What Do Policies Value?*

Daniel Björkegren† Columbia University

Joshua E. Blumenstock‡ U.C. Berkeley

Samsun Knight§ University of Toronto

March 12, 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.

†dan@bjorkegren.com

‡jblumenstock@berkeley.edu

§samsundknight@gmail.com

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

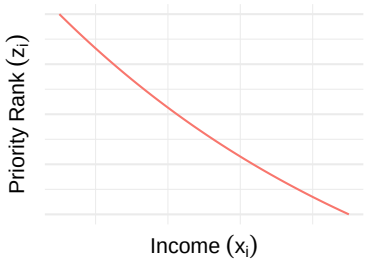
To demonstrate the intuition behind our method, we illustrate with a simple example in Figure 1. Consider the case of a single non-choice outcome and one dimension of heterogeneity, x_i, which corresponds to income. A policymaker allocates a program by ordering households by $z_i = Z(x_i)$, for some 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.

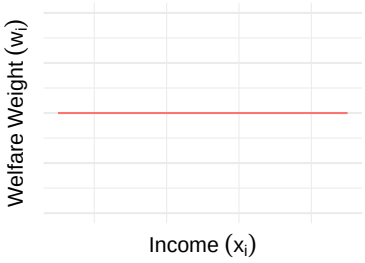
An allocation rule that prioritizes the poor (low x_i)

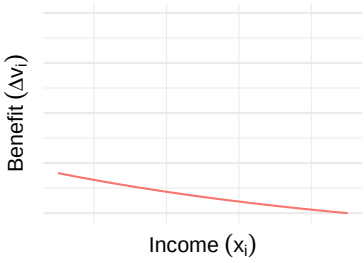


Could result from

(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 (-0.227, -0.038)
log(Income)		-0.237 (-0.252, -0.223)	-0.19 (-0.234, -0.138)
Household Size		0.116 (0.112, 0.119)	0.104 (0.085, 0.118)
Household Head Age		-0.02 (-0.021, -0.018)	-0.016 (-0.02, -0.01)
Education (Middle school or above)		-1.007 (-1.263, -0.85)	-0.727 (-0.952, -0.505)
Impact Weights			
Log consumption (per capita)	β_1		6.07 (4.04, 7.28)
Missed Schooling (per day)	β_3		-0.48 (-1.33, 0.02)
Sickness (per child sick day)	β_2		-0.05 (-0.51, 0.56)
Value Regardless of Impact	α		1
N_{rank}		7767	7767
N_{TE}		.	6784
Hypothesis Tests			p-value
Egalitarian	$\boldsymbol{\gamma} \equiv 1$		3.01e-16
Not Paternalistic	$\boldsymbol{\beta} \equiv 0$		5.12e-10
Egalitarian and Not Paternalistic	$\boldsymbol{\gamma} \equiv 1$, $\boldsymbol{\beta} \equiv 0$		3.96e-22

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.

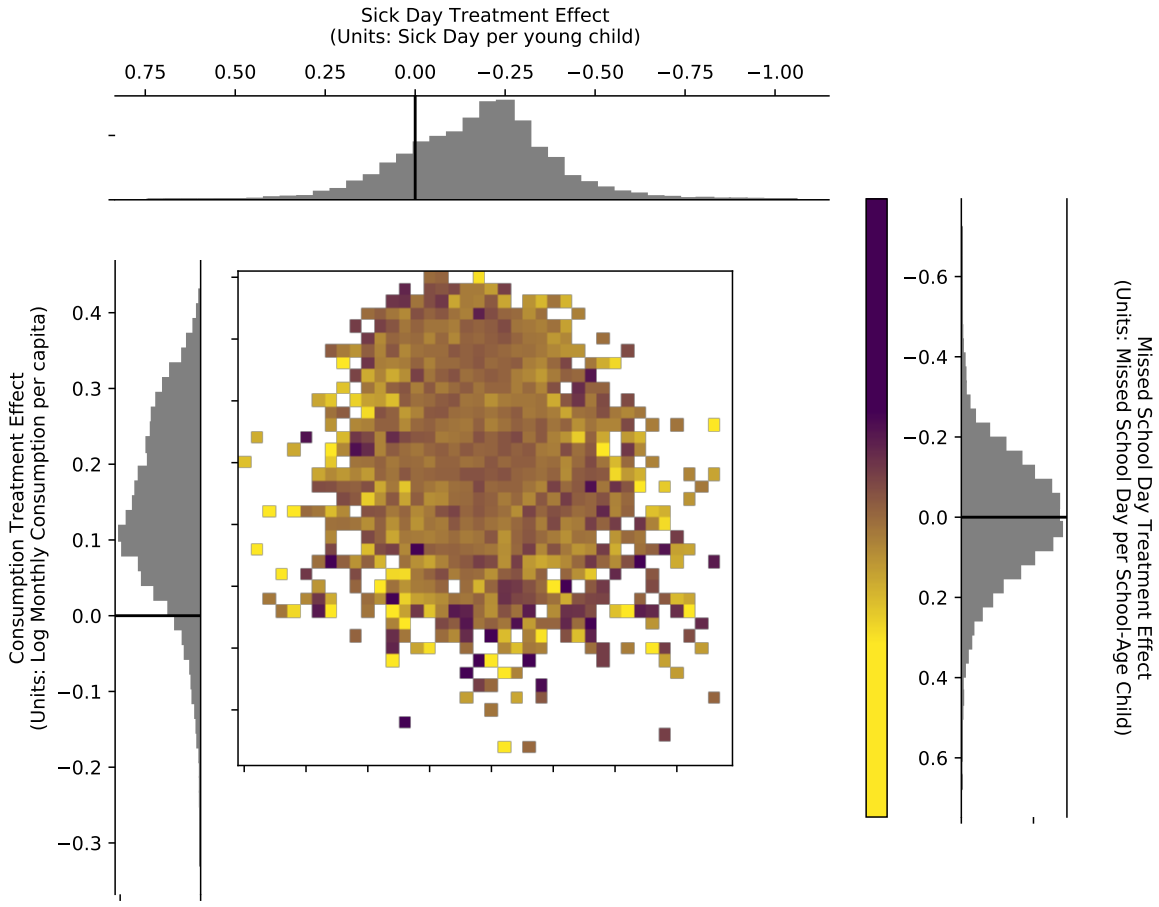
Table 1 column 2 reports the preferences that are consistent with the ranking $\boldsymbol{z}$. The first block of rows shows the implied welfare weights ($\boldsymbol{\gamma}$), and the second block shows implied impact weights ($\boldsymbol{\beta}$ and α). Because the policy ranked all households, 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 $log(\gamma)$				
Indigenous		-0.174 (-0.227, -0.038)	0.062 (0.011, 0.196)	0.065 (0.057, 0.072)
log(Income)		-0.19 (-0.234, -0.138)	-0.072 (-0.11, -0.039)	-0.071 (-0.257, 0.116)
Household Size		0.104 (0.085, 0.118)	0.086 (0.075, 0.096)	0.015 (-0.018, 0.049)
Household Head Age		-0.016 (-0.02, -0.01)	-0.001 (-0.004, 0.002)	0.004 (0.002, 0.005)
Educated		-0.727 (-0.952, -0.505)	-0.416 (-0.582, -0.3)	-0.065 (-0.099, -0.03)
Impact Weights				
Log Consumption (per capita)	β_1	6.07 (4.04, 7.28)	2.23 (1.49, 2.72)	4.37 (2.99, 5.75)†
Missed Schooling (per day)	β_3	-0.48 (-1.33, 0.02)	-0.32 (-0.71, -0.07)	-1.11 (-1.6, -0.62)†
Sickness (per child sick day)	β_2	-0.05 (-0.51, 0.56)	-0.01 (-0.21, 0.3)	-0.69 (-1.03, -0.35)†
Value Regardless of Impact	α	1	1	.
N_{rank}		7767	7767	.
N_{TE}		6784	6784	.
$N_{respondents}$		.	.	421*

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 421), 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.

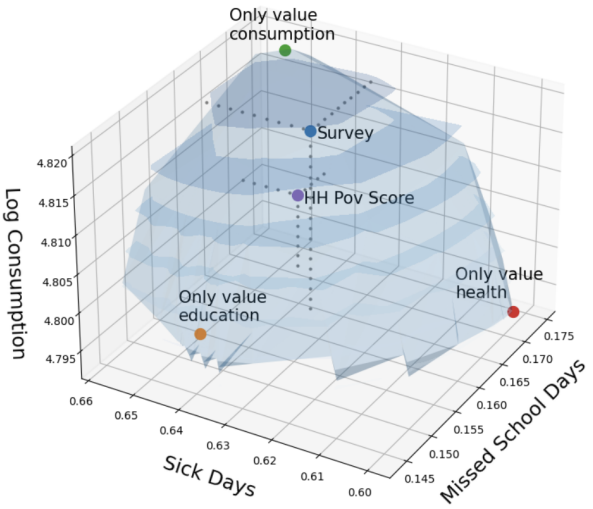
Prioritizing Specific Welfare Outcomes Most real-world policies balance multiple outcomes. For comparison, columns 4-6 of Table 3 present counterfactual allocations that would result in the extreme case where a policy was designed to improve only a single outcome. For instance, a policy that maximized 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	-0.77	-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.

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.

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.

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.

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.

What Do Policies Value?
Online Appendix

Daniel Björkegren Joshua E. Blumenstock Samsun Knight

This document includes additional information for the curious reader.

Contents

S1 Model Extensions

S1.1 Choice Model

The household's optimization of equation (7) yields first order conditions

$$\tilde{b}_{ij} g_j'(y_{ij}) = \eta_i \frac{\partial c}{\partial y_{ij}} \tag{S1}$$

The policy considers the utility of the household according to equation (2). If, for the moment, we consider infinitesimal changes in treatment status (T_i), the derivative of utility from the policy's perspective with respect to treatment is

$$\frac{du_i}{dT_i} = \sum_j \left[b_{ij} g_j'(y_{ij}) \right] \frac{dy_{ij}}{dT_i} + a.$$

Adding and subtracting each side of equation (S1), we obtain

$$\frac{du_i}{dT_i} = \sum_j \left[b_{ij} g_j'(y_{ij}) - \tilde{b}_{ij} g_j'(y_{ij}) + \eta_i \frac{\partial c}{\partial y_{ij}} \right] \frac{dy_{ij}}{dT_i} + a.$$

Recognizing that the budget constraint enforces that $\sum_j \frac{\partial c}{\partial y_{ij}} \frac{dy_{ij}}{dT_i} = \phi_i$, we obtain

$$\frac{du_i}{dT_i} = \sum_j \left[b_{ij} - \tilde{b}_{ij} \right] \underbrace{g_j'(y_{ij}) \frac{dy_{ij}}{dT_i}}_{\approx \Delta v_{ij}} + \eta_i \phi_i + a.$$

When we allow for welfare weights and approximate derivatives with their discrete counterparts, we obtain the generalized equation (9) for ΔS_i.

S1.2 Externalities

S1.2.1 Altruism

This section derives equations for rankings in cases where households share benefits with each other, as would occur with altruistic households. We first consider the household problem. Outcomes are functions of choices, so that each household i chooses y_{ij}, for each j, similar to Section 3.4. Each household i may also choose to contribute amount $\delta_{ii'} \geq 0$ to household i'.

Household i receives utility from its own outcomes, as well as from the outcomes of others,

$$\tilde{u}_i = \sum_j \tilde{b}_{ij} g_j\left(y_{ij}\right) + \tilde{a} \cdot T_i + \sum_{i' \neq i} \frac{\tilde{w}_{ii'}}{\tilde{w}_{ii}} \sum_j \tilde{b}_{i'j} g_j\left(y_{i'j}(\mathbf{T}, \boldsymbol{\delta}_{-i'})\right)$$

subject to the constraint

$$c(\mathbf{y}_i) + \sum_{i'} \delta_{ii'} = \mu_i + \phi_i T_i + \sum_{i'} \delta_{i'i}(\mathbf{T}, \boldsymbol{\delta}_{-i'})$$

where $\tilde{w}_{ii'}$ represents the welfare weight i places on the utility of i'. Let η_i represent the Lagrange multiplier on i's budget constraint.

The first order conditions for y_{ij} yield the same equation (S1) as without altruism, but transfers $\delta_{ii'}$ now link the utilities of households; for any pair i and i',

$$\eta_i \geq \frac{\tilde{w}_{ii'}}{\tilde{w}_{ii}} \eta_{i'}$$

where the inequality arises because transfers must be nonnegative. Together with the budget constraint, these equations define $\delta_{ii'}(\mathbf{T})$ and $y_{ij}(\mathbf{T})$.

