Data & Policy (2024), 6: e3
doi:10.1017/dap.2023.38

CAMBRIDGE
UNIVERSITY PRESS

RESEARCH ARTICLE

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

Stephan Dietrich[1], Daniele Malerba[2] and Franziska Gassmann[1]

[1]UNU-MERIT, Maastricht, Netherlands
[2]IDOS, Bonn, Germany
Corresponding author: Stephan Dietrich; Email: s.dietrich@maastrichtuniversity.nl

Received: 14 March 2023; **Revised:** 12 September 2023; **Accepted:** 24 October 2023

Keywords: Cash transfers; Fair machine learning; Proxy means test; Social protection; Targeting

Abstract

Cash transfer programs are the most common anti-poverty tool in low- and middle-income countries, reaching more than one billion people globally. Benefits are typically targeted using prediction models. In this paper, we develop an extended targeting assessment framework for proxy means testing that accounts for societal sensitivity to targeting errors. Using a social welfare framework, we weight targeting errors based on their position in the welfare distribution and adjust for different levels of societal inequality aversion. While this approach provides a more comprehensive assessment of targeting performance, our two case studies show that bias in the data, particularly in the form of label bias and unstable proxy means testing weights, leads to a substantial underestimation of welfare losses, disadvantaging some groups more than others.

Policy Significance Statement

Cash transfer programs are the most common anti-poverty tool in low- and middle-income countries, reaching more than one billion people globally. Benefits are typically targeted using prediction models. We argue that targeting assessments should carefully consider the characteristics of erroneously targeted households. This is crucial to build social protection systems that align with local redistributive preferences and to avoid discriminatory biases that may be concealed in the data. Our findings and subsequent conclusions prompt us to advocate for a broader discussion, aiming to remove layers of opacity in decision-making and to introduce accountability and evaluation throughout all stages of the lifecycle of social protection policies.

1. Introduction

Cash transfer programs, the most common anti-poverty tool in low- and middle-income countries, have expanded massively over the last decade (Gentilini et al., 2022). In the context of limited budgets, targeting of these programs to the poor is often essential, with many programs relying on data-driven systems to identify eligible households. Because household living standards are difficult to measure and verify, beneficiary selection is often based on methods aimed at ranking households and individuals from poorest to richest. Proxy means testing (PMT) is a popular tool for identifying eligible households based on predicted income or wealth. In this context, policy designers aim to correctly identify 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 shown how flexible machine learning models and novel data sources can reduce targeting errors of such screening systems (McBride and Nichols, 2018; Aiken et al., 2022).

The academic discourse around targeting assessments often focuses on the accuracy of benefit allocation. However, if societies hold preferences for redistribution, a mere comparison of targeting accuracy may offer an incomplete picture. Therefore, performance assessments should place greater weight on targeting errors among poor households than among non-poor households. From a social welfare perspective, providing a transfer to a very poor household holds more value than giving the same transfer to a richer household; thus, the specifics of who is erroneously targeted or excluded are critical. From this vantage point, an increase in prediction accuracy can even result in welfare losses if poorer households are misclassified. Only a few papers evaluate and assess targeting errors (Hanna and Olken, 2018) and cash transfer programs more broadly (Alderman et al., 2019; Barrientos et al., 2022) using a social welfare framework. However, these welfare estimates are calibrated with recent and correct consumption data, and do not account for welfare losses caused by data biases.

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 context, discrimination implies that members of certain societal groups are less likely to benefit from algorithmic decisions than others for reasons unrelated to the targeting criteria. This has spurred discussions on fairness considerations in prediction models (Kleinberg et al., 2015; Corbett-Davies and Goel, 2018; Obermeyer et al., 2019). The origins of biases and algorithmic discrimination often reside in the data used to train models rather than in the estimators themselves. This could, for instance, stem from measurement errors in the form of biased proxy indicators for the true outcome of interest or non-generalizable data (Mehrabi et al., 2021). When applied to targeting error assessments, these data biases are challenging to observe. This would suggest that poor households from algorithmically disadvantaged groups are less likely to be classified as such, leading benefits to be allocated to other, comparatively better-off households. Such allocations have redistributive consequences and result in welfare losses that often remain hidden in the data. For PMTs, this suggests that welfare losses from targeting errors are underestimated.

In this paper, we formalize the welfare implications of targeting errors using a social welfare weight framework and demonstrate how increases in prediction accuracy can even result in welfare losses. Subsequently, building on the work of Gazeaud (2020) and McBride and Nichols (2018), we construct our own PMTs using data from Tanzania and Malawi. These case studies highlight how actual welfare losses may arise from systematic measurement errors, leading to skewed targeting errors. As an illustrative example, we examine household size to shed light on the underlying mechanisms that contribute to an uneven distribution of welfare losses. Our analysis reveals that reporting bias and instability in PMT weights disadvantage smaller households, making them disproportionately more likely to be inaccurately classified as non-poor. Consequently, they would not be identified as eligible by the PMT. Our findings indicate that welfare assessments are often substantially underestimated, with the data biases analyzed accounting for up to half of the targeting error-related welfare losses in our two case studies.

The application of PMT in targeting anti-poverty programs has not been without scrutiny and has faced criticism on multiple fronts. Recent discussions on PMT can be broadly divided into three categories. The first focuses on the relationship between the targeting process and the trade-off between equity and efficiency in beneficiary selection (Brown et al., 2018; Hanna and Olken, 2018; Premand and Schnitzer, 2021). The second addresses the challenges inherent in the algorithmic process itself, examining the efficacy of the prediction process using new estimators and data sources (Brown et al., 2018; McBride and Nichols, 2018; Aiken et al., 2022; Aiken et al., 2023). A third strand delves into the presence of errors in measuring the dependent variable (Gazeaud, 2020) or misreporting of PMT variables (Banerjee et al., 2020)—issues that extend beyond the confines of PMT but represent persistent underlying problems. However, the welfare implications of the combined effects of data biases and an understanding of which groups are systematically disadvantaged remain largely uncharted territories. We identify three pivotal contributions of our paper to both academic and policy dialogues on PMTs.

First, this paper aims to contribute to discussions about the use of data-driven decision-making systems in the public policy domain. The ever-increasing availability of data is offering new opportunities to efficiently target policies to those most in need of public support. Such tools include poverty screening methods based on satellite imagery, cell phone data, or social media (Blumenstock, 2016; Ayush et al., 2020; Ledesma et al., 2020; Aiken et al., 2022). While this is undeniably a promising advancement, the growing complexity in these systems can also heighten the risk of unobserved problematic biases in benefit allocation due to black-box procedures. Typically, PMT weights are deliberately not disclosed for valid reasons, even though individuals might deduce or form beliefs about these weights (Camacho and Conover, 2011; Banerjee et al., 2020). Nevertheless, it is reasonable for citizens to seek information about the general targeting procedures. In preparation for this project, and to gain insight into the extent of opacity in PMTs, we examined 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 utilizing PMTs (see Annex A). Merely half of these programs share information on their targeting methodology, and we found only a single instance where full PMT weights were published. This paper showcases how typically unobservable biases lead to welfare losses that are unequally distributed, impacting some societal groups more detrimentally than others. These results call for closer scrutiny and more transparency in targeting procedures ("fairness through awareness"; Dwork et al., 2012). However, they also raise the question of whether, in certain contexts, targeting should be regarded as a prediction problem in the first place. Alternatively, might other targeting approaches that do not rely on predictions perform better in terms of social welfare, especially when data biases and legitimacy considerations are factored in? This paper does not provide a definitive answer, as it hinges on contextual factors and societal preferences. Still, the findings suggest a need for broader discussion about the application and implications of using PMT systems.

Second, and relatedly, the paper delves into the growing fair machine learning literature that has gained momentum in recent years with inputs from different disciplines. Several papers have identified biases in data used in public domains, such as policing, law, and health, which resulted in and may have even reinforced discriminatory practices (Barocas and Selbst, 2016; Lum and Isaac, 2016; Obermeyer et al., 2019). In this paper, we define data deviations from the unobserved reality, or the ground truth, as bias. Underreporting of assets by households to appear less wealthy is an example. We view discrimination as the problematic cases of bias that systematically harm certain groups more than others. While household size, used here as an illustrative example, is not a protected class like race or gender, the issue remains significant. After all, a child born into a household has no control over its size. Several papers have discussed sources of discrimination and proposed indicators for algorithmic unfairness (Gajane and Pechenizkiy, 2017; Corbett-Davies and Goel, 2018; Rambachan et al., 2020; Ferrer et al., 2021; Mehrabi et al., 2021). If biases are embedded in the data, statistical indicators derived from these data may fail to highlight true imbalances. In this paper, we demonstrate that welfare losses due to targeting errors are substantially underestimated because of these biases. These results indicate that opacity in procedures combined with a purely data-driven approach might mask discriminatory practices. By examining two selected sources of biases, this paper seeks to contribute to the discussion of fairness considerations in social protection systems.

