What Do Policies Value?*

Daniel Björkegren† Columbia University

Joshua E. Blumenstock‡ U.C. Berkeley

Samsun Knight§ University of Toronto

September 2, 2025

Abstract

When a policy prioritizes one person over another, is it because they benefit more, or because they are preferred? This paper develops a method to uncover the values consistent with observed allocation decisions. We estimate how much each individual benefits from an intervention, and then reconcile the allocation with (i) the welfare weights assigned to different people; (ii) heterogeneous treatment effects of the intervention; and (iii) weights on different outcomes. We demonstrate this approach by analyzing Mexico's PROGRESA anti-poverty program. The analysis reveals that while the program prioritized certain subgroups — such as indigenous households — the fact that those groups benefited more implies that the program did not actually assign them a higher welfare weight. We also find evidence that the policy valued outcomes differently from households. The PROGRESA case illustrates how the method makes it possible to audit existing policies, and to design future policies that better align with values.

JEL classification: I38, Z18, H53, O10

Keywords: targeting, welfare, heterogeneous treatment effects

*We thank Luk Yean and Jolie Wei for excellent research assistance, and Yassine Sbai Sassi and Demian Pouzo for help with econometrics. Thank you to Joseph Cummins, Brian Dillon, John Friedman, Ted Miguel, Teddy Mekonnen, Sendhil Mullainathan, Paul Niehaus, Jonathan Roth, Yang Xie, seminar audiences, anonymous referees, and the editor for helpful suggestions. We thank the JP Morgan Chase Research Assistant Program at Brown University for financial support.

1 Introduction

The values behind policy decisions are not always transparent. When governments decide which households receive welfare benefits, or universities select which students to admit, they do not always articulate a rationale behind those decisions. Even when a rationale is given for a policy, it may be difficult to verify. In particular, certain people may be prioritized either because they are expected to benefit the most from the policy, or because they are favored, irrespective of how much they are likely to benefit. This distinction has important implications (Nichols and Zeckhauser, 1982; Coate and Morris, 1995): all members of society may agree on a ranking of who benefits most along some objective metric, but may disagree on how much welfare weight to assign to different entities.

This paper develops a method to infer social preferences that are consistent with observed or proposed policies. This method involves first obtaining estimates of heterogeneity in treatment effects (who benefits the most), and then, in a second stage, separating those from implied welfare weights (who is valued) and how different outcomes are valued, given the policy's allocation. This approach makes it possible to shift the debate from one about means — who should receive what — to one about ends: what are the impacts we desire, and which populations are most important?

We consider a common form of policy, in which some treatment is allocated based on a score or ranking. The allocation could be based on poverty scores in the case of welfare programs, or explicit rankings in the case of applicants for college admission or small business grants. We show that the ranking implies a set of inequalities that can be used to back out the implied value that it places on different outcomes and different entities. Our method can also be used if one only observes the binary decision of who is eligible and who is not.

Intuitively, if a policy allocates benefits to one type of entity who benefits little from the allocation, rather than to a different type that benefits greatly, that suggests the policy implicitly places higher welfare weight on the first type. Or, if a policy consistently allocates to applicants whose health improves as a result of the intervention — instead of applicants whose consumption increases — that implies the policy implicitly highly values health.

To illustrate how this method can be used to interrogate a real-world policy,

we apply it to historical data from PROGRESA, one of the world's largest (and best-studied) anti-poverty programs. We first estimate the heterogeneous treatment effects of the program. Consistent with prior work, we find evidence of treatment effect heterogeneity — for instance, that indigenous households benefit most from the program (cf. Djebbari and Smith, 2008). Our main estimates use OLS but we also demonstrate alternative methods for estimating treatment effects (Wager and Athey, 2018).

We then use our method to estimate the preferences consistent with the observed ranking of households and its heterogeneous effects on consumption, child health, and school attendance. We find that indigenous households were more likely to be allocated the program, but because they benefit so much more, the policy does not actually implicitly place higher welfare weight on them, and if anything is consistent with assigning them *lower* welfare weights than non-indigenous households. Our results also suggest that the program's design is consistent with assigning extra value to poorer, larger, and less educated households. These valuations, estimated using our method, are similar to the stated preferences of Mexican residents, as measured by hypothetical allocation questions in a survey we conducted in 2023. We additionally recover estimates of how the policy implicitly values impacts on consumption, health, and schooling. While a utilitarian policy would defer to the choices made by households, a paternalistic policy may attempt to override these preferences — if, for example, it preferred that parents made different choices for their children. Our estimates strongly reject non-paternalism, suggesting the policy values these outcomes differently from household decision makers. This preference for paternalism is echoed in the responses of Mexican residents.

Our final set of empirical results illustrate how this approach can further be used to evaluate counterfactual policies and preferences. In the PROGRESA case, we show what *would have occurred* had the program designers placed higher value on certain types of impacts (e.g., health vs. education) or certain types of households (e.g., equal welfare weights). This analysis suggests that, for instance, a policymaker who cared exclusively about impacts on schooling should prefer a policy that prioritizes richer households; a policymaker that valued only consumption impacts would instead prioritize indigenous households. More broadly, we show where these counterfactual

policies lie relative to the Pareto frontier that characterizes improvements across the three focal welfare outcomes.

After presenting the empirical results, we discuss more general settings where our approach may be useful. This framework can be used retrospectively, to audit existing programs and elucidate the values they imply, thereby facilitating more critical discussion of implemented policies. However, it can also be used prospectively, to help ensure that future policies better reflect the preferences of policymakers and constituents, thus providing a sort of decision aid to imperfectly rational policymakers. We demonstrate both uses in the case of PROGRESA. In both settings, the main requirements are that (i) there exists a way to estimate how different entities would be affected by the policy, and (ii) that the policy designer can articulate which household characteristics should be permitted to influence preferences. The former is a practical issue: treatment effect heterogeneity is most easily estimated when a randomized control trial facilitates impact evaluation on a subset of the population, as might occur with a pilot study, but could in principle be obtained through non-experimental approaches (e.g., Kent et al., 2020; Johansson et al., 2018). The latter is more subtle, as it entails considerations both theoretical (e.g., the values of constituents) and empirical (i.e., to permit identification). In particular, the full application of our method requires an exclusion restriction that there exist characteristics that describe heterogeneity but which do not directly enter the preferences of the policy, though we show variants of the method that do not require an exclusion restriction.

Taken as a whole, this approach makes it possible to invert the discussion about policies and programs. Rather than debate the means of the policy (who is eligible, how large are the benefits?), this framework makes it possible to debate the ends (how much do we value health, education, or consumption? Should poor families be prioritized over middle class families?). The framework can be applied to a wide range of settings where policymakers allocate scarce resources and heterogeneous treatment effects can be estimated.

Related Literature

This paper contributes to literature on optimal targeting and taxation (Nichols and Zeckhauser, 1982; Barr, 2012; Fleurbaey and Maniquet, 2018), including work

comparing targeted policies to universal basic income (Alatas et al., 2012; Hanna and Olken, 2018). It can be viewed as a response to Ravallion (2009), which argues that targeting poverty directly may not be sufficient for impact, and suggests that it may be better to target based on desired outcomes. In that sense, our work relates closely to Haushofer et al. (2022), which asks how targeting on treatment effects compares to targeting on baseline poverty. Their empirical analysis suggests that those who are most impacted by a Kenyan cash transfer are not always the poorest. Our paper focuses on the inverse problem of estimating the welfare function consistent with an observed policy. The two approaches are thus complementary; ours also extends from a specified utility function defined over a single outcome to a general welfare function that can rationalize targeting based on household characteristics as well as impacts on multiple outcomes. Our empirical results also engage with research on the effects and allocation of cash transfer programs (Behrman and Todd, 1999; Skoufias et al., 2001a; Gertler, 2004; John Hoddinott, 2004; Coady, 2006; Djebbari and Smith, 2008; Alderman et al., 2019). We build on this work by showing how effects can be used to audit policymaker priorities, and improve the design of future policies.

Our approach also relates to a growing literature that takes a given welfare function as fixed, and considers what are the best decisions to take. Kitagawa and Tetenov (2018) computes optimal assignment of treatment with experimental data, and Athey and Wager (2020) with observational data. Gechter et al. (2019) assesses how well different ex ante treatment assignments maximize a given welfare function under ex post experimental data. Wang (2020) considers the theoretical problem of allocating resources given heterogeneous aid agency preferences over individuals, and describes allocation queues as a solution to a combinatorial problem. This literature faces a central problem: what notion of welfare do, or should, societies maximize? Our paper takes a step towards answering this question, by solving the reverse problem: estimating welfare functions consistent with observed decisions.

It is increasingly common to construct indices summarizing multiple outcomes as a more nuanced measure of welfare (Greco et al., 2019). A persistent question in assembling these indices is what weight to apply to each component. These weights have economic meaning: how valuable is one component relative to another? Common approaches are geometric: setting equal values to each component (UNDP, 1990),

or analyzing how components vary together in observational data, using a principal component analysis (Filmer and Pritchett, 2001; McKenzie, 2005). We derive weights that have an economic interpretation using revealed preferences, how policies implicitly make trade-offs. A related approach is to set weights to optimally predict some gold standard measure of utility, if one is available (Jayachandran et al., 2021).

Also related is a recently expanding 'inverse optimum' public finance literature that estimates the redistributive preferences that are consistent with observed income tax policies. Bourguignon and Spadaro (2012) and Hendren (2020) infer the weight on different households implied by a tax schedule, based on the distortions required to transfer them resources. Saez and Stantcheva (2016) generalize welfare weights to reconcile popular notions of fairness with optimal tax theory. That literature considers tax policies that condition on a single covariate (pre-tax income) and affect a single outcome (net-of-tax consumption). Our paper generalizes this approach to arbitrary allocation policies that may condition on a vector of covariates and affect a vector of outcomes. This richer space allows us to back out additional information: how welfare weights depend on a vector of attributes, and the relative value placed on different outcomes (such as consumption, health, or education). It also shows how these welfare questions can be raised across a broad set of domains where heterogeneous treatment effects can be estimated.

More broadly, our efforts also connect with recent computer science scholarship on fairness in machine learning (cf. Dwork et al., 2012; Barocas et al., 2018). Several papers in this literature study the social welfare implications of algorithmic decisions, and how social welfare concerns relate to different notions of fairness (Ensign et al., 2017; Hu and Chen, 2018; Mouzannar et al., 2018; Liu et al., 2018). This relates to work on multi-objective machine learning (Rolf et al., 2020). Kasy and Abebe (2020) describe limitations of fairness constraints, and suggest that algorithms should be optimized for impacts. Also related, Noriega et al. (2018) discuss how different constraints to targeting can impact efficiency and fairness. Our approach is distinct, however, in that we show how using machine learning tools can be used to better characterize and audit the values consistent with a program's observed allocation. We hope that by providing increased visibility into these revealed preferences, future policies can be better aligned with stated preferences and explicit policy objectives.

2 Model

We consider the problem of allocating treatment among N entities, which could be, for example, households, individuals, firms, or regions. For convenience, we refer to entities as households.

A policy ranks each household i in the priority order they will be allocated some benefit or treatment, $T_i \in \{0, 1\}$. This ranking z_i may include ties between households; in the extreme it could simply represent the binary decision of whether household i will be allocated treatment ($z_i \in \{0, 1\}$).