Given household choices, we assume that the policy selects the vector of treatment statuses $\mathbf{T}$ jointly, to maximize S as defined in equations (1) and (2). From the perspective of the policy, the marginal value of treating i includes the effects on i (as in equation (9)), but now also includes spillover effects

$$\frac{dS}{dT_i}(\mathbf{T}) = \sum_{i'} w_{i'} \sum_j \left[b_{i'j} g_j'(y_{i'j}(\mathbf{T}))\right] \frac{dy_{i'j}}{dT_i}(\mathbf{T}) + a,$$

where $w_{i'}$ represents the welfare weight that the policy places on i'. The value of treating i' becomes a function of the treatment status of all households, $\mathbf{T}$, a nuance we will return to.

The equation can be rearranged as follows

$$\begin{aligned}
\frac{dS}{dT_i}(\mathbf{T}) =& w_i \left(\sum_j \left[b_{ij} g'_j(y_{ij}(\mathbf{T})) \right] \frac{dy_{ij}}{dT_i}(\mathbf{T}) + a \right) + \sum_{i' \neq i} w_{i'} \sum_j \left[b_{i'j} g'_j(y_{i'j}(\mathbf{T})) \right] \frac{dy_{i'j}}{dT_i}(\mathbf{T}) \\
=& w_i \left(\sum_j \left[(b_{ij} - \tilde{b}_{ij}) g'_j(y_{ij}(\mathbf{T})) + \eta_i(\mathbf{T}) \frac{\partial c}{\partial y_{ij}}(\mathbf{T}) \right] \frac{dy_{ij}}{dT_i}(\mathbf{T}) + a \right) \\
&+ \sum_{i' \neq i} w_{i'} \sum_j \left[(b_{i'j} - \tilde{b}_{i'j}) g'_j(y_{i'j}(\mathbf{T})) + \eta_{i'}(\mathbf{T}) \frac{\partial c}{\partial y_{i'j}}(\mathbf{T}) \right] \frac{dy_{i'j}}{dT_i}(\mathbf{T}) \\
=& w_i \left(\left[\sum_j \left(b_{ij} - \tilde{b}_{ij} \right) g'_j(y_{ij}(\mathbf{T})) \frac{dy_{ij}}{dT_i}(\mathbf{T}) \right] + \eta_i(\mathbf{T}) \left[\phi_i - \sum_{i'} \left(\frac{d\delta_{ii'}}{dT_i}(\mathbf{T}) - \frac{d\delta_{i'i}}{dT_i}(\mathbf{T}) \right) \right] + a \right) \\
&+ \sum_{i' \neq i} w_{i'} \left(\left[\sum_j \left(b_{i'j} - \tilde{b}_{i'j} \right) g'_j(y_{i'j}(\mathbf{T})) \frac{dy_{i'j}}{dT_i}(\mathbf{T}) \right] + \eta_{i'}(\mathbf{T}) \sum_{i''} \left[\frac{d\delta_{i''i'}}{dT_i}(\mathbf{T}) - \frac{d\delta_{i'i''}}{dT_i}(\mathbf{T}) \right] \right)
\end{aligned}$$

where the last line recognizes that the budget constraint enforces that $\sum_j \frac{\partial c}{\partial y_{ij}} \frac{dy_{ij}}{dT_i} = \phi_i + \sum_{i'} \left(\frac{\partial \delta_{i'i}}{\partial T_i} - \frac{\partial \delta_{ii'}}{\partial T_i} \right)$ and $\sum_j \frac{\partial c}{\partial y_{i'j}} \frac{dy_{i'j}}{dT_i} = \sum_{i''} \left(\frac{\partial \delta_{i''i'}}{\partial T_i} - \frac{\partial \delta_{i'i''}}{\partial T_i} \right)$. The equation accounts for all of the changes in transfers that result when i receives treatment. It can be approximated as

$$\Delta S_i(\mathbf{T}) \approx \overbrace{w_i \left(\sum_j \left[\left(b_{ij} - 1_{\{j \in \mathbb{J}_{choice}\}} \cdot \tilde{b}_{ij} \right) \Delta v_{ij}(\mathbf{T}_{-i}) \right] + \phi_i \eta_i(\mathbf{T}) + a \right)}^{\text{Gross benefit to } i} \tag{S2}$$

$$+ \underbrace{\sum_{i' \neq i} w_{i'} \sum_j \left[(b_{i'j} - 1_{\{j \in \mathbb{J}_{choice}\}} \tilde{b}_{i'j}) \Delta v^{ext}_{ii'j}(\mathbf{T}_{-i}) \right]}_{\text{Spillover internalities}} \tag{S3}$$

$$+ \underbrace{\sum_{i' \neq i} \left[-w_i \eta_i(\mathbf{T}) \Delta \delta_{i:ii'}(\mathbf{T}_{-i}) + w_{i'} \eta_{i'}(\mathbf{T}) \sum_{i''} \Delta \delta_{i:i''i'}(\mathbf{T}_{-i}) \right]}_{\text{Spillover transfers}} \tag{S4}$$

The first term captures gross benefits to household i, which are equivalent to equation (9) evaluated at $\mathbf{T}$. The second term captures internality benefits from altruism to each other household i', and the third term captures the welfare effects from net transfers on the budget constraint. $\Delta \delta_{i:i'i''}(\mathbf{T}_{-i})$ represents the change in transfers from i' to i'' when i changes from untreated to treated.[1]

Note that the policy may prefer not targeting a household that is altruistic, if that household values other households very differently than the policy. To see this, note that the

[1] This equation represents the difference between the welfare at $T_i = 1$ and at $T_i = 0$, but because values $\boldsymbol{\eta}(\mathbf{T})$ are approximated around a given value of T_i, the overall equation is a function of $\mathbf{T}$, not just $\mathbf{T}_{-i}$.

last term will depend on the correlation between the policy's preferences over households, and transfers (which will depend on the treated household's preferences). Altruism will lower the ranking of a household i if it would make net transfers to households i' ($\Delta\delta_{i:ii'} - \Delta\delta_{i:i'i} > 0$) which the policy prefers less ($w_i\eta_i > w_{i'}\eta_{i'}$).

The adjusted equation (S2) provides a starting point to extend our method. One would need to measure the impacts of treating household i not only on itself $\Delta v_{ij}(\mathbf{T}_{-i})$, but also on other households, on both outcomes $\Delta v^{ext}_{ii'j}(\mathbf{T}_{-i})$ and transfers $\Delta\delta_{i:i'i''}(\mathbf{T}_{-i})$. Those impacts should be computed as a function of the treatment status of others, $\mathbf{T}_{-i}$.

A conceptual challenge arises because $\Delta S_i(\mathbf{T}_{-i})$ now depends on $\mathbf{T}_{-i}$, so its implied ranking $\boldsymbol{z}(\mathbf{T})$ may also depend on who is ultimately treated. In practice, policies typically report a single ranking $\boldsymbol{z}$ which does not depend on who is ultimately treated. However, a single ranking can be rationalized with additional assumptions. One approach would be to assume separability, so that the marginal benefit of treating each household does not depend on others' treatment status, $\Delta S_i(\mathbf{T}_{-i}) \equiv \Delta S_i$. Alternately, one could assume the $\mathbf{T}$ under which the ranking $\boldsymbol{z}$ is evaluated: for instance, that the ranking is consistent with the final allocation $\mathbf{T}$. A challenge is that estimating these objects would require a sophisticated experiment, which may need to be informed by the end result (the treatment allocation $\mathbf{T}$).

Spillovers in PROGRESA The PROGRESA experiment, which randomized eligibility at the village level, measures a slightly different object. Imagine i indexed households within a village, ordered by decreasing score z_i, with i^* being the cutoff household, so that that household and all below receive benefits. The design estimates the effect on each eligible $i \leq i^*$ household of treating itself *and other eligibles* in the village: $\mathbb{E}_{i\leq i^*}\left[v_{ij}(\mathbf{T}_{\leq i^*}) - v_{ij}(\mathbf{0})\right]$ for $\mathbf{T}_{\leq i^*} = [\underbrace{1,...,1}_{i^*},\underbrace{0,...,0}_{N-i^*}]$. This corresponds to the Δv_{ij} we estimate in Section 4. In a model without spillovers, this object does not depend on the treatment status of others, $\mathbf{T}_{-i}$.

Using the same experiment, Angelucci and De Giorgi (2009) introduce two additional measures, which capture the average effect on *ineligible* households $i' > i^*$ of treating all eligible households i. One measure, which we denote by $\overline{\Delta v^{ext}_{ii'j}}(\mathbf{T}_{\leq i^*})$, captures average impacts on consumption, mimicking $\mathbb{E}_{i'>i^*}\left[v_{i'j}(\mathbf{T}_{\leq i^*}) - v_{i'j}(\mathbf{0})\right]$. The second, $\overline{\Sigma\Delta\delta_{i'':ii'}}(\mathbf{T}_{\leq i^*})$, captures the average impact on the transfer received by i', mimicking $\mathbb{E}_{i'>i^*}\left[\sum_i(\delta_{ii'}(\mathbf{T}_{\leq i^*}) - \delta_{ii'}(\mathbf{0}))\right]$. One can assume that if i were to become treated, incoming transfers would decrease by an equivalent amount. Note, however, that these estimated effects do not differ based on the characteristics of which i is treated, since they compare treated villages, where all eligibles $i \leq i^*$ are treated, to control villages, where none are. These estimates are thus constant

across i, and so would have only a minor effect on $\Delta S_i(\mathbf{T}_{-i})$ and the ranking it implies.[2]

Given a different experimental design, our approach could be further extended to test for altruism. In particular, if a pilot randomized the type of person i in each village that was treated, and measured the pairwise transfers between each i' and i'', we could estimate $\Delta v^{ext}_{i'ij}(\mathbf{T}_{-i'})$ and $\Delta \delta_{i:i'i''}(\mathbf{T}_{-i'})$ as a function of $\mathbf{T}_{-i'}$. Viviano (2023) proposes additional experimental designs that could better account for spillovers.

S1.2.2 Direct Effects

The choices of households could also differ from those preferred by the policy if household i's outcomes directly affect another household i''s utility. This may lead households to undervalue their impact on other households.

From the perspective of the policy, household i's utility is given by

$$u_i = \sum_j b_{ij} g_j(y_{ij}) + a \cdot T_i + \frac{1}{N-1} \sum_{i' \neq i} \sum_j d_{ii'j} g_j(y_{i'j})$$

where $d_{ii'j}$ represents the value i receives from i''s outcome j (from the perspective of the policy).

From the perspective of the household i, utility is similar

$$\tilde{u}_i = \sum_j \tilde{b}_{ij} g_j(y_{ij}) + \tilde{a} \cdot T_i + \frac{1}{N-1} \sum_{i' \neq i} \sum_j \tilde{d}_{ii'j} g_j(y_{i'j})$$

where $\tilde{d}_{ii'j}$ represents the value i perceives it receives from i''s outcome j. For nonchoice variables $j \notin \mathbb{J}_{choice}$, y_{ij} is a mechanical function of T_i as before. For variables $j \in \mathbb{J}_{choice}$, i selects y_{ij} to maximize $\tilde{u}_i$ subject to the budget constraint (equation (8)). Because it does not internalize its effects on others, the household faces the same first order condition as before, equation (S1), and makes the same choices, regardless of its perception of externalities.

[2] In particular, the ranking implied by equation (S2) with altruism would differ from the ranking implied by (9) in a few respects. The spillover transfers term would be identical except for $w_i \eta_i(\mathbf{T}_{-i})$: that is, although it assumes the same amount of transfer from each i and the same benefits to each i', it would account for different opportunity costs of that transfer from each i based on different welfare weights and constraints. Each element of the spillover internalities term would be identical but the sum will swap out the treated household i and so would differ by one element. Although this may affect the results, it does not capture a primary force that would cause spillovers to alter allocations: that targeting different households may yield different spillovers.

The policy then views the utility of treating i as

$$\Delta S_i \approx \sum_j \Delta v_{ij} \left[w(\mathbf{x}_i) \left(b_{ij} - 1_{\{j \in \mathbb{J}_{choice}\}} \tilde{b}_{ij} \right) + \underbrace{\frac{1}{N-1} \sum_{i' \neq i} w(\mathbf{x}_{i'}) d_{i'ij}}_{\text{Externalities}} \right] + w(\mathbf{x}_i) \left[\eta_i \phi_i + a \right] \tag{S5}$$

which now includes a new term representing the externality on others i'. Without restrictions on externalities $\mathbf{d}$, this new term complicates estimation because the policy's impact on i, Δv_{ij} may now be multiplied by a combination of the welfare weights on all households. However, in cases with sufficiently simple structure on externalities $\mathbf{d}$, direct application of our method yields results with a straightforward interpretation. Consider the following examples:

If there are positive externalities solely within groups of households that share the same welfare weights, such that $d_{ii'j} \equiv 0$ for any $w(\mathbf{x}_{i'}) \neq w(\mathbf{x}_i)$, then our method will estimate a $\boldsymbol{\beta}$ that combines weights on externalities as well as any internalities.[3]

Alternately, if i's outcome only matters to households in one group, then the welfare weights can be interpreted as the weights on that group. For example, consider a vaccination strategy for a contagious disease. Imagine the vaccine does not affect the utility of the people who receive it (so that $b_{ij} \equiv \tilde{b}_{ij} \equiv 0$, and $\phi_i = a = 0$), but mechanically benefits the susceptible people they are in contact with, with a potentially different effectiveness for each target (Δv_{ij}). Each person i is in contact with the same number of susceptible people (n), who are of only one type ($d_{ii'j} \equiv d > 0$ for a subset of i''s with identical $\mathbf{x}_{i'}$, and $d_{ii'j} \equiv 0$ for others; for example, nurses and nursing home staff have contact with elderly people). If our method is estimated on the allocation policy, Δv_{ij} will capture the differential effectiveness on different potential vaccination recipients, w will estimate the welfare weight on the different types of *susceptible* people they are in contact with, and β_j will estimate the average externality ($\frac{1}{N-1} \sum_{i' \neq i} d_{i'ij} = \frac{nd}{N-1}$).