Third, this paper connects the use of social welfare weights to targeting issues in cash transfer programs in low and middle-income countries. Typically, when assessing public social policies, a welfarist approach is adopted where the social planner (often the government) seeks to maximize a social welfare function (Sen, 1977; Saez and Stantcheva, 2016). This planner employs welfare weights, acknowledging that increases in welfare for the worse off carry more weight in terms of social welfare than those for the better off. However, cash transfer programs are often evaluated by observing changes in key outcomes, such as poverty or mean consumption. This method essentially employs a utilitarian approach, as it does not explicitly prioritize improvements among low-income groups (Barrientos et al., 2022). Furthermore, current research does not typically delve into issues of biases and discrimination, even though the primary

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

objective of these policies is often to support the ultra-poor and marginalized (Creedy, 2006; Coady et al., 2020).

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.

2. Targeting Errors and Social Welfare

The demand for anti-poverty programs can be vast, often outstripping available government resources, necessitating targeting. From a purely theoretical standpoint, given a constrained budget, focusing on the poor is the most effective strategy to alleviate poverty. However, the trade-offs between targeting costs and efficiency have been well-documented in literature (Coady et al., 2004; Devereux et al., 2017; Hanna and Olken, 2018). Policy makers face the challenge of selecting a method to identify beneficiaries, but their task is complicated by having incomplete information on household living standards, which hinders accurate ranking of individuals from poorest to richest.

Both vertical and horizontal inefficiencies can diminish the effect of public spending (Atkinson, 2005). Vertical efficiency focuses on targeting accuracy (ensuring only the target group benefits), while horizontal efficiency ensures program comprehensiveness (the entire target group is covered). In anti-poverty programs, efficiency concepts hinge on poverty measurement or the set policy objectives. If a program's main aim is poverty alleviation, that is, raising everyone to a designated poverty line, then its efficiency is gauged by how much it narrows the poverty gap with the allocated budget. If greater importance is given to those most below the poverty line, prioritizing the poorest becomes more efficient. This is reflected in the parameter α of the standard Foster et al. (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 program reduces the poverty headcount rate and assigns transfers to those close to the poverty line. The strength of the poverty reduction objective affects the assessment of targeting efficiency. With a strict objective, the value of a transfer to a non-poor is zero. However, objectives that include the near-poor might assign some value to the transfer. Close to the poverty line, the marginal value is positive but less 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. Since the scores are only approximate actual living standards, they lead to targeting errors, reducing both vertical and horizontal efficiency of the allocated budget.

2.1. *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:

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

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

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:

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

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 following two cases: (a) if a small transfer is given to a household with high welfare weight and (b) 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 (2015). 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 (Campo et al., 2021). Most studies try to reveal the inequality aversion parameter through hypothetical (e.g., using experiments) or actual data (e.g., 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.

2.2. *Targeting errors, bias, and welfare loss*

PMT targeting is based on predictions and generally assumes unbiased data, although Gazeaud (2020) is an exception. However, there is growing evidence and discussion around biases in algorithmic decision-making in the public policy domain, which 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. We view biases as deviations of the observed data from the unobservable truth

[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 et al., 2017; Van der Pol et al., 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.

[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): $SWF = \frac{\sum_{i=1}^{n}(y+b)^{1-p}}{1-p}$.

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

that are systematically related to group affiliations. This leads to distortions in the optimal distribution of benefits among groups, resulting in social welfare losses. The magnitude of these welfare losses is influenced by societal preferences for redistribution in our framework.

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:

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

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 $\hat{E}\left[Y^*_{t=0}\right]$ 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 differences 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, $E\left[\tilde{Y} - Y^*|G=1\right] \neq E\left[\tilde{Y} - Y^*|G=2\right]$. 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 us 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 $\left(E\left[\tilde{Y}_{t=0}|G_a=1\right] < E\left[\tilde{Y}_{t=0}|G_b=2\right]\right)$ 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^*, Δy_b, 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 regarding 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.

3. Data

We examine PMTs constructed using experimental data from Tanzania and Malawi. These data have been employed in previous PMT studies (McBride and Nichols, 2018; Gazeaud, 2020), allowing us to

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

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 the period of the year data were collected. *Recall* and *diary* refer to the consumption data collection module. (n=11280 in Malawi; n=4032 in Tanzania).

benchmark our results against theirs and to rely on a predefined set of PMT variables. Moreover, both hypothetical case studies offer insights that let us explore label bias and the instability of PMT weights.

3.1. Malawi

The 2004/5 Second Integrated Household Survey comprises 11,280 households. The survey spanned 12 months, during which enumerators interviewed one enumeration area per month in randomly selected areas.[4] We leverage the staggered data collection to examine how the timing of data collection influences screening outcomes. Data collection occurred during both the lean (October–March) and harvest seasons (April–September). The proportion of interviews conducted during both periods is balanced, and we observe no significant differences in time-invariant household characteristics between households surveyed in each period. Please refer to Annex A for summary statistics of all PMT variables, consistent with the approach in McBride and Nichols (2018).

In Table 1, we summarize the poverty headcount for smaller and larger households by data collection season. Initially, 65% reported per capita consumption below the consumption poverty line, with poverty rising from 61% in the harvest period to 69% in the lean period. Next, we define smaller households using the median household size of four members (mean is 4.5). Poverty is more prevalent among larger households (80%) than smaller households (52%). For smaller households, poverty increases by roughly 11 percentage points (pp) between the harvest and lean seasons. In contrast, for larger households, this increase is only 7pp, suggesting the relative seasonal change is more pronounced for smaller households.

3.2. Tanzania

Data were obtained from the Survey of Household Welfare and Labour in Tanzania project. This project experimentally tested and compared the consistency of consumption reports using various household survey modules. The survey covered all 4,029 households and included a consumption experiment comprising eight different consumption questionnaire treatments, each randomly assigned to roughly 500 households. These treatments altered the approach (either recall or diary) and the duration of recall (ranging from 7 days to 12 months). The modules were: (a) a long list of items with a 14-day recall, (b) a long list of items with a 7-day recall, (c) a subset list with a 7-day recall, (d) a collapsed list with a 7-day recall, (e) a long list of items with monthly recall, (f) a 14-day household

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

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

diary with frequent visits, (g) a 14-day household diary with infrequent visits, and (h) a 14-day personal diary with frequent visits.

For clarity, we categorize treatments into diary and recall modules but also analyze results separately for each treatment. The experiment spanned seven districts in Tanzania from September 2007 to August 2008. A multistage sampling strategy was employed, selecting villages based on probability-proportional-to-size. Sub-villages were chosen randomly, and within these sub-villages, three households each were randomly assigned one of the eight modules. The experiment's results allow insights into the severity of issues arising from different sources of nonrandom error in consumption measurement (Beegle et al., 2012; Caeyers et al., 2012).

In our analysis, we use the same PMT input variables as Gazeaud (2020). Please refer to Annex B for a list accompanied by summary statistics. In Table 1, we highlight the poverty headcount for smaller and larger households and the method (recall or diary) used to gather consumption data. In total, 41% of households reported per capita consumption below the $1.25/day poverty line. However, for diary-recorded consumption, the figure is 37%, as opposed to 44% from recall data. This difference is due to diary entries typically registering higher values since recall methods can underestimate actual consumption. This deviation is more pronounced among larger households (those with over five members) at 8pp, compared to smaller households (five or fewer members) at 3pp.

Among the smaller households, only 28% fall beneath the poverty line, but this figure jumps to 59% for larger households. This disparity becomes even more stark depending on which consumption module is used: recall methods produce higher reported poverty rates with a 7pp gap compared to diary-based modules. Intriguingly, this difference is primarily attributed to larger households, where the gap between recall and diary methods swells to 12pp. This notable discrepancy is likely due to respondents in larger households being less informed of all consumption activities, resulting in underreporting.

4. Prediction Model and Targeting Errors

In line with current PMT practices, we first construct models to predict household consumption and then classify households as poor based on the predicted consumption. This approach mirrors current PMT procedures, even though directly training models to classify poor households might be more straightforward. We tested various specifications across different model classes but centered our discussion on a simple linear model as a benchmark and a gradient boosting model that performed best. In the linear regression model, all standardized PMT variables are used as inputs. Subsequently, we employ the xgboost library to train a gradient boosting model. Details of the parameter tuning process can be found in Annex C.