We attempt to reconcile that ranking with an implicit welfare function

$$S = \sum_i S_i \tag{1}$$

$$S_i = w(\mathbf{x}_i) \cdot u_i(T_i)$$

where each household i is valued from the perspective of the policy according to some utility $u_i(T_i)$, scaled by some differential welfare weight $w(\mathbf{x}_i)$ based on its characteristics $\mathbf{x}_i$ (boldface indicates vectors, throughout).

The utility of household i from the perspective of the policy can be decomposed into components

$$u_i(T_i) = \sum_j b_{ij} v_{ij}(T_i) + a \cdot T_i \tag{2}$$

where v_{ij} represents the utility of household i arising from component j, and b_{ij} represents the implied value of that component. For simplicity, we here consider "non-choice" components of utility v_{ij}, where i does not directly choose their level of j (e.g., an immune system response to a vaccine). We will later generalize to "choice" outcomes over which i has some ability to influence the outcome (such as consumption and savings) in Section 3.4. We also allow treatment to provide some base value irrespective of its impact on outcomes, denoted by a.[1]

Imagine we knew the impact of treatment on household i's component of utility j: $\Delta v_{ij} \coloneqq v_{ij}(1) - v_{ij}(0)$. The welfare impact of treating household i could then be

[1]For intuition: if a is large in magnitude, the ranking between households is explained mostly by differences in welfare weights; if a is small or zero, the ranking depends also on impacts.

written

$$\Delta S_i = w(\mathbf{x}_i) \cdot \left(\sum_j b_{ij} \Delta v_{ij} + a \right) \tag{3}$$

If the cost of treating each household is the same, the ranking of each household, z_i, can then be reconciled with its implied welfare impact plus a shock ϵ_i, as long as there exists a weakly increasing function f that preserves the ranking of households,

$$z_i = f(\Delta S_i + \epsilon_i). \tag{4}$$

The shock may represent measurement error in estimates of welfare, or mistakes in the allocation.

2.1 Intuition

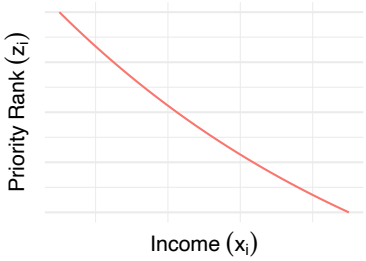
To demonstrate the intuition behind our method, we illustrate with a simple example in Figure 1. Consider the case of a single non-choice outcome and one dimension of heterogeneity, x_i, which corresponds to income. A policymaker allocates a program by ordering households by $z_i = Z(x_i)$, for some function Z that prioritizes poor households. As shown in Figure 1, depending on how treatment effects Δv_i vary with x_i, the same allocation could result from (1) higher welfare weights on the poor, (2) equal welfare weights, or (3) higher welfare weights on the rich. Likewise, in the case where x_i is binary, an allocation to one group can result from (i) higher welfare weights, if that group benefits the same or less; (ii) equal welfare weights, if that group benefits more; or (iii) lower welfare weights, if that group benefits much more.

The next section demonstrates how to empirically recover welfare and impact weights from data when there are multiple dimensions of heterogeneity and multiple outcomes of interest.

3 Estimation

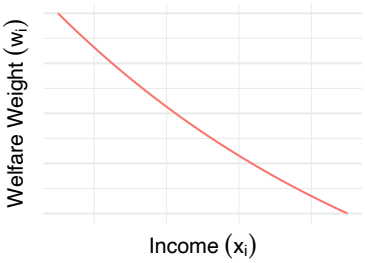
This section describes a procedure to estimate the model (the parameters defining objects Δv_{ij}, a, b_{ij}, and $w(\mathbf{x}_i)$ in equation (3)). We also discuss the conditions under which the parameters are identified.

An allocation rule that prioritizes the poor (low x_i)

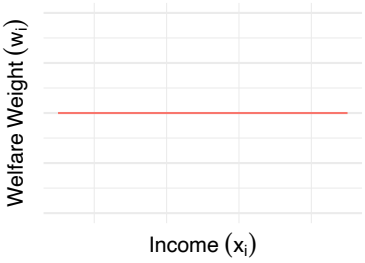


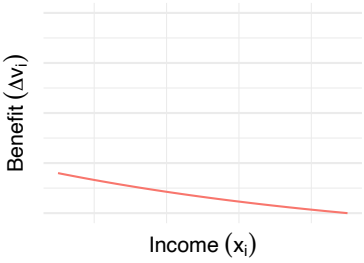
Could result from

(1) Higher welfare weight on the poor if treatment effects are constant



(2) Equal welfare weights on households if treatment effects are higher for the poor





(3) Higher welfare weight on the rich if treatment effects are much higher for the poor

Figure 1: Intuitive Example

3.1 Measurement

We assume that household i's utility from component j can be measured as a function of the observed outcome y_{ij}, i.e., $v_{ij} = g_j(y_{ij})$, with component utility function g_j. At its simplest this function may linear, $g_j(y) = y$, but one may also wish to incorporate diminishing returns, for example $g_j(y) = \log(y)$.[2]

3.2 Procedure

Estimation proceeds in two steps:

First, we obtain a prediction of the effect of treating each household i on each component of utility j. We postulate that the utility on outcome j arises from a process,

$$v_{ij} = v_j(T_i, \tilde{\mathbf{x}}_i) + e_{ij},$$

with some error e_{ij}. This allows treatment effects, $v_j(1, \tilde{\mathbf{x}}_i) - v_j(0, \tilde{\mathbf{x}}_i)$, to be heterogeneous as a function of potentially many covariates $\tilde{\mathbf{x}}_i$. We define the shorthand $\Delta\hat{v}_{ij} = \Delta\hat{v}_j(\tilde{\mathbf{x}}_i)$ to refer to the predicted treatment effect for household i. Heterogeneous treatment effects can be estimated using a variety of methods, including OLS or machine learning approaches that capture nuanced heterogeneity (Wager and Athey, 2018). We illustrate both of these approaches later.

Second, we estimate the preferences that would justify the ranking ($\boldsymbol{z}$), given the predicted effects of treatment on each household, $\Delta\hat{v}_{ij}$. If household i is prioritized over i' ($z_i > z_{i'}$), equation (4) implies

$$\Delta S_i + \epsilon_i > \Delta S_{i'} + \epsilon_{i'}.$$

This problem can be modeled with an ordinal logit likelihood if we make the common assumption that the ranking error is distributed extreme value type-I: $\epsilon_i \sim \sigma \cdot EV(1)$.

To estimate this, consider an empirical analogue to equation (3),

$$\Delta\hat{S}_i = \omega(\mathbf{x}_i) \cdot \left(\sum_j \beta_j(\mathbf{x}_i) \hat{\Delta v}_j(\tilde{\mathbf{x}}_i) + \alpha(\mathbf{x}_i) \right) \tag{5}$$

[2]We assume that these functional forms are known. If the $g_j(\cdot)$ utility functions are incorrectly specified to be linear, then the estimated parameters can in some cases measure the combination of the underlying welfare weights and curvature in utility to a first approximation. See Section 5.2.5.

where each theoretical object is replaced with an empirical analogue (w with ω, b_{ij} with β_j, and a with α). Although our notation here is general, in practice there are some restrictions on these objects. In particular, they cannot all vary as a function of $\mathbf{x}_i$, and must be normalized. (In our application, we assume that β_j are constants, which are defined relative to a constant α with $|\alpha| = 1$. We also assume welfare weights are positive: $\boldsymbol{\omega} > 0$. We describe other options for normalization in Online Appendix S2.) The covariates used to estimate treatment effects ($\tilde{\mathbf{x}}_i$) must also differ from those allowed to determine welfare weights and base values ($\mathbf{x}_i$), as we discuss in the following section (3.3).

Then, the placement of i in the ranking $\boldsymbol{z}$ has likelihood

$$l_i = \frac{\exp\left[\frac{1}{\sigma} \cdot \omega(\mathbf{x}_i)\left(\sum_j \beta_j(\mathbf{x}_i)\hat{\Delta v}_j(\tilde{\mathbf{x}}_i) + \alpha(\mathbf{x}_i)\right)\right]}{\sum_{i' \epsilon \Lambda_i} \exp\left[\frac{1}{\sigma} \cdot \omega(\mathbf{x}_{i'})\left(\sum_j \beta_j(\mathbf{x}_{i'})\hat{\Delta v}_j(\tilde{\mathbf{x}}_{i'}) + \alpha(\mathbf{x}_{i'})\right)\right]} \tag{6}$$

where $\Lambda_i = \{i' | z_{i'} < z_i\}$ is the set of households ranked lower than household i.

The likelihood of the full observed ranking $\boldsymbol{z}$ is therefore

$$L(\boldsymbol{z}, \mathbf{x} | \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma) = \prod_i l_i.$$

We observe a single ordering of all alternatives, which differs from discrete choice settings where partial orderings are observed for multiple decisionmakers. For this type of ranked data, we follow the exploded logit likelihood described by Train (2009). As with many discrete choice models, ours is identified up to a scaling parameter, so we impose $\sigma = 1$. We use maximum likelihood to estimate the $\boldsymbol{\omega}$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$ that best match the observed data $\{\boldsymbol{z}, \mathbf{x}, \{\Delta\hat{v}_{ij}\}_{ij}\}$.

For outcomes y_j that are not choices, the estimated $\boldsymbol{\omega}$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$ correspond with those in the theoretical model: $\boldsymbol{\omega}$ will capture the welfare weights $\boldsymbol{w}$; $\boldsymbol{\beta}_j$ the weights on outcome $\boldsymbol{b}_{ij}$; and $\boldsymbol{\alpha}$ the base value a. For outcomes y_j that are choices, the interpretation is slightly different: $\boldsymbol{\beta}_j$ will capture the *difference* between how the policy and households value the outcome j, and $\boldsymbol{\alpha}$ will additionally capture any relaxation of the constraint on choices. When the magnitude of $\boldsymbol{\alpha}$ is normalized to 1, the value of outcome j captured by $\boldsymbol{\beta}_j$ will be defined relative to this base value. We discuss this interpretation in Section 3.4, parameterization in Section 3.5, and other

more nuanced cases in Section 5.

Confidence intervals are computed using a Bayesian bootstrap (Rubin, 1981) over the entire procedure, which accounts for uncertainty in both treatment effects and preference parameters. We generate bootstrap samples by reweighting (rather than resampling) households, compute treatment effects, and then welfare and impact weights.[3]

In many settings, we may not observe a full ranking or score, but rather a binary allocation of beneficiaries and non-beneficiaries ($T_i \in \{0, 1\}$). This corresponds to a ranking with two levels, so the same procedure can be applied, though it will tend to have less statistical power. We provide an empirical illustration of this setting in Section 5.2.1.