S1.3 Continuous Treatment

If treatment is continuous and not binary, the policy selects $T_i \in [0, \infty)$ for each i. Utility from the perspective of the policy can still be written as in equation (2), with T_i redefined as continuous, but the procedure changes in two ways.

[3] In that case, $\beta_j(\mathbf{x}_i)$ will approximate $\left(b_{ij} - 1_{\{j \in choice\}} \tilde{b}_{ij} \right) + \frac{1}{N-1} \sum_{i' \neq i} w(\mathbf{x}_{i'}) d_{i'ij}$.

First, rather than estimating the effects of treatment as the difference $v_{ij}(1) - v_{ij}(0)$, we seek to estimate each function $v_{ij}(T)$, which we assume is concave. Given experimental variation in T that is not just binary, we can obtain estimates of the slope, $\frac{d\hat{v}_{ij}}{dT}$. Imposing a functional form, such as the quadratic $v_{ij}(T) = c_j(\tilde{\mathbf{x}}_i)T - \frac{1}{2}d_j(\tilde{\mathbf{x}}_i)T^2$, can simplify estimation.

Second, in an optimal allocation $\mathbf{T}$, the marginal returns from the perspective of the policy across individuals will be equated,

$$\frac{\partial S_i}{\partial T_i} = w(\mathbf{x}_i) \cdot \left(\sum_j b_{ij} \frac{dv_{ij}}{dT_i} + a \right) \equiv \xi$$

to a constant we call ξ.

Given estimates of these returns $\frac{d\hat{v}_{ij}}{dT_i}$, one can then estimate ω, $\boldsymbol{\beta}$, and α to minimize the distance to some constant return ξ at the observed treatment levels $\mathbf{T}$,

$$\omega(\mathbf{x}_i) \cdot \left(\sum_j \beta_j(\mathbf{x}_i) \frac{d\hat{v}_{ij}}{dT_i} + \alpha(\mathbf{x}_i) \right) \equiv \xi.$$

With quadratic utility, for instance, this yields

$$\omega(\mathbf{x}_i) \cdot \left(\sum_j \beta_j(\mathbf{x}_i) \left[\hat{c}_j(\tilde{\mathbf{x}}_i) - \hat{d}_j(\tilde{\mathbf{x}}_i) T_i \right] + \alpha(\mathbf{x}_i) \right) \equiv \xi.$$

Since these objects can be scaled arbitrarily we can simply select a convenient scale; e.g., $\xi = 1$. This recovers the welfare weights ω, the weights on different outcomes $\boldsymbol{\beta}$, and the benefit irrespective of outcomes α.

S1.4 Generalized Curvature in Utility Components

If utility functions are assumed to be linear ($\hat{g}_j(y) = y$) but the true utility functions $g_j(y)$ have curvature, the true impact of the program on utility component j is then:

$$\Delta v_{ij} = g_j(y_{ij}^1) - g_j(y_{ij}^0)$$

Taking a Taylor approximation from the factual level y_{ij}, we have $g_j(y_{ij} + \delta) \approx g_j(y_{ij}) + \delta \cdot g_j'(y_{ij})$. Thus for any $g_j(\cdot)$ we have:

$$\Delta v_{ij} \approx g_j(y_{ij}) - g_j(y_{ij}) + \Delta y_j(\tilde{\mathbf{x}}_i) \cdot g_j'(y_{ij}) = \Delta y_j(\tilde{\mathbf{x}}_i) \cdot g_j'(y_{ij})$$

We can then express the utility benefit of treating i in a nonchoice setting as:

$$\Delta S_i \approx \underbrace{w(\mathbf{x}_i)}_{\omega(\mathbf{x}_i)} \left[\sum_j \underbrace{b_{ij} g_j'(y_{ij})}_{\beta_j(\mathbf{x}_i, \{y_{ij}\})} \hat{\Delta y}_j(\tilde{\mathbf{x}}_i) + a \right]$$

This implies that if we do not specifically account for curvature and estimate a linear model, the welfare and impact weights we estimate ($\boldsymbol{\omega}$ and $\boldsymbol{\beta}$) are approximately a combination of the underlying welfare and impact weights ($\boldsymbol{w}$ and $\boldsymbol{b}$), respectively, and any curvature in the utility functions (g_j'), as long as the baseline value of the outcome (y_{ij}) is included as a characteristic along which these weights can vary ($\mathbf{x}_i$). If the true utility is linear, then $\boldsymbol{\omega}$ coincides with $\boldsymbol{w}$ and $\boldsymbol{\beta}$ with $\boldsymbol{b}$. Otherwise, utility curvature multiplies the weights.

S2 Identification

This section discusses conditions that are sufficient for nonparametric identification of $\omega(\cdot)$, $\boldsymbol{\beta}(\cdot)$, and $\alpha(\cdot)$. The parametric functional form in the paper permits slightly different assumptions.[4]

Let $\tilde{\mathbf{x}} = (\mathbf{x}, \mathbf{x}^+)$, so that $\tilde{\mathbf{x}}$ includes the covariates in $\mathbf{x}$ as well as excluded covariates $\mathbf{x}^+$. Let J be some fixed integer and assume that for $j = 1..J$, $\Delta v_j(\mathbf{x}, \mathbf{x}^+)$ is some known, observed function. For some unknown functions $f(\cdot)$, $\omega(\cdot)$, $\beta_j(\cdot)$, $\alpha(\cdot)$, and some errors ϵ, our data generating process is given by

$$\begin{aligned} z_i &= f\left[\omega(\mathbf{x}_i) \left(\sum_{j=1}^{J} \beta_j(\mathbf{x}_i) \Delta v_j(\mathbf{x}_i, \mathbf{x}_i^+) + \alpha(\mathbf{x}_i) \right) + \epsilon_i \right] \\ &= f\left[\sum_{j=0}^{J} \psi_j(\mathbf{x}_i) \Delta v_j(\mathbf{x}_i, \mathbf{x}_i^+) + \epsilon_i \right] \\ &= f\left[g(\mathbf{x}_i, \mathbf{x}_i^+) + \epsilon_i \right] \end{aligned}$$

where $\Delta v_0(\mathbf{x}_i, \mathbf{x}_i^+) \equiv 1$ and

$$\begin{cases} \psi_j(\mathbf{x}_i) := \omega(\mathbf{x}_i)\beta_j(\mathbf{x}_i) & \forall j = 1, \ldots, J \\ \psi_0(\mathbf{x}_i) := \omega(\mathbf{x}_i)\alpha(\mathbf{x}_i) \end{cases}$$

and $g(\mathbf{x}_i, \mathbf{x}_i^+) := \sum_{j=0}^{J} \psi_j(\mathbf{x}_i) \Delta v_j(\mathbf{x}_i, \mathbf{x}_i^+)$.

[4] We are grateful to Yassine Sbai Sassi for invaluable assistance in developing these arguments.

Our identification argument proceeds in two steps. We first present sufficient conditions to identify g and ψ_j (Theorem 1), which relies on econometric assumptions. We then discuss how ω, β_j, and α could be recovered from ψ_j, which requires restrictions justified by economic theory.

S2.1 Assumptions

Exogeneity

$$\epsilon \perp (\mathbf{x}, \mathbf{x}^+)$$

Regularity/Smoothness

R1) ω, β and α are continuously differentiable

R2) f is strictly increasing and continuously differentiable

R3) ϵ has a continuous distribution, with a continuous positive density $f_\epsilon > 0$

R4) The support of z is an interval (not necessarily bounded)[5]

R5) $(\mathbf{x}, \mathbf{x}^+)$ has a continuous distribution

R6) For any fixed $\mathbf{x}$, the function $\Delta v_j(\mathbf{x}, \cdot)$ is continuously differentiable and square integrable for all j (i.e. $\mathbb{E}(\Delta v_j(\mathbf{x}, \mathbf{x}^+)^2) < \infty$).

Identification

I1) For almost all $\mathbf{x}_1$ there exist $\mathbf{x}_1^+$ and $\mathbf{x}_2^+$ such that $g(\mathbf{x}_1, \mathbf{x}_1^+) \neq g(\mathbf{x}_1, \mathbf{x}_2^+)$

I2) For any given $\mathbf{x}$: $(1, \Delta v_1(\mathbf{x}, \cdot), \ldots, \Delta v_J(\mathbf{x}, \cdot))$ is linearly independent (as a family of functions, with the first element 1 being the constant function equal to 1)[6],[7]

I3) $E(\epsilon) = 0$ and $Var(\epsilon) = 1$.

[5] This condition would be redundant if we assume Δv is continuous and $(\mathbf{x}, \mathbf{x}^+)$ and ϵ are supported on an interval.

[6] Formally, linear independence here means that for all $\mathbf{x}$ and for all scalars $\lambda_0, \ldots, \lambda_J$: $\left[\forall \mathbf{x}^+ : \ \sum_j \lambda_j \Delta v_j(\mathbf{x}, \mathbf{x}^+) = 0\right] \iff \lambda_0 = 0, \ldots, \lambda_J = 0$

[7] Assumption I2) is weaker than alternate assumption I2'): for any $\mathbf{x}$, there exist $\mathbf{x}_0^+ \ldots \mathbf{x}_J^+$ such that the matrix $\left(\Delta v_j(\mathbf{x}, \mathbf{x}_{j'}^+)\right)_{0 \leq j \leq J, 0 \leq j' \leq J}$ is of full rank.

Note that these assumptions are slightly different than we use for the parametric estimator in the paper. This proof for general functions g requires f to be a strictly increasing function (making ties in rankings $\boldsymbol{z}$ unlikely); for the linear g used in the paper it may be possible to relax this.

S2.2 Identifying the ψ's

Theorem 1. *Assuming exogeneity and under conditions R1 to R6, I1 to I3, the functions ψ_j, for $j = 1..J$ are identified. ψ_0 is identified up to a uniform translation.*[8]

Proof. The proof proceeds in two steps: first we show that under the theorem's assumptions, the function g is identified; second, we show that, knowing g, and under our assumptions on Δv, ψ is identified.

By I1, the support of z contains more than 2 points. Without loss of generality, assume 0 and 1 are in that support.[9] Let a and b be two arbitrary (fixed) scalars with $a > b$ and assume $f^{-1}(0) = a$ and $f^{-1}(1) = b$. We begin by showing that for any such known a and b, f and g are identified.[10] We then establish that given assumption I3), the functions f and g are identified up to a uniform translation.

We show that in the model:

$$z = f\left(g(\mathbf{x}, \mathbf{x}^+) + \epsilon\right)$$

with $f^{-1}(0) = a$ and $f^{-1}(1) = b$, the functions f and g are identified. For that, we check that our assumptions imply the assumptions of Corollary 1 in Chiappori et al. (2015):

- **Assumption A1:** Follows immediately from exogeneity and assumption R3.
- **Assumption A2:** Follows from exogeneity and assumption R5.
- **Assumption A3:** Follows from R4, the assumption that 0 is in the support of z is not necessary as pointed out in Chiappori et al. (2015).
- **Assumption A4:** f being continuously differentiable and (strictly) increasing (R2), its derivative is strictly positive and therefore its inverse is also continuously differentiable.
- **Assumption A5:** Follows from R1 and R6 and the fact that $g(\mathbf{x}, \mathbf{x}^+) := \sum_{j=0}^{J} \psi_j(\mathbf{x}) \Delta v_j(\mathbf{x}, \mathbf{x}^+)$.

[8] In fact, one can show the function f and the density f_ϵ are also identified under the same assumptions.
[9] Any other points in the support would work similarly.
[10] Which means that we should index f and g by a and b at this point. We omit that for notational ease.