In practice, PMT scores are often estimated and validated using the same data. This can lead to model overfitting, which in turn can produce inaccurate out-of-sample predictions. To mitigate this risk, we separate training data (N*0.8) and a test set (N*0.2). Our preferred model specifications are chosen by examining how much of the consumption variation (with R2 as the performance metric) the model accounts for, utilizing 10-fold cross-validation. Households in the test set are classified as poor if their predicted consumption is below the poverty threshold.

Figure 1 showcases the prediction results for the test data. For clarity, prediction results are illustrated graphically through multidimensional scaling, representing the similarity of households based on PMT variables in two dimensions. Households with analogous PMT variables are adjacent in this scatter plot. Marker colors indicate if a household is classified as poor (orange) or non-poor (blue). The first row of Figure 1 represents actually measured poor households, while rows 2 and 3 depict predicted poor households from the linear and xgboost models, respectively. The first column presents results for Malawi, and the second for Tanzania.

The data indicate that poverty distribution concerning PMT variables is diverse. Many households differ in their poverty status but share similar PMT variable values, emphasizing the challenge of distinguishing the poor from the non-poor using available PMT variables. As anticipated, the

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

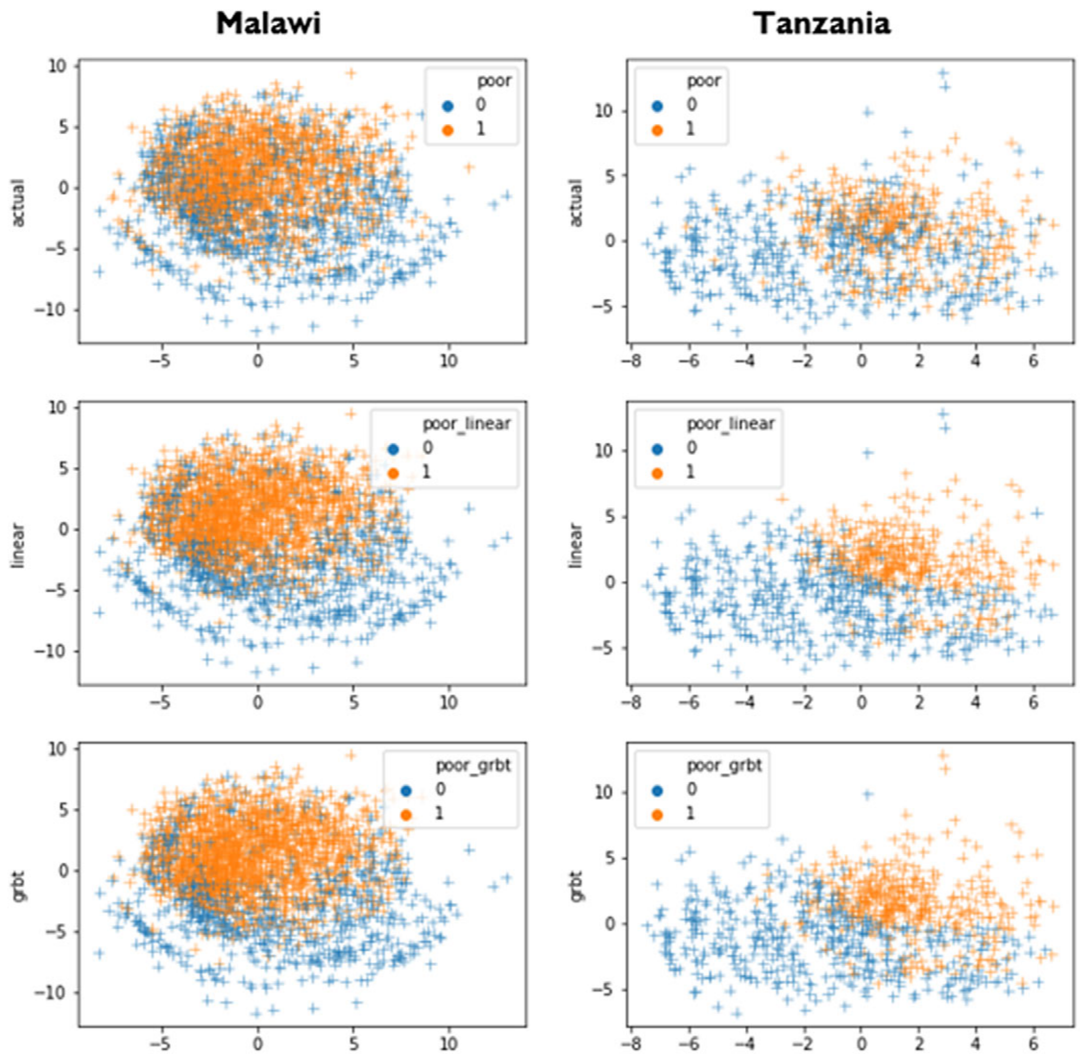

Figure 1. *Actual (first row) and predicted (second and third rows) poor households.*
Notes: *x and y axis show two-dimensional test data of PMT variables: first PMT variables were standardized and thereafter rescaled using multidimensional scaling. Actual refers to true (consumption) poverty status and linear and grbt refer to predicted poverty status with linear and xgboost model.*

Table 2. *PMT and test data used for assessment comparisons*

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

Note: Lean and harvest refer to the period of the year data were 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 which consumption diaries were used or data were collected during the harvest period.

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

Table 3. *Confusion matrix Tanzania*

		Predicted poor	Predicted non-poor
All (diary+recall) (n=805)	Poor	32%	11%
	Non-poor	15%	42%
Diary PMT model (n=300)	Poor	25%	14%
	Non-poor	14%	47%
Recall PMT model (n=300)	Poor	31%	8%
	Non-poor	22%	39%

Note: All predictions based on the xgboost model. *All* refers to the 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. n refers to the sample size of test data.

Table 4. *Malawi confusion matrix*

		Predicted poor	Predicted non-poor
All (n=2255)	Poor	57%	7%
	Non-poor	13%	24%
Harvest model (n=1125)	Poor	52%	8%
	Non-poor	12%	28%
Lean model (n=1125)	Poor	56%	4%
	Non-poor	20%	20%

Note: All predictions based on the xgboost model. *All* refers to the 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.

xgboost model outperforms the linear model in explaining consumption variance, achieving an R2 of 0.62 in test data compared to 0.58 of the linear model for Malawi. Variation in R2 across the 10 folds is relatively low, with a standard deviation of 0.02. Regarding the classification of poor households, the xgboost and linear models correctly classify approximately 81 and 80% of Malawi households, respectively. For Tanzania, R2 values stand at 0.56 and 0.54 for xgboost and linear models, leading to 75% accurate classifications for both. The standard deviation of R2 across the 10 folds is 0.04.

However, despite respectable prediction accuracy, around 7% of the poor households in Malawi data are not identified as such (exclusion error), while 13% of non-poor households are wrongly tagged as poor (inclusion error). For Tanzania, the errors stand at 15% inclusion and 11% exclusion, consistent across both model types. An overview of the classification results for the xgboost model is presented in Tables 3 and 4. Even with similar aggregate performance measures between the xgboost and linear models, the classification of 2.2 and 3.6% of households in Tanzania and Malawi changes depending on the applied model.

In conclusion, the xgboost model slightly outperforms the linear model in predicting consumption. The only marginal performance difference in classifying poor households might stem from the preselected input variables that work well with linear models and because the model was not specifically designed to categorize households but to predict consumption. In both scenarios, predictions are more clustered than the actual distribution of poverty, suggesting that errors in targeting are more likely for specific PMT variable combinations. Subsequently, we compute welfare losses and explore how much they are influenced by data biases.

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

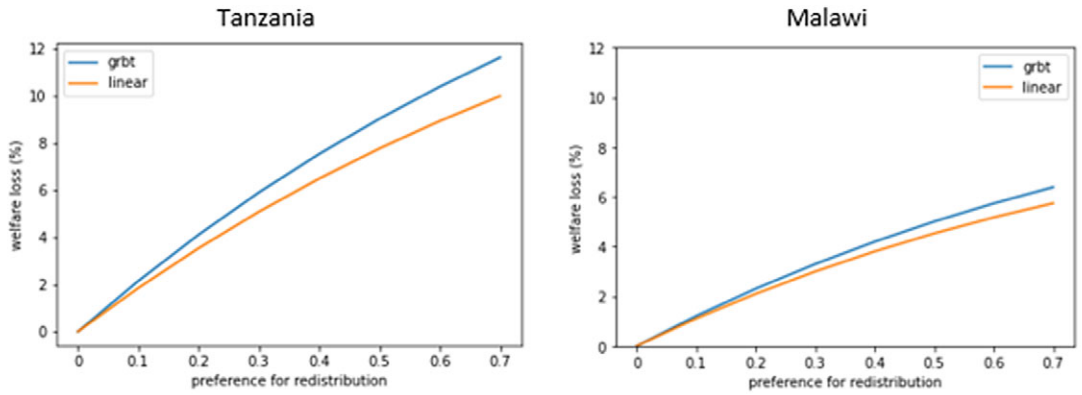

Figure 2. *Marginal welfare loss of unit transfer with linear and xgboost model.*
Notes: *Welfare loss computed as percentage change in welfare in comparison to perfect targeting assuming different levels of inequality aversion.*

4.1. Social welfare loss

For the social welfare loss assessments, we use the test data to simulate an anti-poverty policy, allocating a fixed budget to households predicted as poor based on the PMTs.[5] In line with common practice, the transfer size is uniform for all beneficiaries. The transfer amounts dispensed by each PMT method (linear, xgboost, or perfect targeting) are adjusted to fit within the fixed budget. This means that if a model overpredicts poverty, the transfer size per beneficiary will be reduced compared to a more accurately calibrated model. This consideration is crucial, as traditional targeting assessments often overlook the budgetary implications of overpredictions or the costs at which certain prediction accuracies are attained.