3.3 Identification

Exclusion restriction Preferences are identified based on how the policy's ranking (z_i) varies with the set of characteristics that enter the welfare weights ($\mathbf{x}_i$) and the set that determine treatment effect heterogeneity ($\tilde{\mathbf{x}}_i$). Identification of the full model's parameters ($\boldsymbol{\omega}$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$) requires an 'exclusion restriction', whereby $\mathbf{x}_i$ does not include the full set of characteristics in $\tilde{\mathbf{x}}_i$. To see this, note that without such a restriction, one could set $\alpha \equiv 1$ and $\beta_j \equiv 0$ for all j without empirical loss of generality. Conceptually, an exclusion restriction makes it possible to compare how the policy ranks households who have similar welfare and outcome weights (based on $\mathbf{x}_i$) but would be differentially affected by treatment (based on $\tilde{\mathbf{x}}_i$). For a more formal discussion of identification, see Online Appendix S2.

An exclusion restriction can be justified in settings where there exist covariates that are potentially predictive of treatment effect heterogeneity (and thus may reasonably be included in $\tilde{\mathbf{x}}_i$), but which are unlikely to have been prioritized by a policy. Such exclusions are natural in many settings, as welfare and outcome weights represent preferences, which are commonly coarser than heterogeneity in treatment effects, which may depend on many more idiosyncratic factors. For instance, in the PROGRESA

[3]Random weights are drawn from the distribution $Dirichlet(4, ..., 4)$, following Shao and Tu (1995). The Bayesian bootstrap makes it possible to use treatment effect estimators that hold out part of the sample (like causal forests, which we demonstrate later). For those estimators, standard bootstraps can misestimate if the same observation appears in both training and hold-out samples.

example, the policy is unlikely to have placed different weights on the utility of a child based on the household gender composition – but household composition was one of many correlates of impacts from the program. If there is ambiguity about what to include, it can improve confidence to report sensitivity to different sets.

Conditional preferences without exclusion restriction Alternately, one can impose some of the parameters defining preferences (either $\boldsymbol{\omega}$; or $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$), and use the method to estimate what remaining preferences would be consistent with the allocation. For example, one may wish to know what weights on outcomes ($\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$) would be consistent with the allocation if welfare weights were egalitarian ($\omega(\mathbf{x}_i) \equiv 1$). Or, it may be informative to derive the welfare weights ($\boldsymbol{\omega}$) consistent with the allocation given reasonable weights on outcomes, such as if the policy only valued a single outcome (such as health), or if it valued outcomes according to external estimates (e.g., by calibrating $\boldsymbol{\beta}(\mathbf{x}_i)$ and $\alpha(\mathbf{x}_i)$ to estimates from the medical literature). If outcomes are choices, it would be natural to consider a restriction that the policy is not paternalistic (and thus values easing household constraints uniformly (so $\alpha(\mathbf{x}_i) \equiv \alpha$) but $\beta_j = 0$ for all outcomes j that are choices). In Section 4.3.2, we illustrate how these different restrictions can be applied.

Unobservables Our approach reveals the preferences that are consistent with a potential policy $\boldsymbol{z}$, given estimates of the policy's impact $\Delta\hat{\mathbf{v}}$. Our estimates will recover an observed component of welfare, ΔS_i, that is uncorrelated with any unobserved component, ϵ_i. There are several reasons why these implied preferences of the policy might differ from actual preferences.

First, the implied preferences of the policy could differ from the actual policy preferences if the actual ranking is based on correlated unobservables. For example, if a policy is racially biased but an analyst does not allow race to enter modelled preferences, the policy may be found to be consistent with a preference for an income level that is correlated with race. In such settings, the method still reveals preferences that are *consistent* with the policy's values, under the given specification of preferences, just as ordinary least squares recovers the best linear predictor given included variables, even when it omits variables. Similarly, if $\mathbf{x}$ includes both a relevant variable as well as an irrelevant but colinear variable, the method will have imprecise estimates of

the contribution of both, again similar to a standard regression. The specification of preferences (i.e., which variables they are defined over and their functional form) is thus a substantive decision. For this reason, practical applications should include characteristics that may be relevant for differential preference, including those that one believes should be used, as well as characteristics for which there may be concerns of bias.

Second, the implied preferences of the policy that are revealed by our method may differ from the preferences of the policy*maker* if the policymaker has different beliefs about these impacts at the time of the decision. If that were the case, upon observing the results of our method, the policymaker could change the policy to better align with their preferences. The method thus provides a tool for course correction. The method can also be applied in cases where there is no single policymaker—for example, where allocations are the result of deliberations between constituents.

Sufficient variation Identification also requires sufficient variation. Identification of $\boldsymbol{\beta}$ requires that treatment has different impacts on different components of utility. Impact weight $\boldsymbol{\beta}_j$ is identified primarily by the relative ranking of households that are impacted more or less on utility component j. Then, the welfare weights $\boldsymbol{\omega}$ are primarily identified based on how the ranking places households that have different characteristics but achieve similar weighted impact $(u_i(1) - u_i(0))$. If treatment effects were homogeneous, it would not be possible to separately identify $\boldsymbol{\beta}$ and ω.[4] If the treatment effects were heterogeneous but colinear between different components of utility, it would be possible to identify ω but not $\boldsymbol{\beta}$, because the data would not reveal how different components of utility influence the ranking.

3.4 Outcomes that are Choices

In settings where treatment affects choices made by households, the estimates produced by the above procedure have a slightly different interpretation. As before, utility may be derived from outcomes y_{ij} that are not i's choice (e.g., an immune system response to a vaccine), for which $y_{ij}(T_i)$ is a mechanical function. But utility may also

[4]Their combination may be identified, in which case our method would collapse down to a standard exploded logit that does not account for treatment effects.

depend on components that i chooses, and where treatment changes the choice set (for instance, if a cash transfer relaxes the budget constraint).

For each choice outcome $j \in \mathbb{J}_{choice}$, given T_i, household i chooses y_{ij} to maximize its perceived utility

$$\tilde{u}_i = \sum_j \tilde{b}_{ij} g_j(y_{ij}) + \tilde{a} \cdot T_i \tag{7}$$

subject to budget constraint

$$c(\mathbf{y}_{ij \in \mathbb{J}_{choice}}) = \mu_i + \phi_i T_i \tag{8}$$

with associated Lagrange multiplier η_i. The household perceives its value of outcome j as $\tilde{b}_{ij}$, and its base value of being treated as $\tilde{a}$. It faces a weakly convex cost function c that, in the absence of treatment, is constrained to be below μ_i.[5] Treating i alleviates this constraint by amount ϕ_i. Heterogeneity in treatment effects could then arise from households making different choices due to preferences ($\tilde{b}_{ij}$), budgets (μ_i), or efficacy of treatment (ϕ_i).

When choices are made in this manner, the policy will perceive the value of treating household i as

$$\Delta S_i = \underbrace{w(\mathbf{x}_i)}_{\omega(\mathbf{x}_i)} \left(\sum_j \left[\underbrace{\left(b_{ij} - 1_{\{j \in \mathbb{J}_{choice}\}} \cdot \tilde{b}_{ij} \right)}_{\beta_j(\mathbf{x}_i)} \Delta v_{ij} \right] + \underbrace{\phi_i \eta_i + a}_{\alpha(\mathbf{x}_i)} \right), \tag{9}$$

which generalizes equation (3) when some outcomes j are choices. The underbraces highlight the empirical analogues that would result from estimating the main specification (equation (5)). This derivation, shown in Online Appendix S1.1, arises from the envelope theorem.

For outcomes that are not choices, the interpretation of parameters is analogous to before: $\beta_j(\mathbf{x}_i)$ will capture the policy's marginal valuation of that outcome, b_{ij}. However, for outcomes j that are choices, the interpretation is slightly different. Any choices that the policy values in the same way as the household will not be included ($\beta_j(\mathbf{x}_i) = 0$), because the policy will defer to household optimization due to the envelope theorem. Instead, the policy will value the relaxation of the constraint: $\alpha(\mathbf{x}_i)$

[5] In the case where the functions g_j are linear, strict convexity of c is required to ensure an interior optimum.

will pick up this general relaxation ($\phi_i\eta_i$) plus any explicit benefit a. However, if the policy values the choices of i differently from the household (an internality), then $\beta_j(\mathbf{x}_i)$ will also capture the difference in marginal valuation, $b_{ij} - \tilde{b}_{ij}$.

This suggests that the resulting estimates will place weight on nonchoice outcomes that the policy cares about, and choice outcomes that have internalities. One may include other outcomes and statistically test for paternalism ($\beta_j(\mathbf{x}_i) \neq 0$). A policy may also place weight on choices that have externalities, though this leads to a more subtle interpretation, which we discuss in Section 5.

3.5 Parameterization

Our framework will work with general functional forms for $\omega(\mathbf{x}_i)$ as well as for $\beta_j(\mathbf{x}_i)$ and $\alpha(\mathbf{x}_i)$. In the empirical application that follows in Section 4, we model welfare weights as multiplicative,

$$\omega(\mathbf{x}_i) = \Pi_k \gamma_k^{x_{ik}}.$$

We impose the constraint $\boldsymbol{\gamma} > \mathbf{0}$, so that an outcome cannot be a good for some households and a bad for others.

We use simpler functional forms for preferences because our empirical example uses a sample that is not large enough to differentiate all of the dimensions of heterogeneity that our model allows. We model the relative weight on outcome j and the constant term as the same for all households, $\beta_j(\mathbf{x}_i) \equiv \beta_j$, $\alpha(\mathbf{x}_i) \equiv \alpha$, and $|\alpha| = 1$.[6] The first implies that the wedge in marginal valuations is the same for all households (that is, $b_{ij} - \tilde{b}_{ij} \equiv \Delta b_j$). The second implies that any relaxation in the constraint for choice outcomes is the same for each household ($\phi_i\eta_i \equiv \overline{\phi\eta}$ for some fixed $\overline{\phi\eta}$).[7] The third implies that estimated weights on outcome j will be defined relative to any value $\phi_i\eta_i + a_i$.

[6]The model can identify the sign of α, but when we bootstrap the procedure, the sign may switch between draws (i.e., treatment may be a good and the policy favors certain households, or a bad, and the policy disfavors those households). This leads to bimodal confidence intervals that are difficult to interpret. In our baseline model, $\alpha = 1$ achieves superior fit to $\alpha = -1$, so we restrict to the positive sign ($\alpha = 1$) for all results.

[7]More generally, it implies that the sum of any relaxation in the constraint for choice outcomes plus the base value is the same for each household, $\phi_i\eta_i + a_i \equiv \overline{\phi\eta + a}$, if one allowed a_i to vary.

4 Application

To illustrate how our method can be used in applied settings, we consider the case of PROGRESA, a large conditional cash transfer program in Mexico.

4.1 Background on PROGRESA

First implemented by the Mexican federal government in 1997, PROGRESA provided cash transfers to poor households. Transfers averaged 197 pesos per month (approximately $20 USD at the time). Although transfers were conditioned on regular doctor's visits and/or regular school attendance (John Hoddinott, 2004), roughly 99% of enrolled households met these conditions (Simone Boyce, 2003).[8]

Policy documents emphasize the objectives of alleviating poverty and improving the health and education status of poor children in poor households. Coady (2003) also notes the potential for PROGRESA to "bring about important behavioral change," suggesting a possible mismatch between the natural preferences of household decisionmakers and policymakers.