- **Assumption A6:** Following Chiappori et al. (2015) (equation 5), define

$$\begin{cases} \phi_i(z|\mathbf{x},\mathbf{x}^+) := -\frac{\partial g(\mathbf{x},\mathbf{x}^+)}{\partial x_i} f_\epsilon(f^{-1}(z) - g(\mathbf{x},\mathbf{x}^+)) & \text{for } i \leq dim(\mathbf{x}) \\ \phi_{i+dim(\mathbf{x})}(z|\mathbf{x},\mathbf{x}^+) := -\frac{\partial g(\mathbf{x},\mathbf{x}^+)}{\partial x_i^+} f_\epsilon(f^{-1}(z) - g(\mathbf{x},\mathbf{x}^+)) & \text{for } i \leq dim(\mathbf{x}^+) \end{cases}$$

  Assumption A6 requires that there exists some $i = 1..dim(\mathbf{x}) + dim(\mathbf{x}^+)$ and there exists $(\mathbf{x},\mathbf{x}^+)$ such that $\phi_i(z|\mathbf{x},\mathbf{x}^+) = 0$ for all z. Given that by assumption R3, $f_\epsilon > 0$, therefore $[\exists z : \phi_i(z,\mathbf{x},\mathbf{x}^+) = 0] \iff \frac{\partial g(\mathbf{x},\mathbf{x}^+)}{\partial x_i} = 0$ so the set of points $(\mathbf{x},\mathbf{x}^+)$ such that $\phi_i(z,\mathbf{x},\mathbf{x}^+) = 0$ is non empty for some i by assumption I1.

- **Assumption A7:** Since both $\mathbf{x}$ and $\mathbf{x}^+$ are exogenous, the assumption is satisfied in our setting.[11]

Therefore, g is identified. By assumption I2, this implies that ψ_j is identified for $j = 0..J$. To show that formally, fix x and note that by the linear independence of the family $(\Delta v_j(\mathbf{x},\cdot))_{j=0}^J$ (I2) and the fact that $(\Delta v_j(\mathbf{x},\cdot))_{j=0}^J$ are all square integrable (R6), then they form a base for a finite dimensional Euclidean space with the usual inner product in L^2: $< \Delta v_i(\mathbf{x},\cdot), \Delta v_j(\mathbf{x},\cdot) >= E(\Delta v_i(\mathbf{x},\mathbf{x}^+)\Delta v_j(\mathbf{x},\mathbf{x}^+))$. A well known result from linear algebra is that the Gram matrix $G(\mathbf{x}) := (< \Delta v_i(\mathbf{x},\cdot), \Delta v_j(\mathbf{x},\cdot) >)_{0\leq i\leq J, 0\leq i\leq J}$ is invertible (i.e. of full rank $J+1$) if and only if the family $(\Delta v_j(\mathbf{x},\cdot))_{j=0}^J$ is independent (see for instance Horn and Johnson (2012) theorem 7.2.10).
Denoting $\psi(\mathbf{x}) := (\psi_0(\mathbf{x}),\ldots,\psi_J(\mathbf{x}))'$, note that

$$G(x)\psi(\mathbf{x}) = (E(\Delta v_1(\mathbf{x},\mathbf{x}^+)g(\mathbf{x},\mathbf{x}^+)),\ldots,E(\Delta v_J(\mathbf{x},\mathbf{x}^+)g(\mathbf{x},\mathbf{x}^+)))'$$

therefore

$$\psi(\mathbf{x}) = G(\mathbf{x})^{-1}(E(\Delta v_1(\mathbf{x},\mathbf{x}^+)g(\mathbf{x},\mathbf{x}^+)),\ldots,E(\Delta v_J(\mathbf{x},\mathbf{x}^+)g(x,\mathbf{x}^+)))'$$

which identifies $\psi(\mathbf{x})$ as desired.[12]

Finally, remember that all the objects identified thus far - f, g and ψ_j - are identified up to a choice of pre-images a and b. Let's assume two such pre-images (a,b) and $(\tilde{a},\tilde{b})$

[11] To formally check that, one can add some "virtual" unrelated regressor ζ (say an independent coin flip) to the set of regressors that we label as "the potentially endogenous regressor" $X_{-I} := \zeta$, following Chiappori et al. (2015)'s notation. Then we let the instruments be $Z = X_{-I} = \zeta$. This would allow us to identify g as a function of *all* regressors, including X_{-I}. However, given that by construction, X_{-I} is a independent of everything else, g is independent of X_{-I} (or ζ).

[12] If instead of I2), we use the stronger assumption I2') introduced in the footnote from last page, then identifying ψ would not require that we go through the theorem on Gram matrices.

rationalize our data and prove they must be equal.

Our data is generated by the two (statistically indistinguishable) models:

$$f_{a,b}^{-1}(z) = g_{a,b}(\mathbf{x}, \mathbf{x}^+) + \epsilon_{a,b}$$

and

$$f_{\tilde{a},\tilde{b}}^{-1}(z) = g_{\tilde{a},\tilde{b}}(\mathbf{x}, \mathbf{x}^+) + \epsilon_{\tilde{a},\tilde{b}}$$

Define the function Θ by $\Theta = \mu + \lambda f_{a,b}^{-1}$, with $\lambda := \frac{f_{\tilde{a},\tilde{b}}^{-1}(1) - f_{\tilde{a},\tilde{b}}^{-1}(0)}{f_{a,b}^{-1}(1) - f_{a,b}^{-1}(0)} = \frac{\tilde{a}-\tilde{b}}{a-b}$ and $\mu := f_{\tilde{a},\tilde{b}}^{-1}(0) - \lambda f_{a,b}^{-1}(0)$ so that $\Theta(0) = f_{\tilde{a},\tilde{b}}^{-1}(0) = \tilde{a}$, $\Theta(1) = f_{\tilde{a},\tilde{b}}^{-1}(1) = \tilde{b}$ and :

$$\Theta(z) = \mu + \lambda g_{a,b}(\mathbf{x}, \mathbf{x}^+) + \lambda \epsilon_{a,b}$$

therefore, by our earlier identification result, it must be that $\Theta = f_{\tilde{a},\tilde{b}}^{-1}$, i.e. $f_{\tilde{a},\tilde{b}}^{-1} = \mu + \lambda f_{a,b}^{-1}$.[13]

By exogeneity and assumption I3): $Var(\epsilon_{a,b}|\mathbf{x}, \mathbf{x}^+) = Var(\epsilon_{\tilde{a},\tilde{b}}|\mathbf{x}, \mathbf{x}^+) = 1$ hence:

$$Var(f_{a,b}^{-1}(z)|\mathbf{x}, \mathbf{x}^+) = Var(f_{\tilde{a},\tilde{b}}^{-1}(z)|\mathbf{x}, \mathbf{x}^+) = 1$$

however, since $f_{\tilde{a},\tilde{b}}^{-1} = \mu + \lambda f_{a,b}^{-1}$, we get that $\lambda^2 = 1$, or $\lambda = 1$ (because $\lambda > 0$), implying $f_{\tilde{a},\tilde{b}}^{-1} = \mu + f_{a,b}^{-1}$. Likewise, $g_{\tilde{a},\tilde{b}} = \mu + g_{a,b}$, which, according to our identification result on the ψ's, implies that $\psi_{0,\tilde{a},\tilde{b}} = \mu + \psi_{0,a,b}$.

□

Considering this second step, note that in absence of such an exclusion restriction, the ψ_j's are not identified. To see this clearly, note that if $\mathbf{x}^+ = \emptyset$, then any function $g(\mathbf{x}_i)$ could be equally well explained by setting $\psi_0(\mathbf{x}_i) = g(\mathbf{x}_i)$ and $\psi \equiv 0$ for $j \geq 1$. Separating the contribution of the different ψ_j's requires variables $\mathbf{x}^+$ to affect g but be excluded from the ψ_j's.

S2.3 Recovering ω, α and the β_j's

Given theorem 1 and that

$$\begin{cases} \psi_j(\mathbf{x}) := \omega(\mathbf{x})\beta_j(\mathbf{x}) & \forall j = 1, \ldots, J \\ \psi_0(\mathbf{x}) := \omega(\mathbf{x})\alpha(\mathbf{x}) \end{cases}$$

[13] Importantly, note that the identification result we established does not rely on the assumption $Var(\epsilon) = 1$.

it becomes clear that the identification of ω, β, and α requires some additional "normalization" assumption to remove the extra degree of freedom in the specification. From an econometric standpoint, the choice of normalization is arbitrary, but the properties of the functions of interest ω, α and the β's are affected by the choice (such as their magnitudes, signs, or curvature/dispersion). Thus, the choice of normalization should be dictated by economic theory.

In the empirical application we apply the following normalization:

Normalization 1: $\beta_j(\mathbf{x}) \equiv \beta_j$ **for all** j**,** $\alpha(\mathbf{x}) \equiv \alpha$**,** $|\alpha| = 1$**, and** $\omega(\mathbf{x}) > 0$**.** The impact weights are the same across households ($\beta_j(\mathbf{x}) \equiv \beta_j$ for all j) and are normalized relative to a constant base value ($\alpha(\mathbf{x}) \equiv \alpha$ with $|\alpha| = 1$). This base value may be positive (making the program a good) or negative (making it a bad). Homogeneity in impact weights implies that if two households would attain the same impacts from the program, but are ranked differently, the difference must come from welfare weights. Welfare weights are positive $\omega(\mathbf{x}) > 0$, so that the program may not be a good for some households and a bad for others.

One could also use alternate assumptions to ease this restriction, for example:

Normalization 2: $\beta_j(\mathbf{x}) \equiv \beta_j$ **for some** j**, and** $\alpha(\mathbf{x}) \equiv 1$**.** This implies that the base value of the program is assumed to be good, and at least one impact weight is homogeneous.[14] When $\alpha(\mathbf{x}) \equiv 1$, the method identifies $\psi_0(\mathbf{x}) = \omega(\mathbf{x}) + \mu$ for some location shift μ. Then for a constant β_j, $\psi_j(\mathbf{x}) = \omega(\mathbf{x})\beta_j = \beta_j\psi_0(\mathbf{x}) - \beta_j\mu$. We can then solve for β_j and μ. Here, the choice was arbitrarily made to set $\alpha \equiv 1$; the choice could have been to set any other β_j to 1 and the same discussion would ensue.

Other normalization assumptions may also be possible.

S3 Data Cleaning Process

The data for the evaluation of PROGRESA is composed of household survey responses from a sample of 506 villages from seven states. Surveys were conducted in three different years: a baseline survey was fielded in October 1997, and two follow-up surveys in October 1998 and November 1999. However, the baseline survey included fewer questions, so we rely on the later surveys for detailed information on household outcomes, only using information from the baseline for data on pre-period income and select household characteristics. Villages were

[14] One could alternately assume it is a bad so that $\alpha(\mathbf{x}) \equiv -1$.

randomly assigned to a treatment group that received the program at the beginning, or a control group that received it two years later. For all households, a poverty index score was computed; all households in treatment villages below a score threshold were eligible for the program's transfers.

We compute a measure of average household monthly per capita consumption based on the survey responses. The October 1998 and November 1999 surveys ask households about the quantity consumed, quantity purchased and amount of money spent on 36 common food items, as well as expenditure for several non-food categories (in weekly/monthly/semi-annual amounts). We use the information regarding quantity purchased and amount spent to construct household-specific prices which are then multiplied by the quantity consumed (this helps to account for the fact that households consume food that is self-produced in addition to bought). If household-specific information is missing, we use locality, municipality or state average prices (the finest level available). We follow Angelucci and De Giorgi (2009) in counting each child as 0.73 people when computing per capita consumption.

S4 Preference Survey

We conducted a survey of Mexican residents to elicit their preferences for different allocations of social welfare programs. We solicited responses to a survey from a nationally representative sample of computer users in Mexico, through a Qualtrics survey panel (see https://www.qualtrics.com/).

S4.1 Survey Design

After obtaining consent and an initial information screen, participants were asked their preferences for allocating benefits to different types of households. The survey was given in (Mexican) Spanish. First, respondents were asked to select, from a list, which attributes the government should consider when prioritizing which households receive cash transfers (age, income, household size, education, whether agricultural, indigenous, gender, and demographics of each person in the household). Second, subjects were asked to make monetary allocation decisions between different households using multiple price lists (see Figure S4 for an example). In each, one focal attribute differed between the households, and two other control attributes were held fixed. We randomized which controls were included, the order they were presented, and the scale of the tradeoff.[15] Each subject filled in one price list for each focal attribute.

[15]Each participant saw base tradeoff numbers multiplied by 1x, 2x, or 3x, selected at random.

Third, for a particular household, subjects were asked to make allocation decisions between directly obtaining money and improvements in education, child health, and consumption using multiple price lists (see Figure S5). The description of the household included three randomly selected control attributes. Fourth, subjects were asked about their preferences for paternalism over childrens' outcomes, on subjective Likert scales and on a quantitative scale. Finally, subjects were asked for basic demographics.