4.2. Welfare losses due to targeting errors

Figure 2 illustrates the marginal welfare loss when transfers are allocated through either the linear or xgboost model, compared to a perfect targeting benchmark. In the simulations, the fixed transfer budget is distributed to (predicted) poor households using our algorithms and is then compared to the benchmark scenario. In this benchmark, transfers are given to all (actual) poor households without any targeting errors. In the case of perfect targeting, each poor household receives a one-unit transfer.[6]

We compute the welfare loss, considering different levels of inequality aversion. If societies have no preference for redistribution (ρ=0), there is no welfare loss from targeting errors according to our framework. In this context, whether a rich or poor household receives a transfer has no impact on social welfare. As inequality aversion increases, welfare losses raise because societies emphasize the exclusion of poorer households more than inclusion errors. As a result, identical prediction accuracy can lead to varying welfare losses, depending on where in the welfare distribution errors occur.

Unsurprisingly, the findings indicate that welfare losses are more significant in Tanzania than in Malawi due to differences in prediction accuracy. However, it is surprising that in Malawi, the welfare loss with the xgboost model exceeds that of the linear model, despite the former model’s marginally superior classification accuracy. While the xgboost model excels in predicting consumption and classifying poor households, its welfare losses surpass those of the linear model. This implies that the linear model results in reduced welfare losses, outperforming the xgboost model in this context when positive welfare weights are applied. These welfare losses stem from the biases in benefit allocation and the transfer size, determined by the fixed budget and the count of predicted beneficiaries.

[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.

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

The transfer size significantly affects redistribution levels and, consequently, welfare losses. If, instead of allocating transfers to all predicted poor households, we give benefits to a set proportion of households with the lowest welfare scores, we can eliminate the notion that differences in transfer sizes contribute to welfare loss disparities. The outcomes of this analysis are depicted in Figure A3 in Annex C, confirming that transfer size primarily accounts for the gaps between the xgboost and linear models.

Finally, we focus on the distribution of welfare losses, specifically investigating how label bias and unstable predictors contribute to the systematic underestimation of welfare losses.

4.3. Welfare loss due to measurement errors

As outlined in the conceptual framework, we identify measurement error and the temporal instability of PMT weights as potential drivers of welfare losses. In this section, we assess the extent to which these factors contribute to actual welfare losses. Using the experimental data from Tanzania, we address welfare losses due to label bias by examining consumption data collected with recall versus diary consumption measurement methods. Therefore, we trained two distinct PMT models: one using recall data and the other using diary consumption data. Both models are then validated using the same test data; that is, we classify the same households with PMTs built from either diary or recall data. Any disparities in classification between the PMTs can thus be attributed to variations in PMT weights.

Similarly, with the staggered data collection in Malawi, we consider welfare losses due to PMT instability by applying two algorithms—one trained with lean season data and the other with harvest season data—to the same test data. In addition, to elucidate the underlying mechanisms, we demonstrate that our PMTs tend to disadvantage smaller households (refer to Annex C for a comprehensive overview of feature importance and coefficients in the models).

Table 2 details the distinct training data used to create the PMT pairs and the test data used to validate these models. It is crucial to emphasize that we consistently use the same households to validate and compare the PMT pairs. For consumption measurement, we rely on diary test data as a benchmark, and for seasonal stability, we use harvest season data. We subsequently focused solely on the xgboost model, but the results were qualitatively similar when using the linear model.

4.4. Label bias

Household consumption reports are often viewed as unbiased proxies for actual consumption. Nonetheless, recall bias can skew these reports, and this distortion has been found to be more pronounced in larger households where individual respondents might not accurately represent the consumption of other household members (Gibson and Kim, 2007; Beegle et al., 2012). To gauge the extent of this potential distortion in poverty screening decisions, we delve into the experimental data from Tanzania, similar to Gazeaud (2020). In our primary analysis, we differentiate between data gathered using consumption diaries and recall methods. We regard consumption diaries as a more accurate measurement technique, with individual consumption diaries under frequent supervision being seen as the gold standard approach (Beegle et al., 2012; Caeyers et al., 2012; Gazeaud, 2020). We construct two PMT models, each trained exclusively on either recall or diary consumption data. Subsequently, we use diary consumption test data to validate and contrast the two PMTs. The variation in welfare losses between the PMTs serves as an indicator of the welfare loss attributed to consumption measurement errors.

Table 3 displays the confusion matrix for the comprehensive model and the outcomes using separate PMTs for recall and diary data. (For a detailed view of accuracy, precision, and recall of all models, refer to Annex C.) Households are more likely to be predicted as poor if the PMT is based on recall data rather than diary data. The forecasted poverty rate in the (identical) test data stands at 53% with recall and 39% with the diary PMT, a difference likely stemming from underreporting in recall modules (Beegle et al., 2012; Caeyers et al., 2012; Gazeaud, 2020). Consequently, many households that are not poor were classified as such by the recall PMT, leading to a higher rate of inclusion errors. Conversely, exclusion errors were reduced with the recall PMT compared to the diary PMT. On the whole, the precision of both the diary and

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

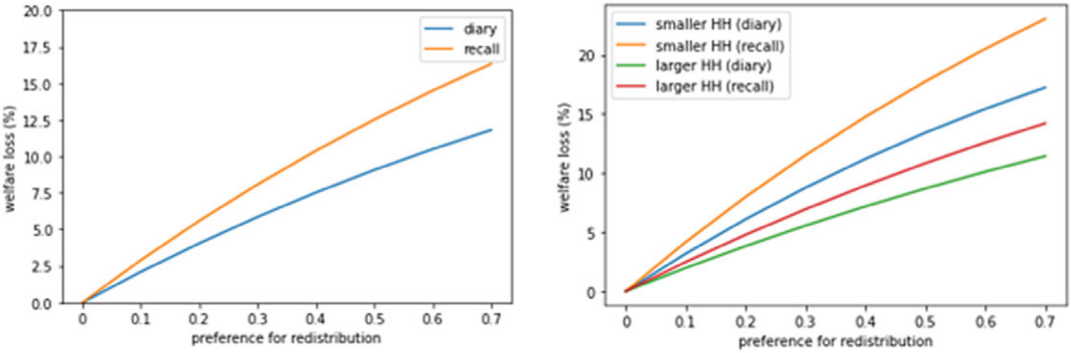

Figure 3. *Welfare loss depending on consumption measurement module and by household size.* Notes: *Welfare loss computed as percentage change in welfare in comparison to perfect targeting assuming different levels of inequality aversion. Evaluation based on the same test data, but models were trained with data only including recall or diary data.*

recall PMTs is fairly comparable, tallying up to 72 and 70%, respectively. However, these metrics overlook the distributive implications and associated welfare losses.

The left side of Figure 3 presents the simulated welfare losses for both PMT estimations. It reveals that welfare losses are notably greater for the model trained using recall data. The difference is as much as 5pp, representing nearly 30% of the total welfare loss. This indicates that, when utilizing a PMT trained with recall data, we underestimate the actual welfare losses by roughly one-third. If we exclusively employ data from the diary treatment with high supervision frequency, which is considered the gold standard in the literature, the discrepancy widens, accounting for almost half of the welfare loss.[7]

Why is the welfare loss so significantly underestimated? While the accuracy of both models is similar, the recall PMT tends to overpredict poverty. This means that the transfer size is reduced, resulting in diminished levels of redistribution. Consequently, the beneficiaries chosen under the diary PMT receive larger transfers. As these beneficiaries are more likely to belong to the poorest segments, welfare losses with the diary PMT decrease as the preference for redistribution rises.