PROGRESA was a targeted program that offered benefits only to eligible households. Within poor communities, the program ranked households based on a 'household poverty score' proxy means test that incorporated a variety of different characteristics (such as household structure, indigenous languages, occupation, income, housing materials, etc.).[9] The score was computed in three steps. First, each household was classified as poor or not poor based on per capita income. Second, that poverty classification was approximated using discriminant analysis based on household characteristics (Skoufias et al., 1999). Third, the list of eligible households was presented in meetings in each community for review; a small number of households changed classification as a result. Our focus is on understanding which underlying values are consistent with the allocation resulting from this method of determining eligibility.

[8]For simplicity, our analysis does not account for the conditionality of the transfer. For a more detailed discussion of PROGRESA and its background, see Emmanuel Skoufias (2008), and Simone Boyce (2003).

[9]The program defined poor communities as those with a high 'village marginality index', computed based on the proportion of households living in poverty, population density, and health and education infrastructure. We focus on the preferences implied by household poverty scores, which were the basis for determining which households within a community were eligible for the program.

During its initial implementation, PROGRESA administrators used a staggered roll-out to randomize when villages could enroll in the program: of the 506 villages included in the evaluation, 320 were randomly assigned to treatment, and initiated into the program in summer 1998. 186 communities were assigned to control and were not initiated into the program until 2000. Behrman and Todd (1999) show that, prior to roll-out, treatment and control communities were statistically indistinguishable across a wide array of observable covariates.

Data

Our analysis relies on two distinct sources of data. The main data comes from household surveys conducted in October 1998 (midline) and November 1999 (endline). These capture household demographics, socioeconomic characteristics, health care utilization, and educational attendance for 14,801 households over the experiment period. Our main analysis focuses on the endline sample of 7,767 households over which our outcomes are defined (N_{rank}), who have at least one child aged 5 or below and at least one child aged 6-16. Within this sample, the transfer given to each household was nearly identical, so we assume the cost of treating each household is identical.[10] We present midline summary statistics for these households in Online Appendix Table S1.[11]

The second data source is a survey that we conducted in 2023 to understand the preferences of Mexican residents over how households should be prioritized for social assistance. We surveyed a sample of 429 Mexican residents to elicit preferences for which types of households should receive transfers, and what types of program impacts were most desirable, in a manner similar to Saez and Stantcheva (2016). The survey asked respondents which household attributes should be considered in the design of

[10]Given the transfer schedule in Skoufias et al. (2001c), 87.2% of households received the upper-bound payment of 750 pesos and 99.2% of households received between 725-750 pesos.

[11]This survey was conducted 1 year after treatment. While there was a baseline survey in 1997, it was more limited and did not include all of the relevant covariates; see Online Appendix Section S3. We note a caveat to the external validity of our approach when using these data to study the values implied by PROGRESA. Since PROGRESA was only targeted at poor villages (i.e., those with a low 'village marginality index'), and because only a subset of households in poor communities were potentially eligible for the program (i.e., households with a high poverty score and with eligible children), the treatment effects we estimate are local to this subpopulation of Mexico. Thus, subsequent inferences about welfare weights should also be interpreted as weights within this subpopulation and may not necessarily generalize to the full Mexican population.

such a program, and relied on multiple price lists to elicit indifference points. We also ask about the degree to which society should entrust household decisionmakers to make the decisions best for children. For a complete description of this survey, see Online Appendix S4.

We focus on the three welfare outcomes (i.e., y_j in our framework) that were emphasized in policy documents and for which the most robust impacts of the program have been documented (Parker and Todd, 2017): (i) *consumption* per-capita; (ii) *child health*, measured as the average number of sick days per child aged 0-5; and (iii) *school attendance*, calculated as the average number of school days missed per child aged 6-16.[12] In our main specification, we allow consumption to enter with logs ($g_0(y_{consumption}) = \log(y_{consumption})$), and allow the other two outcomes to enter the welfare function linearly ($g_j(y_j) = y_j$ for $j > 0$).[13] Note that the program could also have impacted other outcomes not measured; our method will assume that such impacts are either zero or not valued. In Section 4.5.1, we discuss implications and extensions of this simplifying assumption.

We consider welfare weights (i.e., $\boldsymbol{\omega}(\mathbf{x}_i)$) over log of income; number of people; and the household head's age, indigenous status, and whether they completed middle school.

4.2 Characterizing the Decision Rule

As a first step, we characterize the decision rule by indicating which types of households are observed to be ranked higher than others. Table 1 column 1 reports these results, where the contribution of household characteristics to the final ranking $\boldsymbol{z}$ is estimated with a logit ranking model (i.e., our model's likelihood equation (6) with constraints $\boldsymbol{\beta} \equiv 0$ and $\alpha = 1$, estimating the constrained weights $\tilde{\boldsymbol{\gamma}}$). We report coefficients transformed by logarithm ($\log(\tilde{\boldsymbol{\gamma}})$), which can be interpreted as the implied percentage

[12]The review article Parker and Todd (2017) notes that while estimated impacts on consumption, health, and school attendance are robust to adjustments for multiple hypothesis testing, impacts on other outcomes are sensitive to such testing. Specific studies that have estimated significant treatment effects on all three outcomes using the same survey data include John Hoddinott (2004); Emmanuel Skoufias (2008); Simone Boyce (2003); Djebbari and Smith (2008).

[13]A logarithmic functional form for consumption represents a natural benchmark, as Gandelman and Hernandez-Murillo (2015) fails to reject a level of risk aversion consistent with logarithmic utility in Mexico, based on self-reported wellbeing. We also consider robustness to a linear functional form for consumption in Section 4.5.1.

Table 1: What Values are Consistent with the PROGRESA Decision Rule?

		Household Poverty Score 1999	
		Decision Rule	**Implied Preferences**
		(Prioritization)	Welfare Weights
Welfare Weights $log(\boldsymbol{\gamma})$			
Indigenous		0.606 (0.581, 0.634)	-0.174 (-0.225, -0.042)
log(Income)		-0.237 (-0.252, -0.223)	-0.19 (-0.234, -0.138)
Household Size		0.116 (0.112, 0.119)	0.104 (0.082, 0.118)
Household Head Age		-0.02 (-0.021, -0.018)	-0.016 (-0.021, -0.01)
Education (Middle school or above)		-1.007 (-1.263, -0.85)	-0.727 (-0.944, -0.502)
Impact Weights			
Log consumption (per capita)	β_1		6.07 (4.04, 7.19)
Missed Schooling (per day)	β_3		-0.48 (-1.33, -0.03)
Sickness (per child sick day)	β_2		-0.05 (-0.52, 0.56)
Value Regardless of Impact	α		1
N_{rank}		7767	7767
N_{TE}		.	6784
Hypothesis Tests			p-value
Egalitarian	$\boldsymbol{\gamma} \equiv 1$		3.70e-31
Not Paternalistic	$\boldsymbol{\beta} \equiv 0$		3.99e-13
Egalitarian and Not Paternalistic	$\boldsymbol{\gamma} \equiv 1, \boldsymbol{\beta} \equiv 0$		1.33e-64

Notes: 'Decision Rule' column is computed using our method, without treatment effects included in the estimation. 'Implied Preferences' column is calculated using our method, using OLS to estimate heterogeneous treatment effects (see also Figure 2). 95% confidence intervals, in parentheses, are computed using a two-step Bayesian bootstrap procedure that accounts for uncertainty in both treatment effects and preference parameters. Dirichlet bootstrap weights are drawn and then treatment effects are estimated using these bootstrapped weights, and welfare and impact weights are estimated using the same weights. N_{rank} is the number of observations used in estimating the final ranking, N_{TE} describes the number of observations used in estimating the heterogeneous treatment effects, which are then projected to the full sample based on covariates.

changes implied, with 95% confidence intervals in parentheses. For convenience, in the remainder of the paper, we will refer to characteristics as having positive weight if this quantity is above zero (indicating a welfare weight above one), or negative otherwise (indicating a welfare weight below one). These results suggest that households that are indigenous are ranked 60.6 log points higher. It also suggests that each 10% increase in income corresponds with a 2.37% decrease in rank. Each additional household member is associated with a 11.6% increase in ranking. However, the conventional regression in column 1 does not describe *why* these households are ranked highly; it could be that they benefit more (higher treatment effects) or that they are favored (higher welfare weights).

4.3 Results: Estimating What Policies Value

Our main empirical results show how our method can recover the implied values of the PROGRESA allocation.

4.3.1 Heterogeneity in Treatment Effects

As has been documented in prior work, the PROGRESA program significantly impacted several measures of household and child welfare. Among eligible households, we estimate that PROGRESA, on average, increased the log of household monthly consumption by 0.149 (SE=0.015), reduced the number of sick days per child by 0.165 (SE=0.051), and had little effect on the number of school days missed per child (with an average effect of -0.0053, SE=0.028).

However, these treatment effects were heterogeneous. We recover this heterogeneity first by estimating the OLS specification

$$v_{ij} = \theta_{0j} + \boldsymbol{\theta}_{xj}\tilde{\mathbf{x}}_i + (\theta_{Tj} + \boldsymbol{\theta}_{Txj}\tilde{\mathbf{x}}_i)T_i + e_{ij}. \tag{10}$$

We then form predicted treatment effects given

$$\Delta\hat{v}_j(\tilde{\mathbf{x}}_i) = \hat{\theta}_{Tj} + \hat{\boldsymbol{\theta}}_{Txj}\tilde{\mathbf{x}}_i$$

We select variables $\tilde{\mathbf{x}}_i$ to match the specification of heterogeneity in Djebbari and Smith (2008) but omit poverty scores and the village marginality index (and their respective interactions), to avoid potential correlated errors with their use in the second stage. Estimation is performed on the set of potentially eligible households ($N_{TE} = 6784$) for whom randomization affects whether they were given the program. Figure 2 shows that there is considerable heterogeneity in how different households benefit from PROGRESA. Each of the histograms in the figure indicates the distribution of treatment effects for one of the outcomes: for instance, most of the impacts on absences from school are in the range from -0.4 to 0.4 days per child, and most consumption treatment effects are in the range from -0.1 to +0.4 log of consumption.

The Online Appendix provides further insight into the nature and predictors of treatment effect heterogeneity. In Online Appendix Table S2, we show the coefficient estimates for all outcomes. We observe, for instance, that indigenous status significantly

moderates treatment effects for consumption impacts. Online Appendix Figure S1 shows residualized treatment effects, estimated after removing variation explained by the other covariates, to better illustrate how the predictors relate to treatment effects. Panel (a) suggests that, for instance, consumption treatment effects are negatively correlated with income and larger for indigenous households; likewise, panel (b) indicates that schooling treatment effects are smaller in magnitude for households with more members. However, the effects of treatment also vary by fine categories of household composition, such as the number of men aged at least 55 years, and the number of women aged 20-34 years.

4.3.2 Implied Policy Preferences

Next, given that we predict the policy would have impacts $\Delta\hat{v}_{ij}$ on household i, we use our method to back out the implied preferences consistent with ranking that household at position z_i. Although household demographics are correlated with heterogeneous treatment effects, they are likely only coarsely incorporated into the preferences of policymakers for our sample of households that have children. Thus, we assume that these fine measures of household age and gender composition are excluded from welfare weights. This allows us to separately identify the implied preferences of the policy.