S4.2 Estimation

We use the survey responses to estimate welfare weights $\boldsymbol{\gamma}$ and impact weights $\boldsymbol{\beta}$. To identify $\boldsymbol{\gamma}$, we compare transfer amounts (where other impacts are assumed to be $\Delta g_j(x_i) = 0$). If individual i differs from i' only in attribute k and the crossover point is a transfer to i of a and to i' of b, then

$$\gamma_{-k}^{x_{i,-k}} \gamma_k^{x_{i,k}} a = \gamma_{-k}^{x_{i',-k}} \gamma_k^{x_{i',k}} b$$
$$\gamma_k = \left(\frac{b}{a}\right)^{\frac{1}{x_{i,k} - x_{i',k}}}$$

We identify $\boldsymbol{\beta}$ in three steps. First, we hold fixed household attributes, and ask respondents how that specific household would trade off a money payment against an impact on outcome j, for each outcome. If the crossover point is a and $\Delta g_j(x_i) = b$, then this identifies how that household would trade off the outcome against a relaxation of the budget constraint, $\tilde{b}_j/\eta_i = a/b$. Second, we address the possibility that a policy might value outcomes differently from the households themselves,[16] which identifies the ratio $b_j/\tilde{b}_j$. Third, we combine these estimates to form an estimate of the resulting weight in the decision rule for outcomes that are choices, as in Section 3.4: $\beta_j = b_j - \tilde{b}_j \propto \frac{1}{N} \sum_i \tilde{b}_j/\eta_i \left(b_j/\tilde{b}_j - 1\right)$, where the latter is proportional up to the factor η_i. These unscaled results are reported in Appendix Table S15. In Table 2 column 3, we scale the resulting coefficients so the average magnitude is the same as the estimated coefficients.

[16] The survey contains a separate question about how much the government should care about the welfare of children in low income households. This question allows respondents to select how much the government should weigh childrens' welfare on each outcome *relative to parents*. Possible choices range from half as much as parents to twice as much as parents.

S4.3 Subjective Results

In subjective responses, respondents voiced support for paternalism for children, and belief in externalities. These were assessed on Likert scales to three prompts for each outcome, and are summarized in Table S9. Overall, the majority of respondents agreed that some children may require direct support from the government, and disagreed that it was enough to trust parents to do what is best for children.

S4.4 Validation

The design included several checks to ensure that respondents took the survey seriously. First, prior to the survey, participants were asked, 'We care about the quality of our survey data and hope to receive the most accurate measure of your opinions, so it is important to us that you thoughtfully provide your best answer to each question in the survey. Do you commit to providing your thoughtful and honest answers to the questions in this survey?' Only participants who answered 'I will provide my best answers' were invited to continue with the survey. Second, after reading the instructions, participants responded to five simple questions to validate survey comprehension. In order to complete the study, participants had to respond correctly. Third, the survey included controls to ensure that participants spent adequate time on each question. The submit button for the main exercises appeared only after a 5-second delay.[17] Additionally, participants who were completing the survey too quickly (less than half the median elapsed time in the pilot survey) were removed from our sample, following a standard protocol used by Qualtrics. Fourth, in the final demographic survey, respondents were asked to rate the following three statements along the same Likert scale ranging from 'Strongly Disagree' to 'Strongly Agree': 'I made each decision in this study carefully', 'I made decisions in this study randomly', and 'I understood what my decisions meant.' A careful respondent should agree with the first and last statement but disagree with the middle; agreement or disagreement with all statements reveals that a respondent made careless decisions. We restrict the sample to only respondents who disagreed that they had made decisions randomly. 96% of respondents agreed with the first and last statement, and disagreed with the middle; 60% did so strongly.

There was an optional comment box at the conclusion of the survey; 47% of respondents filled in a comment, suggesting high levels of engagement. Although some respondents used

[17]The implementation of this in Qualtrics made it possible for participants to advance if this time had elapsed, even if a multiple price list question had not been answered. For this reason, a handful of participants did not respond to all questions.

the box to indicate some confusion with the user interface, several responded affirmatively to the approach of basing policy on resident preferences, such as (translated to English):

- 'Excellent survey. Hopefully it will help decide which households need more help.'
- 'I hope this type of survey will be applied throughout the country so that it will be implemented'
- 'Let's hope the government relies on the results'
- 'I hope that there will be more surveys on these and that public resources will actually be allocated for infant feeding and education supporting single parents.'
- 'I thought it was very important because it makes you reflect on the families and needs that exist in the country'
- 'Very good study but they should do one for people between 20 and 58 years old especially with disabilities who are the most forgotten and vulnerable groups...'

S5 Treatment Effect Specification

S5.1 Causal Forests

Our framework allows for treatment effects estimated using alternate methods. In this section, we present results using an alternative estimator to demonstrate this flexibility and robustness.[18] Causal forests (Wager and Athey, 2018) allow treatment effects to differ more flexibly according to household covariates than in the baseline OLS model.

Figure S3 shows the distribution of treatment effects estimated using causal forest across several outcomes. Feature importances are presented in Table S3, which show that impacts on log consumption covary especially strongly with indigenous status, similar to our OLS estimates (Table S2) and Djebbari and Smith (2008).

Most importantly, when we use the estimates of treatment effects from causal forests as $\Delta\hat{v}_j$ in our framework, the primitives that are subsequently estimated are similar to those based on the OLS estimates. This comparison is done in Appendix Table S8 (i.e., comparing column 1, which uses OLS, to column 3, which uses causal forests). We also present causal forest versions of exhibits that parallel our full set of main exhibits and robustness checks in Section S6.

[18] Note that we keep the variable set consistent across estimators, and only change the first stage estimator.

S5.2 Assessing Attenuation and Misspecification

Our second-stage estimator uses first-stage predictions as covariates in a plug-in estimator. To assess the possibility that measurement error in first-stage estimates could bias second-stage estimates, we use Monte Carlo simulations inspired by our empirical setting. Results are shown in Table S11. The first column reports how treatment effects are estimated: either known (no first stage error), OLS (corresponding to our main empirical method), or OLS+SIMEX (an error-corrected version we describe below). We assess results under complex specifications ($K_{True} = K_{Est} = 22$), simple specifications ($K_{True} = K_{Est} = 5$), and a misspecification where the true specification is complex but the estimated model is simple ($K_{True} = 22$ but $K_{Est} = 5$). Bias and absolute error increases when treatment effects must be estimated. However, bias is relatively small when the specification is complex; it is slightly larger when the estimation specification is simple (regardless of whether the DGP is simple or complex).

We next test the SIMEX (SIMulation EXtrapolation) procedure of Cook and Stefanski (1994) to correct for measurement error in first-stage estimates, following the implementation of Chilet (2017). The procedure adds a series of different amounts of noise to the first stage, estimates the second stage for each, and then extrapolates to the second stage the estimates that would result if noise were altogether removed from the first stage. We apply SIMEX in three steps:

1. First, for each of our heterogeneous treatment effect dimensions, we add simulated normal measurement error ϵ_{ij} to each $\Delta\hat{v}_{ij}$, where ϵ_{ij} is drawn from the distribution $N(0, \zeta \cdot \sigma^2_{ij})$, for $\zeta \in \{0, 0.5, 1, 2\}$, and where σ^2_{ij} is the jth entry of the Jx1 prediction variance σ^2_i of our first-stage estimates for that particular observation i. We approximate this variance using the covariance matrix of the first stage coefficients, taking $\sigma^2_i = \tilde{\mathbf{x}}_i \Sigma_j$, where Σ_j is the covariance matrix of the first-stage coefficients and $\tilde{\mathbf{x}}_i$ is the row of first-stage covariates corresponding to observation i. Because this only accounts for variance in the parameters, it is likely to underestimate the total error and thus moderate the correction. We take 25 independent draws r of measurement error for each ζ, and then estimate second stage models using $\Delta\hat{v}_{ijr} = \Delta\hat{v}_{ij} + \epsilon_{ijr}$ as our first-stage data for each given draw.

2. Second, for each second-stage parameter, we estimate the estimated parameter value as a quadratic function of ζ, based on the $25 \times 4 = 100$ observations of second-stage model estimates for each ζ across 25 draws. This estimated quadratic model over ζ gives us a

smooth function, $f_\theta(\zeta)$, of how each second-stage parameter θ estimate changes with measurement error, for $\theta \in \{\beta, \omega, \alpha\}$.

3. Then, $\theta^{SIMEX} = f_\theta(-1)$ corresponds to the estimate of parameter θ, removing the forecasted impact of measurement error (as captured by the covariance matrix).

When we apply this SIMEX procedure in Monte Carlo simulations, it reduces bias in estimates arising from first-stage measurement, particularly in $\boldsymbol{\beta}$, as shown in Table S11. The gap between the standard and SIMEX estimates may thus be helpful as a diagnostic for the amount of bias.

We apply SIMEX to the PROGRESA application in Table S12. An illustration of the estimated function and the average second-stage parameter estimates is presented in Figure S6. We find that coefficient estimates are qualitatively similar regardless of whether SIMEX is applied. The SIMEX estimates yield slightly larger and more-significant weights on indigenous status and log consumption, suggesting that our main estimates may be slightly conservative due to modest attenuation bias. Given how modest this correction is, we maintain the standard OLS specification as our main first-stage specification. In applications where estimates are more substantially affected, one might use a bias correction for first stage measurement error.

References

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.

Chiappori, Pierre-André, Ivana Komunjer, and Dennis Kristensen, "Nonparametric 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.

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.

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

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.

S6 Additional Tables

Table S1: Descriptive Statistics (Midline)

	October 1998 mean
Head of household:	
... Is indigenous	0.38
... Age	41.61
... Education (Middle school or higher)	0.07
... Is male	0.94
... Is an agricultural worker	0.63
Household size	
... Number of children less than 6 years old	1.93
... Number of children 6-16 years old	2.76
... Number of adults 17+ years old	3.12
Log monthly average per capita consumption (log pesos)	5.10
Average number of days a school-age child misses school	0.32
Average number of days a young child is sick	1.08
Assigned to treatment group	0.60
N	7430

Notes: Table shows midline statistics (October 1998) for households present in our endline estimation sample (November 1999), among households surveyed in both. Sample restricted to households with children in the relevant age categories for health and schooling at endline (at least one child of age 0-5 y.o. and at least one 6-16 y.o.). Middle school education defined as 8 years or more of education. Number of days a young child is sick, and number of days a school-age child misses school, are computed as an average over the number of children in the respective age group in the household.

Table S2: Treatment Effect Coefficient Estimates: OLS

	Log Consumption (Monthly avg. per person, in pesos)	Schooling (Avg. days school missed per child)	Health (Avg. sick days per child)
Treatment	0.1289 (0.123)	-0.3548 (0.246)	-0.7254 (0.447)
Treatment X head indigenous	0.1539 (0.029)	0.0336 (0.058)	0.0502 (0.105)
Treatment X log(Income 1997)	-0.0097 (0.022)	-0.0095 (0.043)	0.0614 (0.079)
Treatment X num adults	-0.0341 (0.041)	0.0053 (0.082)	-0.1029 (0.149)
Treatment X head age	0.0005 (0.002)	0.0078 (0.004)	0.0032 (0.006)
Treatment X head education	0.0128 (0.062)	0.1207 (0.124)	-0.0819 (0.225)
Treatment X male head of household	-0.0263 (0.068)	0.0962 (0.135)	0.1592 (0.245)
Treatment X head agricultural worker	0.0373 (0.031)	-0.1063 (0.062)	-0.0795 (0.112)
Treatment X num child less than 2 yrs	-0.0271 (0.017)	0.0961 (0.034)	0.0096 (0.062)
Treatment X num child 3 to 5 yrs	-0.0288 (0.02)	-0.0562 (0.04)	0.107 (0.073)
Treatment X num child 6 to 10 yrs	0.0421 (0.016)	0.0116 (0.031)	-0.0124 (0.057)
Treatment X num boys 11 to 14 yrs	0.0116 (0.022)	-0.045 (0.045)	-0.0085 (0.081)
Treatment X num girls 11 to 14 yrs	0.0173 (0.022)	-0.0269 (0.045)	0.0443 (0.081)
Treatment X num boys 15 to 19 yrs	-0.0161 (0.039)	0.0085 (0.079)	0.0187 (0.143)
Treatment X num girls 15 to 19 yrs	-0.0081 (0.038)	0.0139 (0.077)	0.1264 (0.139)
Treatment X num men 20 to 34 yrs	0.035 (0.05)	0.0449 (0.1)	0.2963 (0.182)
Treatment X num women 20 to 34 yrs	0.0127 (0.05)	-0.0129 (0.101)	-0.1266 (0.183)
Treatment X num men 35 to 54 yrs	0.0664 (0.057)	-0.0436 (0.113)	0.0986 (0.205)
Treatment X num women 35 to 54 yrs	-0.0072 (0.057)	0.0253 (0.115)	0.0754 (0.208)
Treatment X num men at least 55 yrs	-0.0446 (0.07)	-0.2913 (0.14)	0.0671 (0.255)
Treatment X num women at least 55 yrs	-0.0026 (0.06)	0.0306 (0.119)	0.0851 (0.217)
Baseline Covariates	X	X	X
R^2	0.159	0.012	0.029
N_{TE}	6784	6784	6784

Notes: OLS coefficients of household characteristics interacted with treatment on three outcome dimensions: log consumption (log monthly per capita consumption), schooling (number of missed school days per child), and health (number of sick days per child). Standard errors in parentheses. Schooling and health sick days / missed school days measured over 28 days prior to survey. Baseline covariates include the covariates without treatment interactions, e.g. head age, as well as a constant term.