In the right panel of Figure 3, we further dissect the results into smaller and larger households (determined by the median household size). As observed earlier, welfare losses are greater for the PMT trained with recall data. However, this difference amounts to about 8pp for smaller households, while for larger households, the discrepancy is less than 3pp. This suggests that consumption measurement error results in a substantial underestimation, approximately 40%, of welfare losses among smaller households. This rate of underestimation is more than double that of larger households, indicating that the welfare losses caused by consumption measurement errors predominantly disadvantage smaller households.

Why are smaller households more affected by consumption measurement errors? According to the original data experiment findings, recall bias is more pronounced in larger households, leading to an underestimation of actual consumption. This bias is reflected in the PMT, meaning that predictors related to household size are underestimated. This results in social welfare losses as smaller households with the same consumption level are less likely to be selected by the PMT. Our choice to compare smaller and larger households is based on existing literature that has highlighted consumption reporting biases. In Annex C, we explore other potential group dynamics, including the age of the household's head, urban versus rural households, and compare these results with PMTs trained without household size

[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 the 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).

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

information. In this context, the difference in welfare loss between larger and smaller households is most evident. However, in other settings, different group characteristics might have a greater impact.

Broadly speaking, this relates to the challenge of measuring household welfare and potential economies of scale from shared resources. In practice, per capita consumption has become the standard for poverty assessments, ensuring consistent results. However, alternatives that consider economies of scale in larger households exist and could have significant distributional implications (Jolliffe and Tetteh-Baah, 2022). To highlight this, we assess our results using a constant-elasticity scale adjustment of household consumption, as shown in Jolliffe and Tetteh-Baah (2022). Essentially, we divide total consumption by the square root of household size rather than just household size (i.e., per capita). To simulate the welfare losses, we adjust the poverty line, so that the poverty rate and the policy budget remain unchanged, but the allocation weights differ. The findings suggest that this alternative method increases welfare losses due to recall bias among smaller households, accounting for 80% of the welfare loss. For a detailed view, please refer to Annex D.

While it is beyond this paper's scope to determine whether per capita consumption is a better measure of welfare, it is important to note that the magnitude of economies of scale varies by context. Our analysis indicates that a seemingly straightforward choice can have substantial effects on allocation imbalances and consequent welfare losses, aspects that are often missed in targeting evaluations and method comparisons.

4.5. Data collection season

Stable predictors ensure that screening outcomes (and associated errors) remain consistent regardless of the screening's exact timing. Frequently, PMT weights are applied to data collected during a different period or year than the training data (see Brown et al., 2018). It is also recognized that in many contexts where PMTs are used, household consumption fluctuates significantly throughout the year (Hopper, 2020). Employing relatively stable household characteristics to predict a fluctuating target variable results in varying errors. To grasp the welfare implications, we utilize the staggered data collection from Malawi, applying a similar strategy to the previous section. We created two separate PMTs with data from the lean and harvest periods and validated both using the same harvest season test data. While we choose harvest season test data to ensure a "pure" validation set, unlike the case of consumption modules previously discussed, there is not a definitive reason to favor lean season data over harvest season test data. Nevertheless, the core message remains consistent.

Table 4 presents a breakdown of the prediction accuracy, as well as inclusion and exclusion errors, for the different PMTs. The model trained with lean season data overestimates actual poverty by 16pp. Consequently, the simulated transfer size per identified poor household is 4% greater with the harvest PMT compared to the lean season PMT. The accuracy level is marginally higher with the harvest season PMT (80%) than with the lean season PMT (77%), primarily because the harvest season PMT has fewer inclusion errors.

To understand the welfare implications, the left side of Figure 4 displays the simulated welfare losses for both PMTs. Contrary to the accuracy of classifications, welfare losses are notably higher for the lean season PMT, suggesting the harvest PMT might be more suitable in this context. Assuming a redistribution preference of 0.7, the lean season PMT underestimates the welfare loss by almost 5pp when applied during the harvest period.

The central reason for this result is the transfer size. Predicted poverty is 12pp higher with the lean season PMT than the harvest season PMT. As a result, the transfer size is about 15% lower. Despite fewer exclusion errors, the harvest season PMT assigns larger amounts to the extremely poor, which leads to more redistribution than with the lean season PMT.

The seasonal poverty dynamics are more pronounced among smaller households. Hence, the right column of Figure 4 indicates the most significant welfare loss difference for smaller households when using the lean season PMT. With the harvest PMT, the welfare loss disparity between smaller and larger households is not substantial. If the lean season PMT is applied during the harvest period, it would underestimate welfare losses by up to 5pp, with smaller households bearing a disproportionately larger share of the welfare losses.

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

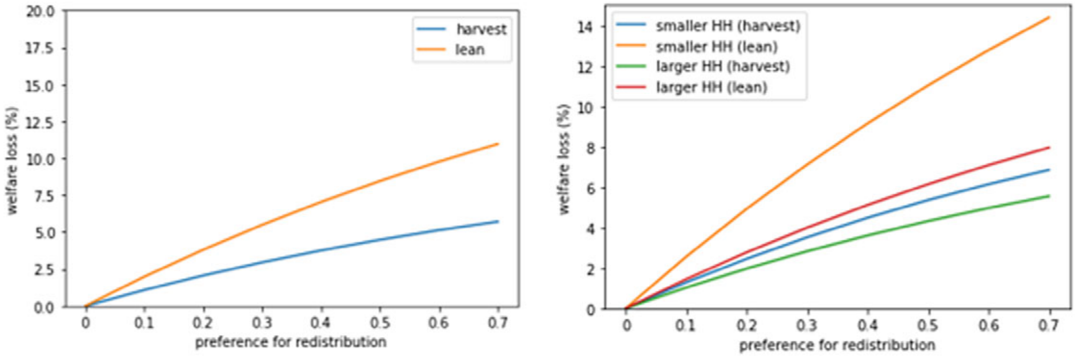

Figure 4. *Welfare loss distribution by household size.*
Notes: *Welfare loss computed as percentage change in welfare in comparison to perfect targeting assuming different levels of inequality aversion. Evaluation based on the same test data, but models were trained with data only including harvest or lean season data.*

Why is the welfare loss difference more significant for smaller households? The difference in both PMTs is more pronounced for smaller households, leading to distinct classifications in 16% of cases, compared to only 5% for larger households. Therefore, the PMT weights for smaller households are less consistent, and prediction errors increase when the PMT is applied to data from a different season.

Why are smaller households more affected by the timing of data collection? Household size is a significant predictor, and even a change of one member can alter classifications for smaller households. The predictor weight, as well as household size, is not constant. Difficulties in measuring household size are documented (Beaman and Dillon, 2012), and in many contexts, it can fluctuate even in the short term. For instance, monthly World Bank High Frequency Phone data in Malawi conducted during the COVID-19 pandemic suggest surprisingly volatile household sizes. 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 underline the challenge of predicting poverty's temporal dynamics with static input data, leading to potential underestimations in welfare losses, especially impacting smaller households.

5. Discussion

The COVID-19 pandemic has spurred 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 Program 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 propose that targeting error evaluations should consider the specifics of misclassified households more closely. Instead of, or alongside, conventional metrics, we employ a social welfare framework. This approach weighs targeting errors based on their position in the welfare distribution and varying levels of societal inequality aversion. Our extended framework underscores that heightened accuracy might actually lead to welfare losses when budgets are static. While offering a broader assessment of targeting performance, we demonstrate that data bias, particularly label bias and unstable PMT weights, can significantly underestimate welfare losses. The size of these often overlooked welfare loss components is alarmingly large, prompting deeper questions about the trustworthiness of targeting evaluations. We further establish that these unseen welfare losses disproportionately affect smaller households.

Our analysis concentrates on two bias sources: target variable measurement and weight stability. However, other biases might be present. Different circumstances might amplify the influence of other factors. Deliberate misreporting to game eligibility thresholds, inconsistent answers from individuals in the same household, or choices in survey sampling design and representation issues might introduce bias.

Our findings suggest that even a singular bias can result in notable underestimations of welfare losses. This is particularly striking when viewed through a social welfare perspective that specifically considers distortionary effects. Merely focusing on prediction accuracy overlooks these discrepancies, as it does not differentiate between inclusion and exclusion errors and neglects to consider the cost at which accuracy is attained (e.g., overprediction vs. underprediction of poverty). One limitation is that our approach requires making assumptions about societal welfare functions and inequality aversion. Moreover, other societal values, such as fairness and risk aversion, could further impact welfare losses, making our estimates at best partial. For instance, unduly affecting smaller households might be deemed unfair by many, leading to inherent welfare losses. While this paper emphasizes distributional implications, future research could explore fairness ratings, thereby highlighting the equity-efficiency trade-off in line with Premand and Schnitzer (2021), and potentially addressing citizens' legitimacy perceptions.