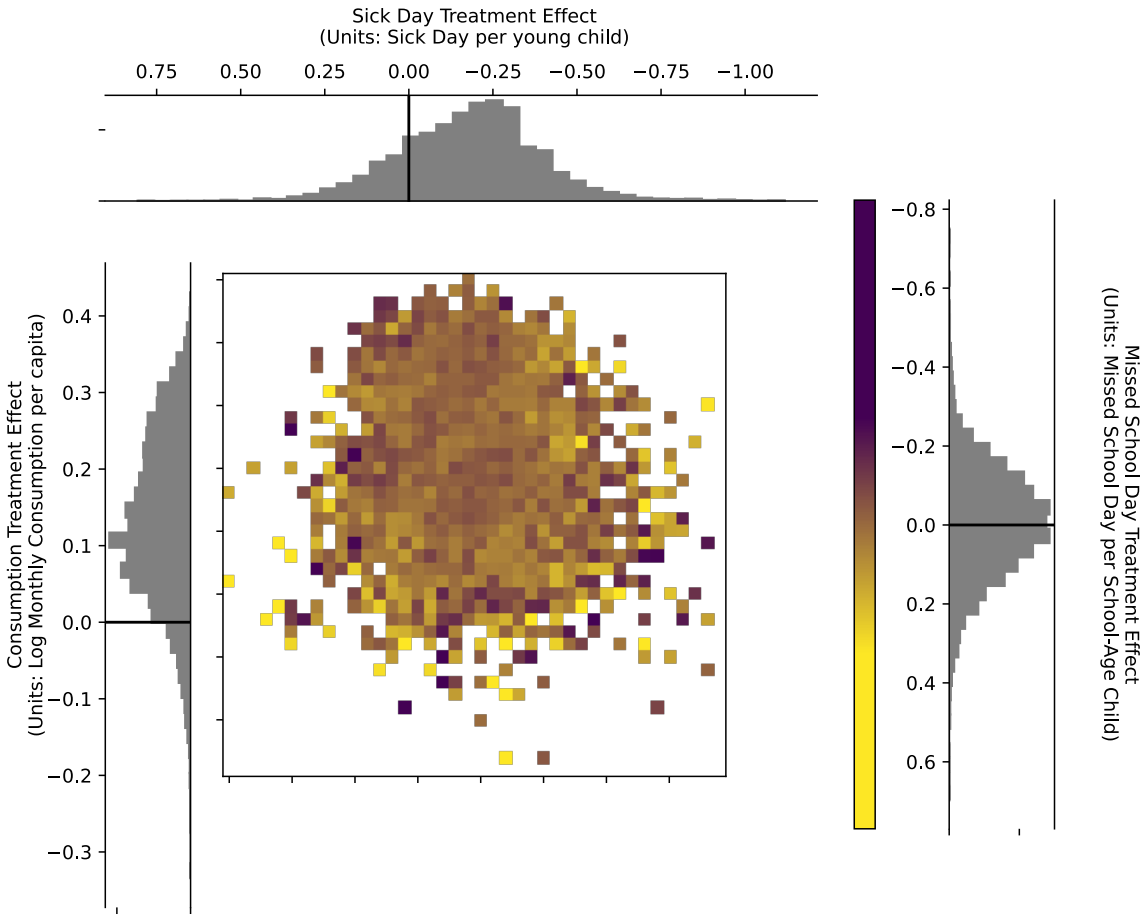
Table 1 column 2 reports the preferences that are consistent with the ranking $\boldsymbol{z}$. The first block of rows shows the implied welfare weights ($\boldsymbol{\gamma}$), and the second block shows implied impact weights ($\boldsymbol{\beta}$ and α). Because the policy ranked all households, we estimate these preferences on this full ranking (N_{rank}).[14]

Accounting for treatment effect heterogeneity leads us to a different understanding of PROGRESA's targeting priorities. For instance, we find that after accounting for the fact that indigenous households benefit more from treatment, the decision rule does not actually place a higher welfare weight on indigenous households; in fact, the estimate suggests that the implied welfare weights may be *lower* (by 17.4%).

The PROGRESA treatment (cash grant) relaxes household budget constraints, which among other things can allow household decisionmakers to improve outcomes

[14]This relies on using the estimated first stage model to extrapolate predicted treatment effects for the 14% of households that were ineligible. This is reasonable if heterogeneity in treatment effects is similar for eligible and ineligible households. In Table S8 (column 2), we show that results are qualitatively similar if we restrict this second step to eligible households.

Figure 2: Distribution of Estimated Treatment Effects



Notes: Heterogeneous treatment effects of PROGRESA, estimated using OLS. Histograms show marginal treatment effects on log consumption (left), sick days among young children (top), and missed school days (right). Center figure shows joint distribution, where each cell corresponds to a combination of consumption and health treatment effects, and is colored according to average treatment effect on attendance. Households without at least one young and one school-age child are omitted from the figure.

for children. For this reason, the outcomes depend on the choices made by households, and the estimates of $\boldsymbol{\beta}$ can be interpreted as the difference between how the policy and household decisionmaker value the outcome, as discussed in Section 3.4. The positive estimate for log consumption thus suggests that the policy places a higher value on this outcome than households. Our estimates of weights on the other impacts are imprecise. For schooling and sickness, the confidence interval includes zero, so we cannot rule out the possibility that the policy's preference coincides with that of household decisionmakers (though for sickness, the confidence interval barely includes zero). Overall, our estimates suggest that, from the perspective of the policymaker (equation (2)), on average 55% of the impact of PROGRESA on household utility comes from simply providing the transfer, irrespective of impacts on measured outcomes (the constant term α). Approximately 45% is derived from the impact on consumption (β_1), and <1% derives from impacts on health and schooling. The ratio α/β_1 suggests that the implied value of providing the program independent of impacts corresponds to 0.16 log points of consumption, or a mean consumption increase of 23.1 pesos per person per month, which is slightly smaller than the average transfer of 33.9 pesos per person per month (John Hoddinott, 2004).

We can also test whether our estimated parameters are consistent with postulated welfare functions. We use Wald tests (with the bootstrapped covariance matrices) to test the null hypothesis that preferences are egalitarian ($\boldsymbol{\gamma} \equiv \mathbf{1}$), non-paternalistic ($\boldsymbol{\beta} \equiv \mathbf{0}$), or both egalitarian and non-paternalistic ($\boldsymbol{\gamma} \equiv \mathbf{1}$ and $\boldsymbol{\beta} \equiv \mathbf{0}$). These results are presented in the bottom panel of Table 1. We reject the hypothesis that our estimated coefficients do not place differential weight on different households and outcomes, across all specifications. We also strongly reject non-paternalism.

4.3.3 Assessing Preferences

Our framework also makes it possible to compare the preferences consistent with alternative policies. For instance, the Mexican government expanded PROGRESA in 2003, changing the poverty score to increase the priority of older and smaller households (Skoufias et al., 2001b). As shown in column 2 of Table 2, by comparing the relative magnitudes of the coefficients in each rule, our method reveals that this new poverty score implicitly switched to having a positive welfare weight for indigenous

Table 2: Assessing Decision Rules

		(1) Implied Preferences (Estimated) 1999 Pov. Score	(2) Implied Preferences (Estimated) 2003 Pov. Score	(3) Stated Preferences (Resident survey)
Welfare Weights $log(\boldsymbol{\gamma})$				
Indigenous		-0.174 (-0.225, -0.042)	0.062 (0.005, 0.196)	0.065 (0.057, 0.072)
log(Income)		-0.19 (-0.234, -0.138)	-0.072 (-0.109, -0.039)	-0.071 (-0.270, 0.129)
Household Size		0.104 (0.082, 0.118)	0.086 (0.075, 0.096)	0.015 (-0.018, 0.048)
Household Head Age		-0.016 (-0.021, -0.01)	-0.001 (-0.004, 0.002)	0.004 (0.002, 0.005)
Educated		-0.727 (-0.944, -0.502)	-0.416 (-0.58, -0.3)	-0.065 (-0.099, -0.030)
Impact Weights				
Log Consumption (per capita)	β_1	6.07 (4.04, 7.19)	2.23 (1.48, 2.72)	4.672 (3.190, 6.153) †
Missed Schooling (per day)	β_3	-0.48 (-1.33, -0.03)	-0.32 (-0.71, -0.07)	-1.189 (-1.723, -0.655) †
Sickness (per child sick day)	β_2	-0.05 (-0.52, 0.56)	-0.01 (-0.21, 0.3)	-0.740 (-1.097, -0.382) †
Value Regardless of Impact	α	1	1	.
N_{rank}		7767	7767	.
N_{TE}		6784	6784	.
$N_{respondents}$		.	.	424*

Notes: Columns 1-2 are estimated using our method, using OLS to estimate heterogeneous treatment effects. Column 3 indicates stated preferences estimated on a survey of Mexican residents; to reduce the impact of outliers we report the median response (for details of this survey, see Appendix S4). † Survey weights scaled to match the scale of estimated impact weights since we did not estimate the scale of idiosyncratic noise in the survey. 95% confidence intervals are reported in parentheses. 'Educated' defined as a household head with a middle school education or above. In the first two columns, confidence intervals are computed using a two-step Bayesian bootstrap procedure that accounts for uncertainty in both treatment effects and preference parameters: dirichlet bootstrap weights are drawn and then treatment effects are estimated using these bootstrapped weights, and welfare and impact weights are estimated using the same weights. N_{rank} describes the number of observations used in estimating the final ranking, N_{TE} describes the number of observations used in estimating the heterogeneous treatment effects, which are then projected to the full sample based on covariates. *: The number of survey respondents differs for different parameters (ranging between 411 and 424), due to incomplete responses. Confidence intervals in column 3 are computed using standard errors from a standard bootstrap over all individuals, with missing values dropped.

households, and placed less welfare weight on lower-income and younger households.

Table 2 also illustrates how the implemented policy (column 1) compares to the median stated preferences of residents, as reported in the survey we conducted in 2023 (column 3). Welfare weights $\boldsymbol{\gamma}$ are estimated from residents' choices of how to prioritize different households in a multiple price list. The welfare weights implied by the implemented policy are similar to resident preferences, but place higher welfare weights on indigenous households. Impact weights $\boldsymbol{\beta}$ are formed by asking how a household would make decisions between an outcome and a cash transfer, and then ask how society should value that outcome relative to the decisionmaker in the household.[15] On average, survey respondents value impacts on the health of children more than they expect household decisionmakers to, and more than the implemented policy does. In separate survey questions, we asked residents to rate statements describing whether the government should directly support children, whether these outcomes have externalities, and whether the government should trust parents to do what is best for children. The responses, summarized in Online Appendix Table S9, are consistent with support for paternalism.

4.4 Counterfactuals

We next consider the reverse problem: given preferences, what would the resulting policy look like? In the PROGRESA example, Table 3 compares the policy's true allocation (column 1) to counterfactual allocations that would have resulted from alternative preferences (columns 2-6). Panel A indicates which preferences are used. We allow the welfare weights to be those estimated from the 1999 policy (columns 1, 4-6), those elicited from the resident survey (column 2), or fixed to weight all households equally (column 3). We allow the impact weights to be those estimated from the 1999 policy (columns 1 and 3), those elicited from the resident survey (column 2), or to only value one outcome (columns 4-6). Panel B indicates the decision rule implied by those preferences, where we take the implied ranking and estimate a logit model, as in column 1 of Table 1. Panel C shows the average outcomes that would be

[15] This combination allows us to estimate the implied weight a policy should place on each outcome, $\tilde{b}_j - b_j$. Because this survey does not estimate the scale of the idiosyncratic error σ, we rescale these survey estimates of $\boldsymbol{\beta}$ to have the same average magnitude as those estimated from the 1999 poverty score. See Online Appendix S4.

expected under the hypothetical policy, assuming the hypothetical policy treated the same number of households as the implemented policy.