Table S3: Feature Importance Estimates: Causal Forest

	Log Consumption Monthly per capita (pesos)	**Schooling** # days missed school per child	**Health** # Sick days per child
head indigenous	0.319	0.009	0.012
log household income 97	0.230	0.182	0.273
head age	0.123	0.319	0.239
num child 3 to 5 yrs	0.011	0.070	0.165
num child less than 2 yrs	0.017	0.157	0.025
num adults	0.093	0.033	0.030
num child 6 to 10 yrs	0.064	0.019	0.052
num men at least 55 yrs	0.014	0.049	0.006
head agricultural worker	0.017	0.034	0.013
num women 20 to 34 yrs	0.006	0.024	0.030
num boys 11 to 14 yrs	0.010	0.023	0.021
num men 20 to 34 yrs	0.011	0.014	0.038
num girls 11 to 14 yrs	0.011	0.013	0.027
num girls 15 to 19 yrs	0.022	0.008	0.016
num boys 15 to 19 yrs	0.021	0.005	0.020
male head of household	0.003	0.019	0.003
num women 35 to 54 yrs	0.005	0.011	0.009
num men 35 to 54 yrs	0.013	0.005	0.006
head education	0.002	0.002	0.008
num women at least 55 yrs	0.007	0.002	0.005
N_{TE}	6784	6784	6784

Notes: Feature importances as estimated from causal forest estimation of heterogeneous treatment impacts of PROGRESA on three outcome dimensions: log consumption (log monthly per capita consumption), schooling (number of missed school days per child), and health (number of sick days per child). Schooling and health sick days / missed school days measured over 28 days prior to survey. Estimates reflect 3 separate causal forest estimations for each respective outcome.

Table S4: Alternative Outcome Specifications

	Household Poverty Score								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Log Welfare Weights $log(\gamma)$									
Indigenous	-0.174 (-0.227, -0.038)	-0.282 (-0.316, 0.05)	0.548 (0.46, 0.608)	-0.175 (-0.229, -0.038)	-0.192 (-0.253, -0.058)	-0.194 (-0.246, -0.056)	0.61 (0.566, 0.651)	0.546 (0.446, 0.604)	0.178 (0.087, 0.378)
log(Income)	-0.19 (-0.234, -0.138)	-0.192 (-0.264, -0.072)	-0.219 (-0.238, -0.166)	-0.192 (-0.234, -0.139)	-0.194 (-0.241, -0.147)	-0.195 (-0.246, -0.149)	-0.238 (-0.243, -0.219)	-0.219 (-0.239, -0.168)	-0.174 (-0.256, -0.137)
Household Size	0.104 (0.085, 0.118)	0.097 (0.069, 0.131)	0.113 (0.096, 0.122)	0.105 (0.084, 0.118)	0.106 (0.084, 0.121)	0.106 (0.087, 0.122)	0.116 (0.11, 0.12)	0.113 (0.097, 0.122)	0.116 (0.091, 0.14)
Head Age	-0.016 (-0.02, -0.01)	-0.016 (-0.022, -0.009)	-0.017 (-0.019, -0.016)	-0.016 (-0.02, -0.011)	-0.018 (-0.024, -0.012)	-0.018 (-0.025, -0.012)	-0.02 (-0.02, -0.018)	-0.018 (-0.019, -0.016)	-0.016 (-0.021, -0.01)
Education	-0.727 (-0.952, -0.505)	-0.831 (-1.277, -0.415)	-0.844 (-1.048, -0.666)	-0.731 (-0.956, -0.515)	-0.785 (-1.147, -0.578)	-0.79 (-1.038, -0.615)	-1.012 (-1.289, -0.87)	-0.84 (-1.012, -0.683)	-0.62 (-0.986, -0.322)
Impact Weights β									
Log Consumption (per capita)	6.07 (4.04, 7.28)			6.08 (4.01, 7.4)	6.68 (4.38, 8.45)	6.69 (4.33, 7.77)			
Log Food Consumption (per capita)		2.92 (-0.28, 5.13)							
Log Non-Food Consumption (per capita)		4.32 (0.39, 5.75)							
Linear Consumption (per capita)									0.02 (0.01, 0.03)
Missed Schooling (per day)	-0.48 (-1.33, 0.02)	-0.67 (-1.83, 0.72)	-0.78 (-1.46, -0.3)	-0.48 (-1.38, 0.0)				-0.78 (-1.49, -0.33)	-0.4 (-1.31, 0.09)
Sickness (per child sick day)	-0.05 (-0.51, 0.56)	-0.05 (-1.01, 1.12)	0.05 (-0.24, 0.39)		-0.06 (-0.58, 0.7)		0.06 (-0.25, 0.47)		-0.04 (-0.47, 0.43)
Value Regardless of Impact	1	1	1	1	1	1	1	1	1
N_{rank}	7767	7767	7767	7767	7767	7767	7767	7767	7767
N_{TE}	6784	6784	6784	6784	6784	6784	6784	6784	6784

Notes: All columns computed using our method, using heterogeneous treatment effects estimated with OLS (see Figure 2). 95% 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.

Table S5: Alternative Welfare Weight Specifications (1)

	Household Poverty Score					
	(1)	(2)	(3)	(4)	(5)	(6)
				Log Welfare Weights $log(\gamma)$		
Indigenous	-0.174 (-0.227, -0.038)	-0.067 (-0.137, 0.065)	-0.113 (-0.221, 0.105)	-0.104 (-0.146, 0.036)	-0.102 (-0.143, -0.002)	-0.035 (-0.17, 0.189)
log(Income)	-0.19 (-0.234, -0.138)		-0.223 (-0.28, -0.14)			
Household Size	0.104 (0.085, 0.118)				0.045 (0.034, 0.052)	
Head Age	-0.016 (-0.02, -0.01)					-0.013 (-0.023, -0.005)
Education	-0.727 (-0.952, -0.505)			-0.53 (-0.694, -0.363)		
				Impact Weights β		
Log Consumption (per capita)	6.07 (4.04, 7.28)	3.04 (1.84, 4.1)	6.25 (3.98, 7.67)	3.52 (2.32, 4.57)	2.6 (1.8, 3.34)	4.35 (2.72, 7.41)
Missed Schooling (per day)	-0.48 (-1.33, 0.02)	-0.66 (-1.36, -0.19)	-1.21 (-2.36, -0.25)	-0.42 (-1.18, -0.03)	-0.55 (-0.98, -0.24)	-0.89 (-2.41, -0.01)
Sickness (per child sick day)	-0.05 (-0.51, 0.56)	-0.0 (-0.39, 0.39)	0.24 (-0.41, 1.07)	-0.09 (-0.42, 0.44)	-0.07 (-0.31, 0.23)	0.22 (-0.4, 1.22)
Value Regardless of Impact	1	1	1	1	1	1
N_{rank}	7767	7767	7767	7767	7767	7767
N_{TE}	6784	6784	6784	6784	6784	6784

Notes: All columns computed using our method, using heterogeneous treatment effects estimated with causal forest (see Figure 2). 95% 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.

Table S6: Alternative Welfare Weight Specifications, Continued (2)

	Household Poverty Score				
	(7)	(8)	(9)	(10)	(11)
			Log Welfare Weights $log(\gamma)$		
Indigenous			-0.148 (-0.213, 0.041)	-0.155 (-0.215, -0.048)	-0.059 (-0.141, 0.078)
log(Income)	-0.219 (-0.277, -0.144)	-0.18 (-0.236, -0.127)	-0.182 (-0.238, -0.127)	-0.184 (-0.242, -0.133)	-0.112 (-0.156, -0.077)
Household Size				0.072 (0.057, 0.085)	0.018 (0.01, 0.032)
Head Age					-0.001 (-0.004, 0.002)
Education		-0.794 (-1.198, -0.512)	-0.816 (-1.286, -0.513)	-0.461 (-0.656, -0.295)	-0.468 (-0.658, -0.357)
Number of Adults					-0.093 (-0.149, -0.058)
Number of 0-5 y.o.					0.195 (0.154, 0.221)
Number of 6-16 y.o.					0.067 (0.047, 0.098)
			Impact Weights β		
Log Consumption (per capita)	5.3 (4.09, 6.24)	5.07 (4.03, 5.9)	6.26 (4.17, 7.66)	4.01 (2.91, 4.75)	2.71 (1.61, 3.93)
Missed Schooling (per day)	-1.22 (-2.27, -0.36)	-0.78 (-1.86, -0.07)	-0.71 (-1.86, 0.01)	-0.58 (-1.14, -0.13)	-0.6 (-1.04, -0.18)
Sickness (per child sick day)	0.22 (-0.41, 1.1)	0.04 (-0.47, 0.96)	0.03 (-0.52, 1.12)	-0.01 (-0.38, 0.5)	-0.56 (-0.7, -0.16)
Value Regardless of Impact	1	1	1	1	1
N_{rank}	7767	7767	7767	7767	7767
N_{TE}	6784	6784	6784	6784	6784

Notes: All columns computed using our method, using heterogeneous treatment effects estimated with causal forest (see Figure 2). 95% 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.

Table S7: Fixed-Parameter Model Estimates

	Household Poverty Score			
	Egalitarian	Only Value Consumption	Only Value Missed School Days	Only Value Sick Days
	Log Welfare Weights $log(\gamma)$			
Indigenous	0	0.293 (0.254, 0.339)	0.688 (0.571, 0.848)	0.542 (0.469, 0.627)
log(Income)	0	-0.173 (-0.19, -0.155)	-0.282 (-0.351, -0.252)	-0.222 (-0.255, -0.173)
Household Size	0	0.104 (0.097, 0.109)	0.119 (0.106, 0.141)	0.114 (0.099, 0.124)
Head Age	0	-0.014 (-0.016, -0.011)	-0.027 (-0.033, -0.023)	-0.018 (-0.021, -0.016)
Education	0	-0.727 (-0.814, -0.617)	-1.365 (-1.986, -1.142)	-0.837 (-1.042, -0.692)
	Impact Weights β			
Log Consumption (per capita)	2.69 (2.08, 3.54)	1	0	0
Missed Schooling (per day)	-0.68 (-1.36, -0.23)	0	-1	0
Sickness (per child sick day)	0.0 (-0.36, 0.4)	0	0	-1
Value Regardless of Impact	1	1	1	1
N_{rank}	7767	7767	7767	7767
N_{TE}	6784	6784	6784	6784

Notes: All columns computed using our method, using heterogeneous treatment effects estimated with OLS (see Figure 2). 95% 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. Column 1 presents results with enforced egalitarianism (equal welfare weights) across households. Columns 2-4 present results allowing priority based only on one outcome, under the assumption that it is valued equally as the value of the program independent of impacts $|\beta_j| = \alpha$.

Table S8: Further Robustness Checks

	Household Poverty Score			
	Full Sample	Only Eligible	Causal Forest TEs	Binary Ranking (Eligible/Ineligible)
	Log Welfare Weights $log(\gamma)$			
Indigenous	-0.174 (-0.227, -0.038)	-0.272 (-0.322, -0.115)	-0.189 (-0.268, 0.021)	0.101 (0.02, 0.322)
log(Income)	-0.19 (-0.234, -0.138)	-0.223 (-0.273, -0.175)	-0.192 (-0.255, -0.135)	-0.085 (-0.161, -0.041)
Household Size	0.104 (0.085, 0.118)	0.126 (0.104, 0.148)	0.097 (0.084, 0.109)	0.051 (0.042, 0.065)
Head Age	-0.016 (-0.02, -0.01)	-0.022 (-0.025, -0.017)	-0.018 (-0.023, -0.013)	-0.004 (-0.011, -0.001)
Education	-0.727 (-0.952, -0.505)	-0.722 (-0.876, -0.355)	-0.679 (-0.826, -0.568)	-0.547 (-0.663, -0.456)
	Impact Weights β			
Log Consumption (per capita)	6.07 (4.04, 7.28)	6.92 (4.28, 7.88)	9.41 (6.93, 11.4)	8.63 (6.45, 10.4)
Missed Schooling (per day)	-0.48 (-1.33, 0.02)	-0.34 (-1.71, 0.25)	-0.02 (-1.21, 0.66)	-0.49 (-1.99, 0.9)
Sickness (per child sick day)	-0.05 (-0.51, 0.56)	0.33 (-0.45, 0.8)	0.26 (-0.45, 0.5)	-0.37 (-1.54, 0.16)
Value Regardless of Impact	1	1	1	1
N_{rank}	7767	6641	7767	7767
N_{TE}	6784	6641	6784	6784

Notes: All columns computed using our method, using heterogeneous treatment effects estimated with OLS (see Figure 2) except where noted. Column 2 presents results using only households determined program-eligible for PROGRESA in both November 1999 and October 1998. Column 3 presents results using causal forest to estimate heterogeneous treatment effects. Column 4 presents results using a binary {0,1} measure of each household's eligibility for PROGRESA as the priority ranking z_i; we use causal forest HTE estimates for this model as we encounter non-convergence issues with OLS (although the OLS model with alternative outcomes specification in Table A4, column 2 is convergent and similar). 95% 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.

