan Dietrich, Daniele Malerba and Franziska Gassmann*