Survey-based Estimates of Resident Preferences Column 2 of Table 3 shows the allocation that would result from imposing the preferences of residents as revealed by the survey. Relative to the actual policy in column 1, the hypothetical policy in column 2 places greater priority on indigenous households, and less priority on households with less education. Other household attributes are similarly prioritized under the two policies. In Panel C, we see that the policy consistent with resident preferences would slightly increase average consumption and slightly reduce average child missed school days and sick days relative to the implemented policy.

Alternate Welfare Weights When welfare weights are set equal across households (column 3), the resulting ranking increases the priority of indigenous households, slightly lowers the priority of large and poor households, and no longer prioritizes households with lower education.

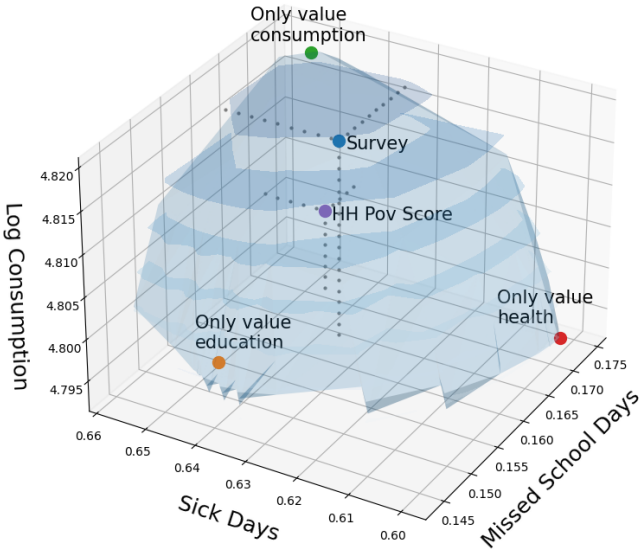
Prioritizing Specific Welfare Outcomes Most real-world policies balance multiple outcomes. For comparison, columns 4-6 of Table 3 present counterfactual allocations that would result in the extreme case where a policy was designed to improve only a single outcome. For instance, a policy that maximized impacts on consumption with no explicit consideration of health or education (column 4) would end up placing greater priority on households where the head is indigenous, and would place higher priority on households with lower income. Alternatively, a policy designed to maximize educational impacts would prioritize smaller households and those with *higher* income (column 5). Finally, if only health impacts were valued, the policy would largely preserve the prioritization of indigenous households, and put smaller emphasis on lower-education households (column 6).

Table 3: Designing Decision Rules

	(1) HH Poverty Score	(2) Resident Preferences	(3) Equal Welfare Weights	(4) Policy only values impact on: Consumption	(5) Policy only values impact on: Education	(6) Policy only values impact on: Health
Panel A: Preferences						
Welfare Weights $\boldsymbol{\gamma}$	Estimated	From survey	Unity	Estimated	Estimated	Estimated
Impact Weights $\boldsymbol{\beta}$	Estimated	From survey	Estimated	Only consumption	Only education	Only health
Panel B: Implied decision rule (priority over covariates, in logs)						
Indigenous	0.606	2.289	1.987	2.372	-0.114	-0.023
log(Income)	-0.237	-0.344	-0.183	-0.375	0.303	0.178
Household Size	0.116	-0.009	-0.022	0.042	-0.137	-0.041
Household Head Age	-0.02	-0.009	-0.012	-0.02	-0.013	-0.036
Education	-1.007	-0.206	0.054	-0.770	-0.532	-0.131
Panel C: Counterfactual outcomes (monthly)						
Log Consumption per capita (pesos)	4.803	4.817	4.819	4.819	4.798	4.794
Missed school (days/child)	0.169	0.162	0.169	0.172	0.146	0.172
Sickness (sick days/child)	0.645	0.634	0.649	0.651	0.641	0.600
Model Log Likelihood	-60930	-61647	-61953	-61327	-61467	-61615
N_{rank}	7767	7767	7767	7767	7767	7767

Notes: Table shows the distributional and outcome effects of designing decision rules using our framework. Panel A indicates which weights are used to prioritize households. Column 1 uses the ranking assigned by PROGRESA. Column 2 uses preferences elicited in a survey we conducted of Mexican residents. For the survey column, we set $\alpha = 1$ and scale survey impact weights to have the same average magnitude as estimated impact weights. Survey weight model likelihood computed using same constant term. Column 3 projects the ranking as though the policy assigned the same welfare weight to all households, so preference results from differences in outcomes. Columns 4-6 indicate what would have happened if the policy used the estimated weights over households but only valued about impacts on education/health/consumption, with $\alpha = 0$. Panel B shows the distributional effects of each column's preferences, by estimating the implied priority ranking across households. Panel C shows each policy's expected average outcomes, calculated using estimates of heterogeneous treatment effects.

Figure 3: Expected Program Impacts under Alternative Preferences



Notes: Figure shows the frontier of possible average welfare impacts that would have resulted from different allocations of PROGRESA. Each axis indicates the expected average impacts for the corresponding welfare outcome. Labeled points indicate specific allocations described in Table 3.

Understanding the policies that would result from extreme preferences can help in understanding the full set of potential policies, and what those policies imply. Figure 3 characterizes the frontier of possible average outcomes that would result from different allocations of PROGRESA. This frontier is shown as a convex hull with contour lines; the labeled points correspond to the policies given in the columns of Table 3. Policies that only value a single outcome lie at the corners of the outcome space. The implemented program ('HH Poverty Score') is close to the allocation consistent with the survey of Mexican residents preferences. All labeled points apply welfare weights so one would not expect either to reach the frontier for unweighted outcomes, but they are close.[16]

More broadly, this method makes it possible to navigate program design in outcome space, rather than implementation space.

[16]The distances from labeled points to the frontier, defined as frontier point coordinates minus allocation point coordinates and in units of (Log Consumption, Sick Days, Missed School Days), are as follows. Survey: (-0.0003, -0.005, 0.001); HH Poverty Score: (-0.010, -0.0009, 0.005); Consumption: (0.0, -0.0001, 0.0001); Health: (0.0001, -0.0002, -0.0003); Education: (-0.0003, 0.0002, -0.003). The distance between the implemented program and the survey is (-0.014, 0.010, 0.007).

4.5 Additional Considerations

4.5.1 Specification of Outcomes

This section briefly discusses which outcomes should be included when modeling welfare from the perspective of the policy. One simple approach is to include outcomes in the framework in order to empirically test whether they in fact influence the decision rule; that is, whether the estimated coefficients differ from zero. As noted previously, this interpretation depends in part on whether the outcomes are choices (and treatment simply alters the choice set); in that case, non-zero weight implies that the policy values the outcome differently from the household.

When multiple reasonable sets of outcomes could be included, it is reasonable to test multiple sets to assess robustness. For instance, Online Appendix Table S4 shows how our main estimates (from Table 2) change if we split the consumption outcome into food and nonfood consumption (column 2); it also includes specifications that include just 1 or 2 outcomes at a time (columns 3-8). We saw previously that consumption explains a substantial portion of the impact on households in our baseline specification. Alternate specifications find similar results so long as they include consumption; specifications that omit consumption find estimates close to that of the raw ranking itself (Table 1 column 1).[17]

Additionally, there may be multiple reasonable functional forms through which outcomes could be valued. Our primary specification uses log consumption, but column 9 of Table S4 presents results using a linear functional form for consumption. Results are again similar: indigenous households have a positive welfare weight, but this weight is still much smaller relative to the weights on other attributes than the ranking alone would suggest.

4.5.2 Specification of Covariates

The set of covariates included when estimating heterogeneous treatment effects ($\tilde{\mathbf{x}}_i$) is flexible: one may include any baseline variables predictive of heterogeneity so long as one takes care to avoid overfitting. The set of covariates $\mathbf{x}_i$ allowed into the welfare

[17] An additional outcome that a policy might value is long term investments, as documented in Gertler et al. (2012), which could have trade-offs with short-term consumption. This analysis could be extended to include investments as an outcome.

weights is more nuanced and should be motivated by theory. As noted, the practical requirement (exclusion restriction) is that the covariates $\mathbf{x}_i$ not include all those in $\tilde{\mathbf{x}}_i$.

When there are multiple reasonable specifications for $\mathbf{x}_i$, it again is reasonable to assess robustness to those different specifications. This is demonstrated for PROGRESA in Online Appendix Tables S5 and S6, which compare specifications with different covariates. Results are almost all qualitatively unchanged.

Absent an exclusion restriction, the framework can be applied by imposing some parameters and estimating the rest, as suggested in Section 3.3. We demonstrate this approach in Online Appendix Table S7, which shows what happens when preferences are assumed to be egalitarian or to only prioritize one particular impact. In the latter case, one may wish to impose impact weights from a scientific literature; for simplicity we assume that for the selected j, $|\beta_j| = \alpha = 1$. We find that results are broadly similar across these specifications, with positive weight on household size and negative weights on household income and head education. This assumes that consumption impacts are valued much less than in our full estimated specifications, and accordingly finds a positive weight on indigenous households, though in most cases it is attenuated compared to the ranking alone.

A caveat: impacts may correlate with unobservables The exclusion restriction is more nuanced for policies that may value households based on components that are difficult to measure. Imagine a policymaker assigns household i a true welfare weight w_i, which may contain components that are not well captured by observable covariates $\mathbf{x}_i$, such as 'neediness'. If those components are correlated with impact on some outcome, $\Delta y(\tilde{\mathbf{x}}_i)$ (say, how much of a grant that a household spends on food consumption), then our method may attribute a weight on this impact that in fact arises from the correlation with the unobservable.[18]

It is the exclusion restriction that opens the door for this problem. $\mathbf{x}_i$ should include all variables that may enter welfare weights, including those that may signal unobservables, if one expects these are valued. The set of variables allowed to enter

[18]For example, imagine that a policy values households based on neediness (w_i), and values simply providing treatment but not its impacts ($a > 0$, $b = 0$). It proxies neediness with food consumption (y). If the correlates of food consumption are omitted from the specification of welfare weights ($\mathbf{x}_i$) then we might estimate $\omega(\mathbf{x}_i) = 1$, $\alpha = 0$, and $\beta = w_i(\Delta y)$ and mistakenly conclude that the policymaker values all people equally, and values impacts on food consumption.

into treatment effects, but which are excluded from $\mathbf{x}_i$, should not include variables that may signal unobservable welfare weights. If a user of this method is unwilling to commit to excluding characteristics from $\mathbf{x}_i$, that would suggest the exclusion restriction may not hold, and $\boldsymbol{\omega}$, $\boldsymbol{\beta}$, and α are not separately identified in their setting. One may still impose part of preferences and estimate the remainder as demonstrated above.