Table S9: Preferences for Paternalism in Survey of Mexican Residents

Prompt	Likert Average (0-4)
Some children may require direct support	
Nutrition and resources	3.6
Education	3.6
Health	3.6
Externalities	
Nutrition and resources	3.6
Education	3.7
Health	3.5
Do not directly help children; trust parents	
Nutrition and resources	1.4
Education	1.3
Health	1.4
$N_{respondents}$	429

Notes: Subjective preferences for paternalism was asked of Mexican residents in an online survey through a series of Likert scale responses to prompts, as detailed in Section S4. Each cell reports the average response (between 0 to 4), when strongly disagree is coded as zero and strongly agree as 4. Direct support assessed with 'Some children may require direct support from the government to ensure that they X', where X was 'are healthy', 'regularly attend school', or 'are fed adequately' for the three different outcomes. Likewise, externalities was assessed with prompt, 'When one child is X, it benefits other members of the community', for X 'healthier', 'better educated', or 'better fed'. Finally, trust parents was assessed with prompt, 'The government should not directly help children, but should instead trust parents to do what is best for their children's X' for X 'health', 'education', or 'nutrition'.

Table S10: Monte Carlo: Precision by Sample Size

N_{HTE}	N_{rank}	MAE_γ	MAE_β
1000	1000	0.1532	0.2125
1000	5000	0.1279	0.1964
1000	15000	0.1256	0.1952
5000	1000	0.1093	0.1585
5000	5000	0.0672	0.0973
5000	15000	0.0596	0.0867
15000	1000	0.1014	0.1463
15000	5000	0.0490	0.0696
15000	15000	0.0380	0.0556

Notes: Table reports the average mean absolute error (MAE) from Monte Carlo simulations that use our method to infer preferences from samples of different sizes. The first two columns indicate the sample size used for estimating treatment effects and the ranking, respectively.

Simulation details: We assume the true policy parameters are $\boldsymbol{\gamma} = [1.08, 0.94]$, $\boldsymbol{\beta} = [-0.22, 1.32]$, inspired by the PROGRESA application. We assume true outcomes are given by $v_{ij} = \theta_{0j} + \boldsymbol{\theta}_{\mathbf{x}j}\tilde{\mathbf{x}}_i + (\theta_{Tj} + \boldsymbol{\theta}_{T\mathbf{x}j}\tilde{\mathbf{x}}_i)T_i + e_{ij}$, which we estimate via OLS from a sample of N_{HTE} households. MAE estimates are averaged over 1000 Monte Carlo draws and across both components of the $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ vectors. Simulation parameters:

Covariates: $\tilde{x}_i \overset{\text{iid}}{\sim} N(1, 0.25)$; $\mathbf{x}_i$ contains the first 2 dimensions of $\tilde{\mathbf{x}}_i$.

Treatment status: $T_i \sim Binomial$ with $p = 0.5$.

Heterogeneous effect: $\boldsymbol{\theta}_T = [-0.14, -0.70]$, $\boldsymbol{\theta}_0 = [0, 0]$, $\boldsymbol{\theta}_{T\mathbf{x}0}$ = [-0.76, 0.08, 0.08, 0.61, 0.77, -0.04, -0.28, 0.12, -0.31, -0.12, -0.35, 0.22, -0.08, 0.28, 0.60, -0.71, 0.17, 0.37, 0.12, 0.35, -0.71, -0.25], $\boldsymbol{\theta}_{\mathbf{x}0}$ = [-0.16, -0.06, -0.40, 0.02, -0.17, 0.14, 0.70, -0.51, 1.18, -0.33, -0.10, 0.28, 0.24, -0.22, 0.03, 0.02, 0.19, 0.04, -0.02, 0.23, -0.71, 0.60], $\boldsymbol{\theta}_{T\mathbf{x}1}$ = [0.62, -0.44, 0.54, -0.78, 0.36, 0.06, 0.48, -1.05, 0.91, 0.19, -0.21, 0.70, -0.13, -0.08, 0.59, -0.25, -0.18, -0.37, -0.87, 0.14, -0.00, 0.33], $\boldsymbol{\theta}_{\mathbf{x}1}$ = [-0.39, -0.07, -0.03, -0.19, -0.70, 0.27, 0.17, 0.05, 0.80, 0.07, 0.59, -0.68, -0.33, 1.08, 0.75, -0.03, -0.93, -0.03, 0.43, 0.42, 0.30, -0.63]. $e_{ij} \sim N(0, 0.75)$ and ranking errors distributed $EV1(0, 1)$.

Table S11: Monte Carlo: Attenuation/Misspecification and SIMEX Correction

TE Est. Method	K_{True}	K_{Est}	$Bias_\gamma$	$Bias_\beta$	MAE_γ	MAE_β
Known	22	22	-0.001	0.006	0.01	0.052
OLS	22	22	0.007	-0.055	0.023	0.157
OLS + SIMEX	22	22	0.006	-0.024	0.025	0.147
Known	5	5	-0.003	0.078	0.055	0.427
OLS	5	5	0.023	-0.158	0.075	0.59
OLS + SIMEX	5	5	0.025	-0.044	0.137	1.18
Known	22	22	-0.001	0.006	0.01	0.052
OLS	22	5	0.016	-0.179	0.05	0.573
OLS + SIMEX	22	5	0.009	-0.024	0.105	1.143

Notes: Table reports the average mean absolute error (MAE) and bias from Monte Carlo simulations that use our method to infer preferences from samples of different sizes. The first three columns indicate the first-stage treatment effect estimation method; the number of covariates in the true heterogeneous effect DGP; and the number of covariates included in the estimated first-stage model. "Known" is the setting where true effects are directly used in the second-stage estimation; "OLS" is our baseline setting, where OLS estimation is used in the first-stage; and "OLS + SIMEX" is the setting where SIMEX adjustment is used on the first-stage estimates (Cook and Stefanski, 1994). For more detail on SIMEX, see Section S5.2.

Simulation details: We assume the true policy parameters are $\boldsymbol{\gamma} = [1.08, 0.94]$, $\boldsymbol{\beta} = [-0.22, 1.32]$, inspired by the PROGRESA application. We assume true outcomes are given by $v_{ij} = \theta_{0j} + \boldsymbol{\theta}_{\mathbf{x}j}\tilde{\mathbf{x}}_i + (\theta_{Tj} + \boldsymbol{\theta}_{T\mathbf{x}j}\tilde{\mathbf{x}}_i)T_i + e_{ij}$, which we estimate via OLS from a sample of N_{HTE} households. $N_{rank} = N_{HTE} = 7500$ in all settings. MAE and bias estimates are averaged over 50 Monte Carlo draws and across both components of the $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ vectors. Simulation parameters:

Treatment status: $T_i \sim Binomial$ with $p = 0.5$.

Covariates are drawn from simulated normal distributions with means and variances to match the covariates from our actual data, with simulated heterogeneous effects that are set to the estimated corresponding OLS point estimate effects for the respective covariate interaction terms in our application. $e_{ij} \sim N(0, 0.75)$ and ranking errors distributed $EV1(0, 1)$.

Table S12: PROGRESA Estimates with SIMEX Measurement Error Correction

	Household Poverty Score	
	Baseline	With SIMEX Adjustment
	Log Welfare Weights $log(\boldsymbol{\gamma})$	
Indigenous	-0.174 (-0.227, -0.038)	-0.252 (-0.381, -0.069)
log(Income)	-0.19 (-0.234, -0.138)	-0.175 (-0.229, -0.101)
Household Size	0.104 (0.085, 0.118)	0.095 (0.07, 0.109)
Head Age	-0.016 (-0.02, -0.01)	-0.014 (-0.024, -0.005)
Education	-0.727 (-0.952, -0.505)	-0.682 (-0.936, -0.399)
	Impact Weights β	
Log Consumption (per capita)	6.07 (4.04, 7.28)	7.38 (4.22, 10.24)
Missed Schooling (per day)	-0.48 (-1.33, 0.02)	-0.55 (-1.88, 0.08)
Sickness (per child sick day)	-0.05 (-0.51, 0.56)	-0.15 (-0.81, 0.77)
Value Regardless of Impact	1	1
N_{rank}	7767	7767
N_{TE}	6784	6784

Notes: All columns computed using our method, using heterogeneous treatment effects estimated with causal forest (see Figure 2). 95% 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. SIMEX adjustment procedure is applied to estimate second column figures, in order to correct for measurement error in first-stage estimates (Cook and Stefanski, 1994; Chilet, 2017). For more details on the SIMEX procedure, see Section S5.2.

Table S13: What Values are Consistent with the PROGRESA Decision Rule?
(with Causal Forest HTE Estimates)

		Household Poverty Score 1999	
		Decision Rule	**Implied Preferences**
		(Prioritization)	Welfare Weights
Welfare Weights $log(\boldsymbol{\gamma})$			
Indigenous		0.606 (0.581, 0.634)	-0.189 (-0.268, 0.021)
log(Income)		-0.237 (-0.252, -0.223)	-0.192 (-0.255, -0.135)
Household Size		0.116 (0.112, 0.119)	0.097 (0.084, 0.109)
Household Head Age		-0.02 (-0.021, -0.018)	-0.018 (-0.023, -0.013)
Education (Middle school or above)		-1.007 (-1.263, -0.85)	-0.679 (-0.826, -0.568)
Impact Weights			
Log consumption (per capita)	β_1		9.41 (6.93, 11.4)
Missed Schooling (per day)	β_2		-0.02 (-1.21, 0.66)
Sickness (per child sick day)	β_3		0.26 (-0.45, 0.5)
Value Regardless of Impact	α		1
N_{rank}		7767	7767
N_{TE}		.	6784
Hypothesis Tests			p-value
Egalitarian	$\boldsymbol{\gamma} \equiv 1$		1.01e-56
Not Paternalistic	$\boldsymbol{\beta} \equiv 0$		8.37e-11
Egalitarian and Not Paternalistic	$\boldsymbol{\gamma} \equiv 1, \boldsymbol{\beta} \equiv 0$		2.15e-90

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 causal forests to estimate heterogeneous treatment effects (see also Figure S3). 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.

Table S14: Assessing Decision Rules
(with Causal Forest HTE Estimates)

		(1)	(2)	(3)
		Implied Preferences (Estimated)		Stated Preferences
		1999 Pov. Score	2003 Pov. Score	(Resident survey)
Welfare Weights $log(\boldsymbol{\gamma})$				
Indigenous		-0.189 (-0.268, 0.021)	0.055 (0.017, 0.188)	0.065 (0.057, 0.072)
log(Income)		-0.192 (-0.255, -0.135)	-0.07 (-0.113, -0.039)	-0.071 (-0.257, 0.116)
Household Size		0.097 (0.084, 0.109)	0.085 (0.078, 0.094)	0.015 (-0.018, 0.049)
Household Head Age		-0.018 (-0.023, -0.013)	-0.003 (-0.006, -0.001)	0.004 (0.002, 0.005)
Educated		-0.679 (-0.826, -0.568)	-0.457 (-0.556, -0.386)	-0.065 (-0.099, -0.03)
Impact Weights				
Log Consumption (per capita)	β_1	9.41 (6.93, 11.4)	3.16 (2.33, 3.57)	4.37 (2.99, 5.75)†
Missed Schooling (per day)	β_2	-0.02 (-1.21, 0.66)	0.18 (-0.34, 0.39)	-1.11 (-1.6, -0.62)†
Sickness (per child sick day)	β_3	0.26 (-0.45, 0.5)	0.42 (0.13, 0.48)	-0.69 (-1.03, -0.35)†
Value Regardless of Impact	α	1	1	.
N_{rank}		7767	7767	.
N_{TE}		6784	6784	.
$N_{respondents}$		.	.	421*

Notes: Columns 1-2 are estimated using our method, using causal forests 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 421), 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.

Table S15: Impact Weights in Survey of Mexican Residents

Outcome	Private Value median $\tilde{b}_{ij}$ (unscaled)	Social Value median b_{ij} (unscaled)	Resulting Parameter median β_j (unscaled)	$N_{respondents}$
Log Consumption (transfer pesos per unit)	735.43 (29.64)	994.23 (65.38)	245.65 (39.60)	411
Education (transfer pesos per missed day of school)	-316.67 (25.39)	-425.00 (24.39)	-62.50 (14.02)	412
Health (transfer pesos per day of sickness)	-375.00 (27.26)	-450.00 (21.76)	-38.91 (9.81)	424

Notes: Preferences derived from survey of Mexican residents, as described in Appendix S4. Bootstrapped standard errors in parentheses.