Our findings and subsequent conclusions prompt us to advocate for a broader discussion, eliminating opacity in decision-making and ensuring accountability and evaluation throughout a social protection program's lifecycle. Does greater awareness ensure fairness? Not necessarily. Much of the welfare losses we pinpointed come from data biases that often remain hidden. While some biases might be addressed, there's still a question about the use of predictions at all. Fairness is subjective, and any attempt to measure potential unfairness in prediction outcomes will always be incomplete. An alternative approach, discussed in prediction fairness conversations, might focus more on the causal mechanisms that necessitate these predictions in the first place.[9] A case in point is Kenya's Hunger Safety Net Program, which expands its provisions when remotely sensed drought indicators cross a threshold. This component aims to prevent households from slipping into poverty due to natural disasters, rather than retroactively determining poverty.

Finally, addressing data concerns is crucial. Many countries rely on household surveys to construct PMT coefficients for targeting beneficiaries. However, these coefficients often get applied to different datasets, such as registries or censuses, to actually decide a household's program eligibility. Our study underscores how using PMT with inconsistent data can amplify welfare losses. Hence, discussions about harmonizing household surveys and administrative data are imperative.

Acknowledgments. 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.

Funding statement. This work received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. The authors declare no competing interests exist.

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

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

Author contribution. Conceptualization: S.D., D.M., F.G.; Data curation: S.D.; Formal analysis: S.D.; Methodology: S.D., D.M.; Software: S.D.; Project administration: S.D.; Supervision: S.D.; Validation: S.D., D.M.; Visualization: S.D.; Writing—original draft: S.D., D.M.; Writing—review and editing: F.G., S.D., D.M.

Data availability statement. The data underlying this article are available in OSF (https://osf.io/nxqku/) DOI:10.17605/OSF.IO/NXQKU.

References

Adler MD (2019) Cost-benefit analysis and social welfare functions. In: White MD (ed), *The Oxford Handbook of Ethics and Economics*. Oxford, USA: Oxford University Press.

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

Aiken EL, **Bedoya G**, **Blumenstock JE and Coville A** (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 H, **Behrman JR and Tasneem A** (2019) The contribution of increased equity to the estimated social benefits from a transfer program: An illustration from PROGRESA/Oportunidades. *World Bank Economic Review 33*(3): 535–550.

Atkinson AB (1970) On the measurement of inequality. *Journal of Economic Theory 2*(3), 244–263.

Atkinson AB (2005) On targeting social security: Theory and Western experience with family benefits. In Van de Walle D and Nead K (eds), *Public Spending and the Poor. Theory and Evidence*. Washington DC: The World Bank.

Ayush K, **Uzkent B**, **Burke M**, **Lobell D and Ermon S** (2020) Generating interpretable poverty maps using object detection in satellite images. Preprint. arXiv:2002.01612.

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

Barocas S and Selbst AD (2016) Big data's disparate impact. *California Law Review 104*, 671–732.

Barrientos A, **Dietrich S**, **Gassmann F and Malerba D** (2022) Prioritarian rates of return to antipoverty transfers. *Journal of International Development 34*(3), 550–563.

Bastagli F, **Hagen-Zanker J**, **Harman L**, **Barca V**, **Sturge G and Schmidt T** (2019) The impact of cash transfers: A review of the evidence from low-and middle-income countries. *Journal of Social Policy 48*(3), 569–594.

Beaman L and Dillon A (2012) Do household definitions matter in survey design? Results from a randomized survey experiment in Mali. *Journal of Development Economics 98*(1), 124–135.

Beegle K, **De Weerdt J**, **Friedman J and Gibson J** (2012) Methods of household consumption measurement through surveys: Experimental results from Tanzania. *Journal of Development Economics 98*(1), 3–18.

Blumenstock JE (2016) Fighting poverty with data. *Science 353*(6301), 753–754.

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

Caeyers B, **Chalmers N and De Weerdt J** (2012) Improving consumption measurement and other survey data through CAPI: Evidence from a randomized experiment. *Journal of Development Economics 98*(1), 19–33.

Camacho A and Conover E (2011) Manipulation of social program eligibility. *American Economic Journal: Economic Policy 3* (2), 41–65.

Campo SD, **Anthoff D and Kornek U** (2021) Inequality aversion for climate policy ZBW - Leibniz Information Centre for Economics.

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

Coady DP, **D'Angelo D and Evans B** (2020) Fiscal Redistribution and Social Welfare: Doing More or More to Do? EUROMOD Working Paper.

Cooke IR, **Queenborough SA**, **Mattison EHA**, **Bailey AP**, **Sanders DL**, **Graves AR**, **Morris J**, **Atkinson PW**, **Trawick P and Freckleton RP** (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), 269–277.

Corbett-Davies S and Goel S (2018) The measure and mismeasure of fairness: A critical review of fair machine learning. Preprint. arXiv:1808.00023.

Creedy J (2006) Evaluating policy: Welfare weights and value judgements. University of Melbourne, Department of Economics: Research Paper 971.

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

Dwork C, **Hardt M**, **Pitassi T**, **Reingold O and Zemel R** (2012) Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. New York, NY, USA: Association for Computing Machinery, pp. 214–226. https://doi.org/10.1145/2090236.2090255

Ferrer X, **van Nuenen T**, **Such JM**, **Coté M and Criado N** (2021) Bias and discrimination in AI: A cross-disciplinary perspective. *IEEE Technology and Society Magazine 40*(2), 72–80.

Foster J, **Greer J and Thorbecke E** (1984) A class of decomposable poverty measures. *Econometrica: Journal of the Econometric Society 52*, 761–766.

Gajane P and Pechenizkiy M (2017) On formalizing fairness in prediction with machine learning. Preprint. arXiv:1710.03184.

Gazeaud J (2020) Proxy means testing vulnerability to measurement errors? *Journal of Development Studies 56*(11), 2113–2133.

Gentilini U, **Almenfi M**, **Orton I**, **Dale P** (2022) Social Protection and Jobs Responses to COVID-19 (No. 33635). The World Bank Group

Gibson J and Kim B (2007) Measurement error in recall surveys and the relationship between household size and food demand. *American Journal of Agricultural Economics 89*(2), 473–489.

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

Hopper R (2020) The Dynamics of Deprivation in Malawi: The Multi-Dimensional Effects of the Lean Season on Children. Available at: https://www.unicef.org/malawi/media/4551/file/Report: The Dynamics of Deprivation in Malawi.pdf.

Jolliffe D and Tetteh-Baah S (2022) Identifying the Poor - Accounting for Household Economies of Scale in Global Poverty Estimates. IZA Discussion Paper No. 15615, Available at SSRN: https://ssrn.com/abstract=4241594 or http://doi.org/10.2139/ssrn.4241594

Kind J, **Botzen WJW and Jeroen CJHA** (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 J, **Ludwig J**, **Mullainathan S and Obermeyer Z** (2015) Prediction policy problems. *American Economic Review 105* (5), 491–495.

Ledesma C, **Garonita OL**, **Flores LJ**, **Tingzon I, and Dalisay D** (2020) Interpretable poverty mapping using social media data, satellite images, and geospatial information. Preprint. arXiv:2011.13563.

Lum K and Isaac W (2016) To predict and serve? *Significance 13*(5), 14–19.

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

Mehrabi N, **Morstatter F**, **Saxena N**, **Lerman K and Galstyan A** (2021) A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR) 54*(6), 1–35.

Obermeyer Z, **Powers B**, **Vogeli C and Mullainathan S** (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*(6464), 447–453.

Okun AM (2015) *Equality and efficiency: The big tradeoff*. Brookings Institution Press.

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

Rambachan A, **Kleinberg J**, **Ludwig J and Mullainathan S** (2020) An economic perspective on algorithmic fairness. *AEA Papers and Proceedings 110*, 91–95.

Saez E and Stantcheva S (2016) Generalized social marginal welfare weights for optimal tax theory. *American Economic Review 106*(1), 24–45.

Sen A (1976) Poverty: An ordinal approach to measurement. *Econometrica: Journal of the Econometric Society 44*, 219–231.

Sen A (1977) On weights and measures: Informational constraints in social welfare analysis. *Econometrica: Journal of the Econometric Society 45*, 1539–1572.

Van der Pol T, **Bos F and Romijn G** (2017) Distributionally Weighted Cost-Benefit Analysis: From Theory to Practice (October 10, 2017). CPD Discussion Paper No. 364, Available at SSRN: https://ssrn.com/abstract=3090393 or http://doi.org/10.2139/ssrn.3090393.