4.5.3 Treatment Effect Specification and Measurement Error

The first stage of our approach can be estimated with a variety of methods and specifications for heterogeneous treatment effects. Using a flexible method, such as a linear estimator with many covariates or causal forests, can reduce the chance of misspecification. However, more flexible methods can result in noisier first stage predictions, which could attenuate or bias second stage estimates. In Online Appendix Section S5, we discuss this in more detail, and present all our results replacing the OLS first stage with causal forests, a nonlinear estimator (Wager and Athey, 2018). Results are all similar. We also assess the potential magnitude of attenuation and misspecification with Monte Carlos and a bias correction technique from the statistics literature (simulation extrapolation, or SIMEX, Cook and Stefanski, 1994). Our PROGRESA estimates remain very similar when we apply this correction. In applications where measurement error has larger effects, one may use corrections, or use a different approach such as jointly modeling both stages of the method in a single likelihood.

5 Broader Applications and Extensions

The PROGRESA example illustrates how our method can be used retroactively to understand the priorities of an observed allocation policy. It thus provides a type of 'value audit', which can reveal the values consistent with an implemented policy. These values can then be compared to the values of constituents, or the stated objectives of policymakers.

The same technique can be used prospectively, to help policy designers iteratively improve the alignment between their values and the values implied by the policies they

adopt. This requires a first step that estimates how much different households would benefit from the policy. In the PROGRESA case, for example, we use data from the first phase of the program roll-out to estimate treatment effect heterogeneity; these results are shown in Figure 2. Then, for any *prospective* policy proposal — which need not be implemented — our method can be used to estimate welfare parameters implied by that proposed policy. For instance, column 2 of Table 2 illustrates how a 2003 update to the original PROGRESA poverty score placed higher weight on wealthier households. Finally, the method can help course-correct, to better align future policies with stated preferences. In our example, this is most directly illustrated in Table 3, which shows the policies that would result from counterfactual preferences.

The method can be applied in a variety of settings. For instance, medical interventions are often scarce; given knowledge about the heterogeneous effects of these treatments, our approach can provide insight into the welfare weights implied by different proposed allocation policies. Likewise, a marketing agency may be interested in targeting promotions to customers who are likely to respond along multiple margins, such as specific purchases or longer-term retention, while also prioritizing specific consumer segments; our approach can help them translate from a menu of possible campaigns to the preferences and values implied by each one.

What do these diverse settings have in common? We identify three main elements that are necessary for our framework to be applied. The first requirement is a practical one: our framework requires an understanding of the (potentially heterogeneous) impacts of a policy on one or more outcomes, in order to obtain the $\Delta\hat{v}_{ij}$ in the first estimation step.[19] These are easiest to estimate when there is a pilot where treatment is randomly assigned to a representative subset of the population of interest; this was the case with PROGRESA, and our analysis in Section 4 shows how to apply the framework in this canonical setting. Absent a randomized intervention, it may be possible to use

[19]Note that private parties may desire to allocate treatment to people who have high outcome *levels*, rather than those who would see the highest *impacts* (e.g., an employer may hire candidates who will have the highest performance, not those whose performance would benefit the most from a job offer). In such cases, our method could be used with two alterations: the welfare function (equation (1)) would sum only over treated (hired) individuals, and as a result one would replace $\Delta\hat{v}_{ij}$ in equation (3) with the predicted outcome that would result if i were treated, $\hat{v}_{ij}(1) = v_{ij} + (1 - T_i)\Delta\hat{v}_{ij}$. If one is willing to assume that treatment effects do not differ between people (so that most heterogeneity arises from levels), then one could replace this with an individual's level v_{ij}, and could use a similar approach without estimating treatment effects.

non-experimental methods for estimating treatment effect heterogeneity (e.g., Kent et al., 2020; Johansson et al., 2018), or even to extrapolate from existing evidence on heterogeneous treatment of similar policies in similar environments. The second requirement is that the implementer must define the outcomes and characteristics that enter into the objective function. This decision has implications for both identification (as discussed in Section 3.3) and for interpreting the downstream analysis (discussed in Section 3.4). Third, the framework requires sufficient data and variation to identify the key parameters of our model, which we discuss next.

5.1 Sample Size Considerations

The sample size requirements for implementing this approach will vary depending on the amount of heterogeneity, noise, and the complexity of the specification of impact and welfare weights. Using Monte Carlo simulations, we provide an example of how error varies with the number of observations used to estimate treatment effects and the ranking. Online Appendix Table S10 provides estimates of mean absolute error over differing sample sizes, assuming treatment effects are linear in parameters and using OLS for the first stage. These simulations suggest that so long as one has a sufficiently large sample over which to estimate treatment effects, one can substantially improve precision by simply observing more rankings between households. Since estimating treatment effects may require running an experiment, such a Monte Carlo exercise can help inform power calculations to ensure that the design is adequately powered both for estimating treatment effects and to use our method to evaluate potential policies.

5.2 Interpretation Under Different Scenarios

Certain settings may require additional nuance in implementation and interpretation.

5.2.1 If Only an Allocation is Observed

In many settings, information about the allocation might be more limited than in our benchmark case where a full ranking is observed. For instance, a tax policy may only have a small number of brackets, or it may only be possible to observe a binary allocation. This may reduce the variation available to estimate preferences, but in

principle our method can still be used. In the PROGRESA example, column 4 of Online Appendix Table S8 demonstrates that when our method is applied to a binary allocation ($z(\mathbf{x}_i) = 1\{i \text{ eligible}\}$), point estimates are similar to those reported in Table 1. Although the point estimate for indigenous is positive, it is smaller relative to the other coefficients than would be implied by the decision rule, and its confidence interval nearly covers zero. Otherwise, most qualitative conclusions are the same.

5.2.2 Continuous Treatment

Our model considers a binary treatment given in rank order. One could extend the framework to consider instead a treatment $T_i \in [0, \infty)$ that may be given in varying quantities. Estimation would differ in two respects. In the first step, one would estimate the slope of each component of utility with respect to the continuous treatment, $\frac{d\hat{v}_{ij}}{dT_i}$ (the continuous analogue of $\Delta\hat{v}_{ij}$). In the second step, one would solve for the parameters that equate the marginal utility of each household i at the observed transfer amounts $\mathbf{T}$, from the perspective of the policy. For more details see Online Appendix S1.3.

5.2.3 Externalities

The interpretation of the method's estimates can change if treating one household affects another household. We explore two stylized cases of how spillovers could arise:

Altruism i may value the utility of i'. Then, if i receives a treatment that expands their choice set, they may use that opportunity to help i'. For example, a household receiving a cash transfer may share resources with its neighbors. In Online Appendix S1.2.1, we derive a formula for ΔS_i that generalizes to choice outcomes with altruism. This formula includes terms for how treating i affects its transfers to i', $\Delta\delta_{ii'}$, and each outcome j of i', $\Delta v^{ext}_{ii'j}$. In the PROGRESA example, there is evidence that treated households share benefits with untreated households in the same village (Angelucci and De Giorgi, 2009), mostly through transfers and loans. Such spillovers would affect the interpretation of our results primarily if they were differential (so that treating household i would have different spillovers than treating household i'); if each household induced the same spillovers, the interpretation would remain mostly

the same because the benefit of treating each household is similarly shifted. In the case of PROGRESA, the experimental design allows only for the estimation of average spillover effects, so we cannot empirically determine if spillovers were differential.[20]

Direct effects i may value the outcomes of i', and thus their treatment status. For example, school admission may take into account peer effects, or a vaccination strategy may prioritize some individuals because of their propensity for contagion to sensitive groups. When outcomes are choices, a policy may wish to correct for each household undervaluing their impact on others. We derive the general formula for ΔS_i with such externalities in Online Appendix S1.2.2.

5.2.4 Manipulation

Households may have incentives to manipulate their reported characteristics $\tilde{\mathbf{x}}_i$ in order to be prioritized. If the ease of manipulating a characteristic differs between households in unobserved ways, a policy that anticipates manipulation may place a weight on it that differs from their preference, to account for manipulation (Frankel and Kartik, 2018; Björkegren et al., 2020). We analyze the initial PROGRESA rule as was implemented in a pilot, so we expect both manipulation by households, and anticipation of manipulation by policymakers, to be negligible. However, manipulation may be relevant in settings where the decision rule is publicized and households are familiar with it. Extending this framework to invert the preferences implied by strategy-robust decision rules is an interesting direction for future work.

5.2.5 Nonlinear Utility Functions

One can alternately consider utility functions of general form, $u_i(\mathbf{v}_i)$ from the perspective of the policy and $\tilde{u}_i(\mathbf{v}_i)$ from the perspective of the household. Then, equation (9)

[20]Differential spillovers could be estimated with a more nuanced experiment that randomized the composition of treated households by village: e.g., in some villages treating indigenous households and others nonindigenous, and tracking how ineligible outcomes compare to those in controls where no one is treated.

generalizes to

$$\Delta S_i \approx \underbrace{w(\mathbf{x}_i)}_{\approx\omega(\mathbf{x}_i)} \left(\sum_j \left[\underbrace{\left(\frac{\partial \tilde{u}_i}{\partial v_{ij}} - 1_{\{j \in \mathbb{J}_{choice}\}} \cdot \frac{\partial u_i}{\partial v_{ij}} \right)}_{\approx\beta_j(\mathbf{x}_i)} \Delta v_{ij} \right] + \underbrace{\phi_i \eta_i + a}_{\approx\alpha(\mathbf{x}_i)} \right)$$

The interpretation generalizes from the previous linear case. β_j captures the policy's marginal valuation of outcome j for nonchoice outcomes ($\frac{\partial \tilde{u}_i}{\partial v_{ij}}$). For choice outcomes, it will capture the difference ($\frac{\partial \tilde{u}_i}{\partial v_{ij}} - \frac{\partial u_i}{\partial v_{ij}}$). When $\tilde{u}_i$ and u_i are linear functions of $\mathbf{v}_i$, then $\beta_j(\mathbf{x}_i)$ and $\alpha(\mathbf{x}_i)$ will correspond to the underlying objects. If they are more nuanced functions, they will represent approximations. As we show in Online Appendix Section S1.4, this linear approximation can affect parameter estimates if the function actually has curvature. This suggests that one should attempt to measure outcomes $\mathbf{v}_i$ in metrics that enter utility approximately linearly.

5.2.6 Heterogeneous Treatment Costs

If the costs of treatment differ between households, the comparisons underlying our method should be adjusted to account for this difference. For example, a policy might treat a single high-cost household i or a combination of other low-cost households. If one wishes to hit the budget constraint exactly, this becomes a combinatorial problem.

6 Conclusion

Policy discussions commonly revolve around the mechanics of implementation, rather than more fundamental notions of utility and welfare weights. This paper demonstrates a way to invert those discussions. We provide a method to recover the primitives consistent with observed policies, using a model of preferences in conjunction with methods for estimating heterogeneous treatment effects, and demonstrate how to convert between welfare and allocation space.

Our main empirical example illustrates how our method can be used to understand the priorities of an allocation policy: that is, we estimate the relative value that PROGRESA placed on different household outcomes (e.g., education vs. health), and calculate the implied welfare weights assigned to different types of households

(e.g., poor vs. indigenous households). We show how this framework can be applied to evaluate the policies that would be implied by counterfactual preferences, such as different relative valuations of household outcomes. Beyond social assistance and welfare policy, we expect that this framework will be relevant to a much broader range of contexts where there is interest in understanding the values implied by a policy or allocation, and in designing policies to better align with values.