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

Annex A. PMT Review

A case study review into transparency in PMT was undertaken in 2021, with East Africa selected due to its numerous programs relying on PMT for targeting and its relevance to this study. This review emphasized areas of methodological rigor and transparency, as identified from the research informing this note. Primarily, the review consulted official program documentation provided by the respective governments, though RCT reviews included documents from universities or consultancies. Given the significance of social protection programs for individual welfare and poverty reduction, only publicly available documents were considered. An email was also sent out soliciting additional public-domain documents and information from individuals or departments responsible for program implementation that might have been missed. The review's findings indicated that the targeting methodology was disclosed in roughly two-thirds of cases, but less than half employed even a basic estimation method. None of the reviewed programs had trained models for out-of-sample predictions. While about half of the key programs disclosed their PMT variables, less than two-fifths made their PMT variable weights public. In the context of this note's findings, this reflects a significant gap in program accountability and transparency. The fact that 36% of programs had a published RCT—excluding non-RCT-based evaluations—suggests that ex post impact assessments are of higher priority to policymakers than ex ante targeting evaluations.

Table A1. *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%)

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

Annex B. Summary Statistics

B.1 Malawi

Table A2. *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.	1,886.24	2,169.89	2,017.55	2,131.86	Music player	0.28	0.10	0.13	0.20
	(1,612.31	1,343.62	1,822.06	1,343.62		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.

B.2 Lean versus harvest season

The survey was conducted over a period of 12 months in 2004/2005 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 A3. *Lean season balance tests*

	Non-lean season	Lean season	Pr Chi2
No cement floor	4,527	4,509	0.92
Cement floor	1,127	1,117	
No electricity	5,321	5,299	0.86
Electricity	333	327	
No flushing toilet	5,503	5,459	0.34
Flushing toilet	151	167	
No grass roof	1,506	1,447	0.27
Grass roof	4,148	4,179	

Note: Chi2 test for differences in distribution of variables between lean and harvest season.

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

B.3 Tanzania

Table A4. *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	2,337.05	2,586.53	2,389.89	2,506.47	Improved roof	0.27	0.43	0.36	0.30
	(1,706.16)	(1468.07)	(1,879.16)	(1,468.07)		(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.

Annex 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 10,000 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 resulting from the xgboost and linear models.

Table A5. *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

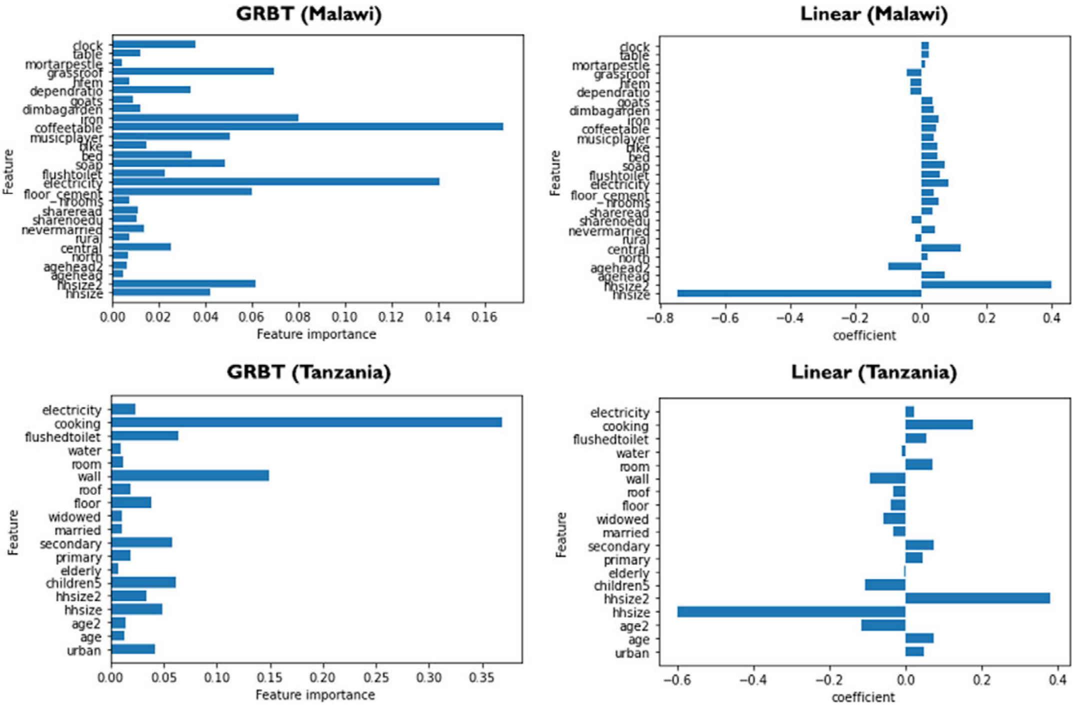

Figure A1. *Feature importance/coefficients.*

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

Table A6. *Performance summary of the PMT models*

	Accuracy	Precision	Recall
	Malawi		
All (n=2255)	80%	82%	89%
Harvest model (n=1125)	80%	82%	87%
Lean model (n=1125)	76%	74%	94%
	Tanzania		
All (n=805)	74%	69%	75%
Diary model (n=300)	72%	64%	64%
Recall model (n=300)	70%	58%	79%

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

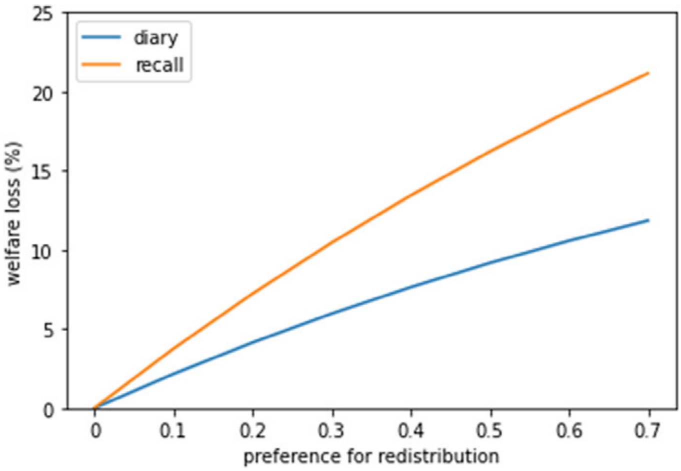

Figure A2. *Marginal welfare loss of unit transfer with diary and recall PMT (only using personal diaries with high supervision frequency treatment to train diary model and for model validation).*

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

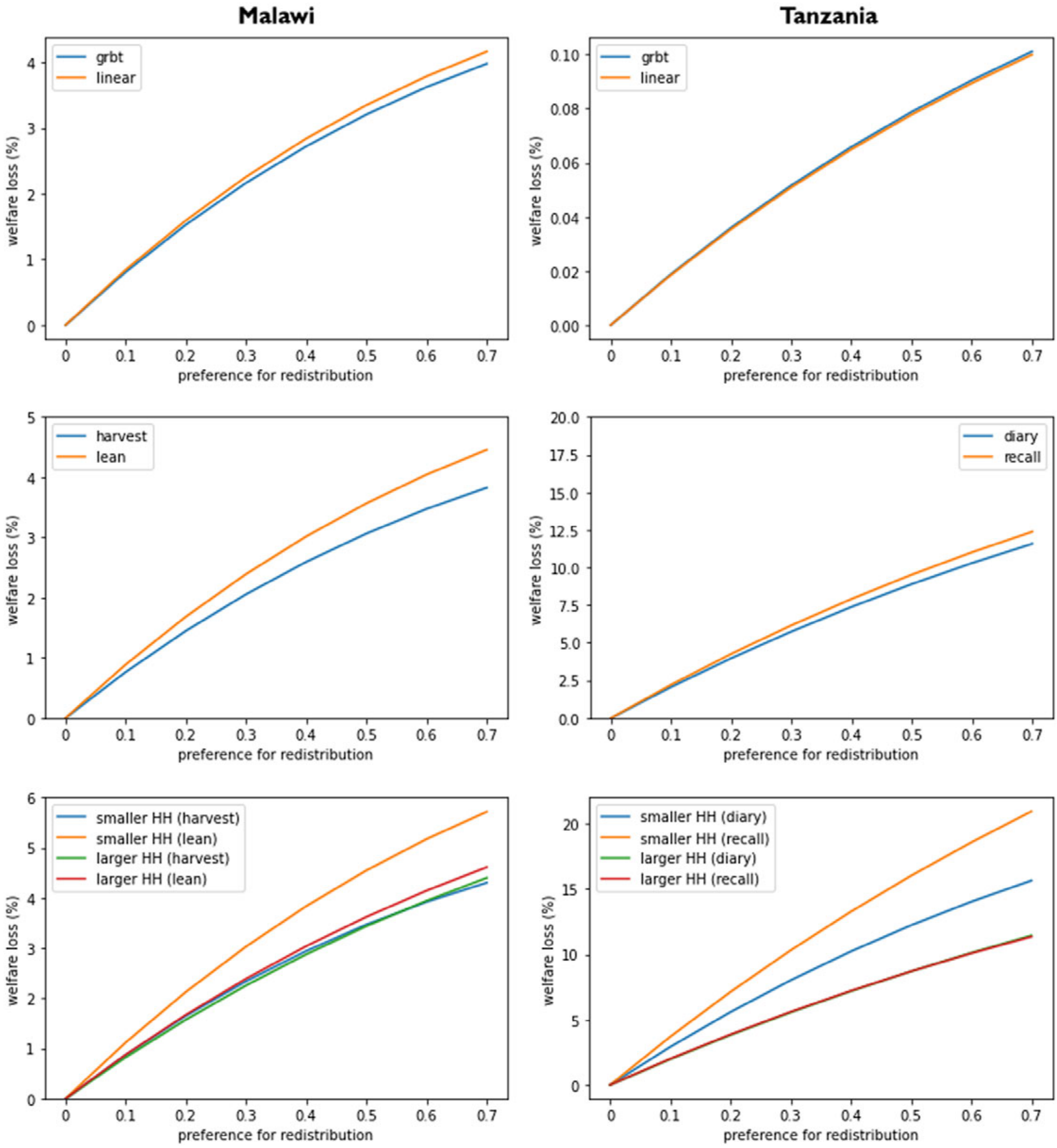

Figure A3. *Welfare loss predictions using a fixed quota instead of fixed poverty line approach.*

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

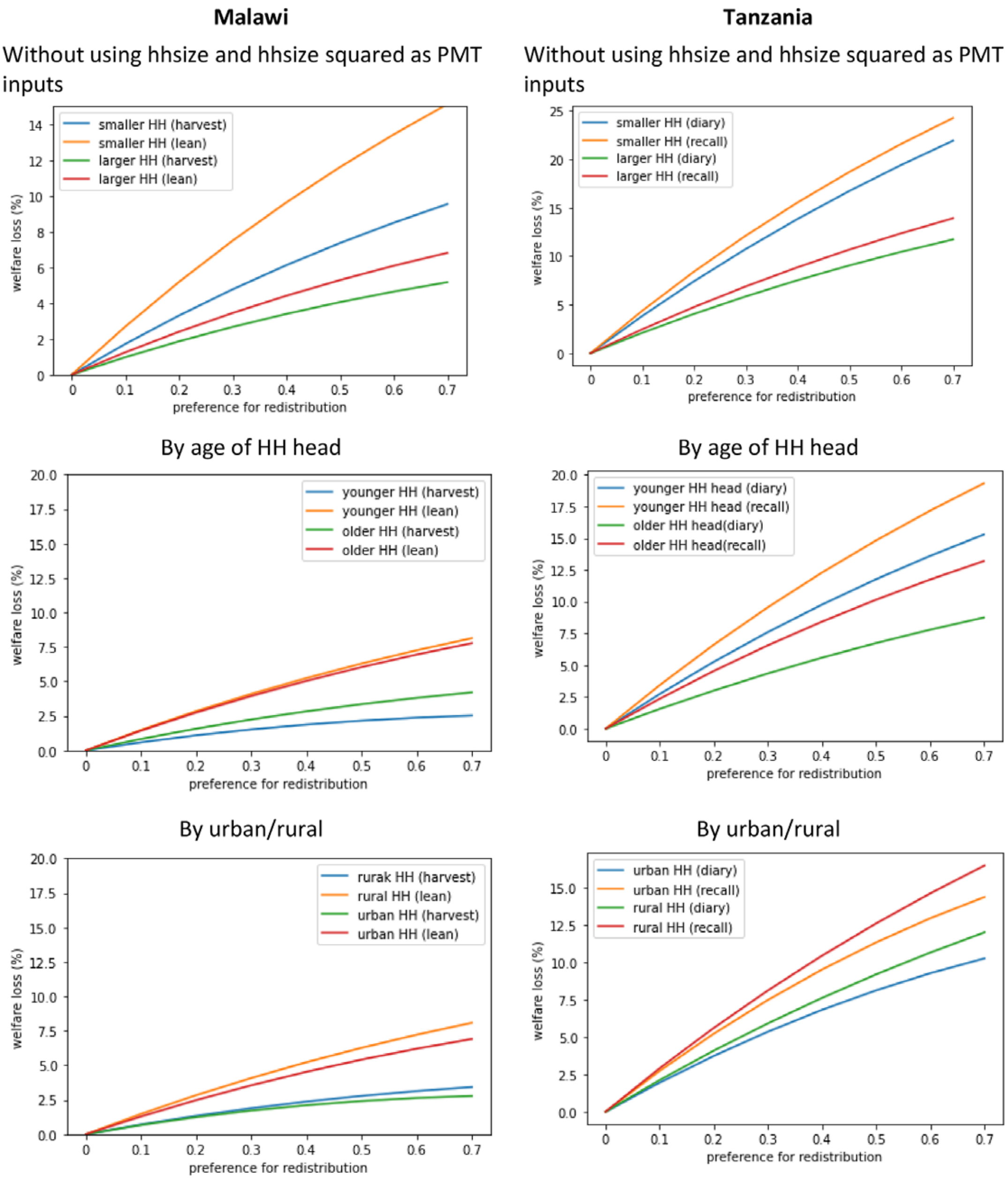

Figure A4. *Welfare losses after removing household size from PMT input list, by distinguishing between urban and rural households, and the age of the household head.*

Annex D. Per Capita versus Adult Equivalent Scale

To account for economies of scale within households, we follow Jolliffe and Tetteh-Baah (2022) 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, that is, 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 shows the resulting welfare losses of the two PMTs overall and separately by household size.

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

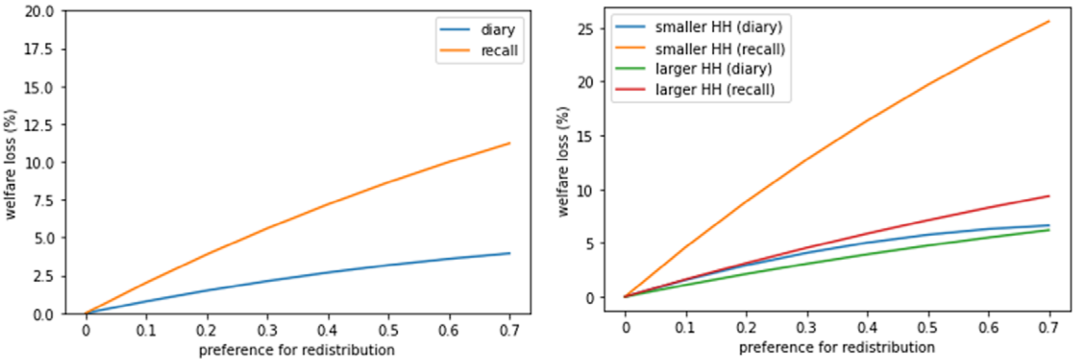

Figure A5. *Welfare losses if consumption is converted to account for household economies of scale.* Notes: *Welfare loss computed as percentage change in welfare in comparison to perfect targeting assuming different levels of inequality aversion. Evaluation based on the same test data, but models were trained with data only including recall or diary data.*

Annex 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 1-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, 9 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 A7. *Month-on-month variation in household size, Malawi phone survey*

		Household size m+1								
		1	2	3	4	5	6	7	8	9
Household size m	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	0.15

Source: World Bank's high frequency phone surveys.
Note: Resampling of household size is based on monthly phone survey data. Marker colors show standard deviation of original prediction minus prediction with resampled household size. Results based on Monte Carlo simulation with 1000 iterations.

https://doi.org/10.1017/dap.2023.38 Published online by Cambridge University Press

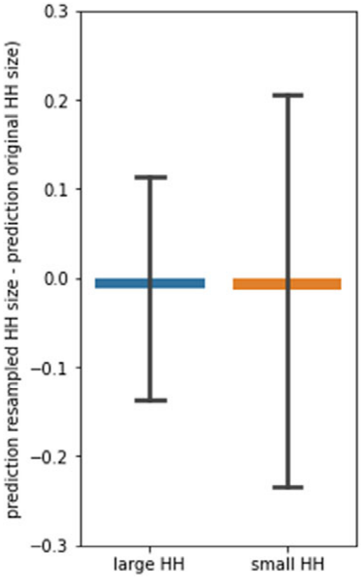

Figure A6. *Variance of predicted poverty due to household size sampling variability.*

Cite this article: Dietrich S, Malerba D and Gassmann F (2024). Predicting social assistance beneficiaries: On the social welfare damage of data biases. *Data & Policy*, 6: e3. doi:10.1017/dap.2023.38