This framework could be used in several ways. To begin, it could be used to characterize the realized allocations of an existing program, to provide an indication of the preferences they imply. This, in turn, can provide a way to audit existing programs, to help hold policymakers accountable for past decisions – and in particular, to evaluate whether an implemented allocation reflects the stated goals of the policy, or the preferences of constituents. Perhaps most importantly, this approach can be used to adjust proposed policies to better align with those goals.

Data Availability Statement

The data and code underlying this research are available on Zenodo at `https://dx.doi.org/10.5281/zenodo.15557913`.

References

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, and Julia Tobias, "Targeting the Poor: Evidence from a Field Experiment in Indonesia," *American Economic Review*, June 2012, *102* (4), 1206–1240.

Alderman, Harold, Jere R Behrman, and Afia Tasneem, "The Contribution of Increased Equity to the Estimated Social Benefits from a Transfer Program: An Illustration from PROGRESA/Oportunidades," *The World Bank Economic Review*, October 2019, *33* (3), 535–550.

Angelucci, Manuela and Giacomo De Giorgi, "Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption?," *American Economic Review*, February 2009, *99* (1), 486–508.

Athey, Susan and Stefan Wager, "Policy Learning with Observational Data," *arXiv:1702.02896 [cs, econ, math, stat]*, September 2020. arXiv: 1702.02896.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan, *Fairness and Machine Learning*, fairmlbook.org, 2018.

Barr, Nicholas, *Economics of the welfare state*, Oxford university press, 2012.

Behrman, Jere R. and Petra E. Todd, "Randomness in the experimental samples of PROGRESA (education, health, and nutrition program)," *International Food Policy Research Institute, Washington, DC*, 1999.

Björkegren, Daniel, Joshua E. Blumenstock, and Samsun Knight, "Manipulation-Proof Machine Learning," *arXiv:2004.03865 [cs, econ]*, April 2020. arXiv: 2004.03865.

Bourguignon, François and Amedeo Spadaro, "Tax–benefit revealed social preferences," *The Journal of Economic Inequality*, March 2012, *10* (1), 75–108.

Boyce, Paul Gertler Simone, "An Experiment in Incentive-Based Welfare: The Impact of PROGRESA on Health in Mexico," in "," Vol. 85 Royal Economic Society 2003.

Chiappori, Pierre-André, Ivana Komunjer, and Dennis Kristensen, "Non-parametric identification and estimation of transformation models," *Journal of Econometrics*, 2015, *188* (1), 22–39.

Chilet, Jorge Ale, "Gradually Rebuilding a Relationship: The Emergence of Collusion in Retail Pharmacies in Chile," 2017.

Coady, David, "Alleviating structural poverty in developing countries: The approach of PROGRESA in Mexico," 2003.

Coady, David P., "The Welfare Returns to Finer Targeting: The Case of The Progresa Program in Mexico," *International Tax and Public Finance*, May 2006, *13* (2-3), 217–239.

Coate, Stephen and Stephen Morris, "On the Form of Transfers to Special Interests," *Journal of Political Economy*, December 1995, *103* (6), 1210–1235.

Cook, J. R. and L. A. Stefanski, "Simulation-Extrapolation Estimation in Parametric Measurement Error Models," *Journal of the American Statistical Association*, 1994, *89* (428), 1314–1328.

Djebbari, Habiba and Jeffrey Smith, "Heterogeneous impacts in PROGRESA," *Journal of Econometrics*, July 2008, *145* (1), 64–80.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel, "Fairness through awareness," in "Proceedings of the 3rd innovations in theoretical computer science conference" ACM 2012, pp. 214–226.

Ensign, Danielle, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian, "Runaway Feedback Loops in Predictive Policing," *arXiv:1706.09847 [cs, stat]*, June 2017. arXiv: 1706.09847.

Filmer, Deon and Lant H. Pritchett, "Estimating Wealth Effects Without Expenditure Data—Or Tears: An Application To Educational Enrollments In States Of India*," *Demography*, February 2001, *38* (1), 115–132.

Fleurbaey, Marc and Francois Maniquet, "Optimal income taxation theory and principles of fairness," *Journal of Economic Literature*, 2018, *56* (3), 1029–79.

Frankel, Alex and Navin Kartik, "Muddled Information," *Journal of Political Economy*, November 2018, pp. 000–000.

Gandelman, Nestor and Ruben Hernandez-Murillo, "Risk Aversion at the Country Level," SSRN Scholarly Paper ID 2646134, Social Science Research Network, Rochester, NY 2015.

Gechter, Michael, Cyrus Samii, Rajeev Dehejia, and Cristian Pop-Eleches, "Evaluating Ex Ante Counterfactual Predictions Using Ex Post Causal Inference," *arXiv:1806.07016 [stat]*, July 2019. arXiv: 1806.07016.

Gertler, Paul, "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment," *The American Economic Review*, 2004, *94* (2), 336–341.

Gertler, Paul J., Sebastian W. Martinez, and Marta Rubio-Codina, "Investing Cash Transfers to Raise Long-Term Living Standards," *American Economic Journal: Applied Economics*, January 2012, *4* (1), 164–192.

Greco, Salvatore, Alessio Ishizaka, Menelaos Tasiou, and Gianpiero Torrisi, "On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness," *Social Indicators Research*, January 2019, *141* (1), 61–94.

Hanna, Rema and Benjamin A. Olken, "Universal Basic Incomes versus Targeted Transfers: Anti-Poverty Programs in Developing Countries," *Journal of Economic Perspectives*, November 2018, *32* (4), 201–226.

Haushofer, Johannes, Paul Niehaus, Carlos Paramo, Edward Miguel, and Michael Walker, "Targeting impact versus deprivation," *Working Paper*, 2022.

Hendren, Nathaniel, "Measuring economic efficiency using inverse-optimum weights," *Journal of Public Economics*, July 2020, *187*, 104198.

Hoddinott, Emmanuel Skoufias John, "The Impact of PROGRESA on Food Consumption," *Economic Development and Cultural Change*, 2004, *53* (1), 37–61.

Horn, Roger A. and Charles R. Johnson, *Matrix Analysis*, 2 ed., Cambridge University Press, 2012.

Hu, Lily and Yiling Chen, "Welfare and Distributional Impacts of Fair Classification," *arXiv:1807.01134 [cs, stat]*, July 2018. arXiv: 1807.01134.

Jayachandran, Seema, Monica Biradavolu, and Jan Cooper, "Using Machine Learning and Qualitative Interviews to Design a Five-Question Women's Agency Index," Technical Report w28626, National Bureau of Economic Research March 2021.

Johansson, Fredrik D., Uri Shalit, and David Sontag, "Learning Representations for Counterfactual Inference," June 2018. arXiv:1605.03661 [cs, stat].

Kasy, Maximilian and Rediet Abebe, "Fairness, equality, and power in algorithmic decision making," in "ICML Workshop on Participatory Approaches to Machine Learning" 2020.

Kent, David M., Jessica K. Paulus, David van Klaveren, Ralph D'Agostino, Steve Goodman, Rodney Hayward, John P.A. Ioannidis, Bray Patrick-Lake, Sally Morton, Michael Pencina, Gowri Raman, Joseph S. Ross, Harry P. Selker, Ravi Varadhan, Andrew Vickers, John B. Wong, and Ewout W. Steyerberg, "The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement," *Annals of Internal Medicine*, January 2020, *172* (1), 35–45. Publisher: American College of Physicians.

Kitagawa, Toru and Aleksey Tetenov, "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 2018, *86* (2), 591–616. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA13288.

Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt, "Delayed Impact of Fair Machine Learning," in "Proceedings of the 35th International Conference on Machine Learning," Vol. 80 of *Proceedings of Machine Learning Research* Stockholm, Sweden 2018, pp. 3156–3164.

McKenzie, David J., "Measuring inequality with asset indicators," *Journal of Population Economics*, June 2005, *18* (2), 229–260.

Mouzannar, Hussein, Mesrob I. Ohannessian, and Nathan Srebro, "From Fair Decision Making to Social Equality," *arXiv:1812.02952 [cs, stat]*, December 2018. arXiv: 1812.02952.

Nichols, Albert L. and Richard J. Zeckhauser, "Targeting Transfers through Restrictions on Recipients," *The American Economic Review*, 1982, *72* (2), 372–377.

Noriega, Alejandro, Bernardo Garcia-Bulle, Luis Tejerina, and Alex Pentland, "Algorithmic Fairness and Efficiency in Targeting Social Welfare Programs at Scale," *Bloomberg Data for Good Exchange Conference*, 2018.

Parker, Susan W. and Petra E. Todd, "Conditional Cash Transfers: The Case of *Progresa/Oportunidades*," *Journal of Economic Literature*, September 2017, *55* (3), 866–915.

Ravallion, Martin, "How Relevant Is Targeting to the Success of an Antipoverty Program?," *The World Bank Research Observer*, 2009, *24* (2), 205–231.

Rolf, Esther, Max Simchowitz, Sarah Dean, Lydia T. Liu, Daniel Björkegren, Moritz Hardt, and Joshua Blumenstock, "Balancing Competing Objectives with Noisy Data: Score-Based Classifiers for Welfare-Aware Machine Learning," in "" 2020.

Rubin, Donald B., "The Bayesian Bootstrap," *The Annals of Statistics*, 1981, *9* (1), 130–134. Publisher: Institute of Mathematical Statistics.

Saez, Emmanuel and Stefanie Stantcheva, "Generalized Social Marginal Welfare Weights for Optimal Tax Theory," *American Economic Review*, January 2016, *106* (1), 24–45.

Shao, Jun and Dongsheng Tu, *The Jackknife and Bootstrap* Springer Series in Statistics, New York, NY: Springer, 1995.

Skoufias, Emmanuel, Benjamin Davis, and Jere R. Behrman, "An evaluation of the selection of beneficiary households in the education, health, and nutrition program (PROGRESA) of Mexico," *International Food Policy Research Institute, Washington, DC*, 1999.

_ , _ , and Sergio de la Vega, "Targeting the Poor in Mexico: An Evaluation of the Selection of Households into PROGRESA," *World Development*, October 2001, *29* (10), 1769–1784.

_ , Sergio de la Vega, and Benjamin Davis, "Targeting the poor in Mexico," *FCND dicussion papers 103*, 2001.

_ , Susan W. Parker, Jere R. Behrman, and Carola Pessino, "Conditional Cash Transfers and Their Impact on Child Work and Schooling: Evidence from the PROGRESA Program in Mexico [with Comments]," *Economía*, 2001, *2* (1), 45–96. Publisher: Brookings Institution Press.

Skoufias, Vincenzo Di Maro Emmanuel, "Conditional Cash Transfers, Adult Work Incentives, and Poverty," *Journal of Development Studies*, 2008, *44* (7), 935–960.

Train, Kenneth E., *Discrete Choice Methods with Simulation*, 2 ed., Cambridge: Cambridge University Press, 2009.

UNDP, "Human Development Report 1990: Concept and Measurement of Human Development," Technical Report 1990.

Viviano, Davide, "Policy design in experiments with unknown interference," *Working Paper*, 2023.

Wager, Stefan and Susan Athey, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, July 2018, *113* (523), 1228–1242.

Wang, Fan, "The Optimal Allocation of Resources Among Heterogeneous Individuals," *Available at SSRN*, 2020.