Table S16: Designing Decision Rules
(with Causal Forest HTE Estimates)

	(1) HH Poverty Score	(2) Resident Preferences	(3) Equal Welfare Weights	(4) Policy only values impact on: Consumption	(5) Education	(6) 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	1.699	1.566	1.871	-0.367	3.144
log(Income)	-0.237	-0.291	-0.176	-0.343	1.003	-0.232
Household Size	0.116	0.047	0.034	0.137	-0.177	0.123
Household Head Age	-0.02	-0.002	-0.008	-0.024	-0.12	-0.13
Education	-1.007	-0.086	0.055	-0.846	-0.088	-0.683
Panel C: Counterfactual outcomes (monthly)						
Log Consumption per capita (pesos)	4.802	4.812	4.813	4.812	4.796	4.796
Missed school (days/child)	0.172	0.168	0.170	0.170	0.158	0.175
Sickness (sick days/child)	0.640	0.629	0.638	0.637	0.646	0.610
Model Log Likelihood	-60945	-61944	-62286	-61402	-61514	-61534
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.

Table S17: Alternative Outcome Specifications (with Causal Forest HTE Estimates)

	Household Poverty Score								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Log Welfare Weights $log(\gamma)$									
Indigenous	-0.189 (-0.268, 0.021)	-0.193 (-0.264, -0.018)	0.588 (0.511, 0.611)	-0.179 (-0.262, 0.037)	-0.189 (-0.282, 0.022)	-0.18 (-0.267, 0.034)	0.602 (0.553, 0.626)	0.591 (0.52, 0.618)	0.224 (0.198, 0.282)
log(Income)	-0.192 (-0.255, -0.135)	-0.198 (-0.276, -0.144)	-0.241 (-0.251, -0.208)	-0.189 (-0.275, -0.136)	-0.192 (-0.276, -0.135)	-0.189 (-0.279, -0.137)	-0.237 (-0.253, -0.219)	-0.24 (-0.255, -0.223)	-0.18 (-0.235, -0.137)
Household Size	0.097 (0.084, 0.109)	0.095 (0.082, 0.105)	0.115 (0.107, 0.119)	0.095 (0.081, 0.109)	0.097 (0.082, 0.107)	0.095 (0.081, 0.107)	0.116 (0.11, 0.121)	0.115 (0.112, 0.118)	0.105 (0.092, 0.114)
Head Age	-0.018 (-0.023, -0.013)	-0.018 (-0.023, -0.013)	-0.018 (-0.02, -0.015)	-0.018 (-0.022, -0.013)	-0.018 (-0.023, -0.014)	-0.018 (-0.023, -0.013)	-0.02 (-0.02, -0.018)	-0.018 (-0.02, -0.015)	-0.018 (-0.021, -0.012)
Education	-0.679 (-0.826, -0.568)	-0.691 (-0.784, -0.565)	-0.978 (-1.157, -0.829)	-0.661 (-0.809, -0.545)	-0.68 (-0.802, -0.573)	-0.662 (-0.796, -0.568)	-1.002 (-1.222, -0.853)	-0.984 (-1.18, -0.844)	-0.708 (-0.827, -0.612)
Impact Weights β									
Log Consumption (per capita)	9.41 (6.93, 11.4)			9.14 (6.74, 11.03)	9.42 (6.85, 12.14)	9.18 (6.83, 12.14)			
Log Food Consumption (per capita)		4.62 (2.35, 6.77)							
Log Non-Food Consumption (per capita)		5.77 (3.34, 6.94)							
Linear Consumption (per capita)									0.04 (0.03, 0.05)
Missed Schooling (per day)	-0.02 (-1.21, 0.66)	-0.35 (-1.32, 0.23)	-0.53 (-1.68, 0.47)	-0.06 (-1.23, 0.56)				-0.53 (-1.67, 0.47)	0.18 (-0.9, 0.81)
Sickness (per child sick day)	0.26 (-0.45, 0.5)	0.25 (-0.46, 0.47)	-0.05 (-0.53, 0.54)		0.26 (-0.41, 0.48)		-0.06 (-0.55, 0.54)		0.09 (-0.41, 0.38)
Value Regardless of Impact	1	1	1	1	1	1	1	1	1
N_{rank}	7767	7767	7767	7767	7767	7767	7767	7767	7767
N_{TE}	6784	6784	6784	6784	6784	6784	6784	6784	6784

Notes: All columns computed using our method, using heterogeneous treatment effects estimated with causal forest (see Figure S3). Confidence intervals are computed using a two-step Bayesian bootstrap procedure that accounts for uncertainty in both treatment effects and preference parameters. 95% 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.

Table S18: Alternative Welfare Weight Specifications (1)
(with Causal Forest HTE Estimates)

	Household Poverty Score					
	(1)	(2)	(3)	(4)	(5)	(6)
				Log Welfare Weights $log(\gamma)$		
Indigenous	-0.189 (-0.268, 0.021)	-0.11 (-0.153, 0.069)	-0.096 (-0.173, 0.128)	-0.127 (-0.17, 0.059)	-0.154 (-0.195, 0.011)	-0.081 (-0.127, 0.133)
log(Income)	-0.192 (-0.255, -0.135)		-0.189 (-0.269, -0.122)			
Household Size	0.097 (0.084, 0.109)				0.041 (0.035, 0.047)	
Head Age	-0.018 (-0.023, -0.013)					-0.008 (-0.014, -0.004)
Education	-0.679 (-0.826, -0.568)			-0.468 (-0.536, -0.4)		
				Impact Weights β		
Log Consumption (per capita)	9.41 (6.91, 11.48)	4.74 (3.36, 5.07)	7.52 (5.28, 9.52)	4.32 (3.25, 4.5)	5.69 (4.08, 6.76)	5.13 (3.76, 5.54)
Missed Schooling (per day)	0.26 (-0.44, 0.47)	0.16 (-0.25, 0.28)	0.39 (-0.26, 0.7)	0.03 (-0.28, 0.17)	0.3 (-0.2, 0.57)	0.14 (-0.29, 0.3)
Sickness (per child sick day)	-0.02 (-1.23, 0.65)	0.02 (-0.71, 0.41)	-0.22 (-1.42, 0.65)	-0.18 (-0.79, 0.22)	0.3 (-0.72, 0.65)	-0.08 (-0.81, 0.35)
Value Regardless of Impact	1	1	1	1	1	1
N_{rank}	7767	7767	7767	7767	7767	7767
N_{TE}	6784	6784	6784	6784	6784	6784

Notes: All columns computed using our method, using heterogeneous treatment effects estimated with causal forest (see Figure S3). 95% 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.

Table S19: Alternative Welfare Weight Specifications, Continued (2)
(with Causal Forest HTE Estimates)

	Household Poverty Score				
	(7)	(8)	(9)	(10)	(11)
			Log Welfare Weights $log(\gamma)$		
Indigenous			-0.118 (-0.187, 0.098)	-0.152 (-0.228, 0.029)	-0.033 (-0.064, 0.103)
log(Income)	-0.199 (-0.268, -0.13)	-0.159 (-0.22, -0.099)	-0.146 (-0.225, -0.093)	-0.141 (-0.212, -0.099)	-0.122 (-0.16, -0.09)
Household Size				0.057 (0.05, 0.065)	0.029 (0.025, 0.04)
Head Age					-0.003 (-0.004, -0.001)
Education		-0.664 (-0.823, -0.539)	-0.638 (-0.809, -0.515)	-0.397 (-0.471, -0.345)	-0.518 (-0.586, -0.463)
Number of Adults					-0.118 (-0.146, -0.102)
Number of 0-5 y.o.					0.133 (0.115, 0.142)
Number of 6-16 y.o.					0.097 (0.09, 0.105)
			Impact Weights β		
Log Consumption (per capita)	6.7 (5.41, 8.05)	6.29 (5.22, 7.49)	7.27 (5.2, 8.83)	5.44 (4.07, 6.18)	3.24 (2.11, 3.57)
Missed Schooling (per day)	-0.38 (-1.41, 0.66)	-0.46 (-1.39, 0.52)	-0.26 (-1.41, 0.53)	-0.46 (-1.3, 0.23)	-0.37 (-1.04, 0.24)
Sickness (per child sick day)	0.36 (-0.29, 0.69)	0.26 (-0.25, 0.58)	0.29 (-0.28, 0.57)	0.1 (-0.29, 0.29)	-0.3 (-0.53, -0.07)
Value Regardless of Impact	1	1	1	1	1
N_{rank}	7767	7767	7767	7767	7767
N_{TE}	6784	6784	6784	6784	6784

Notes: All columns computed using our method, using heterogeneous treatment effects estimated with causal forest (see Figure S3). 95% 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.

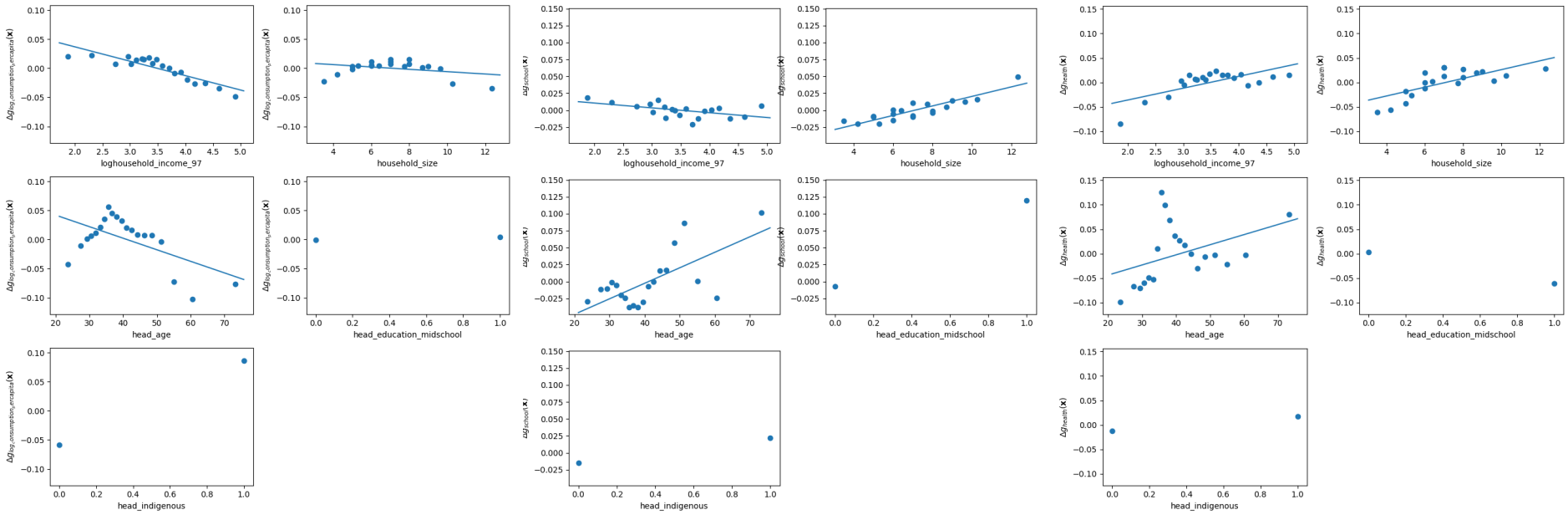
Table S20: Fixed-Parameter Model Estimates
(with Causal Forest HTE Estimates)

	Household Poverty Score			
	Egalitarian	Only Value Consumption	Only Value Missed School Days	Only Value Sick Days
	Log Welfare Weights $log(\gamma)$			
Indigenous	0	0.368 (0.344, 0.417)	0.582 (0.542, 0.652)	0.599 (0.545, 0.677)
log(Income)	0	-0.185 (-0.202, -0.172)	-0.25 (-0.275, -0.221)	-0.282 (-0.346, -0.244)
Household Size	0	0.103 (0.099, 0.106)	0.115 (0.112, 0.12)	0.12 (0.113, 0.13)
Head Age	0	-0.016 (-0.017, -0.014)	-0.017 (-0.02, -0.016)	-0.022 (-0.025, -0.018)
Education	0	-0.776 (-0.828, -0.707)	-0.987 (-1.14, -0.854)	-1.089 (-1.3, -0.953)
	Impact Weights β			
Log Consumption (per capita)	3.86 (3.08, 4.75)	1	0	0
Missed Schooling (per day)	-0.13 (-0.82, 0.36)	0	-1	0
Sickness (per child sick day)	0.13 (-0.25, 0.29)	0	0	-1
Value Regardless of Impact	1	1	1	1
N_{rank}	7767	7767	7767	7767
N_{TE}	6784	6784	6784	6784

Notes: All columns computed using our method, using heterogeneous treatment effects estimated with causal forest (see Figure S3). 95% 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. Column 1 presents results with enforced egalitarianism (equal welfare weights) across households. Columns 2-4 present results allowing priority based only on one outcome, under the assumption that it is valued equally as the value of the program independent of impacts $|\beta_j| = \alpha$.

S7 Additional Figures

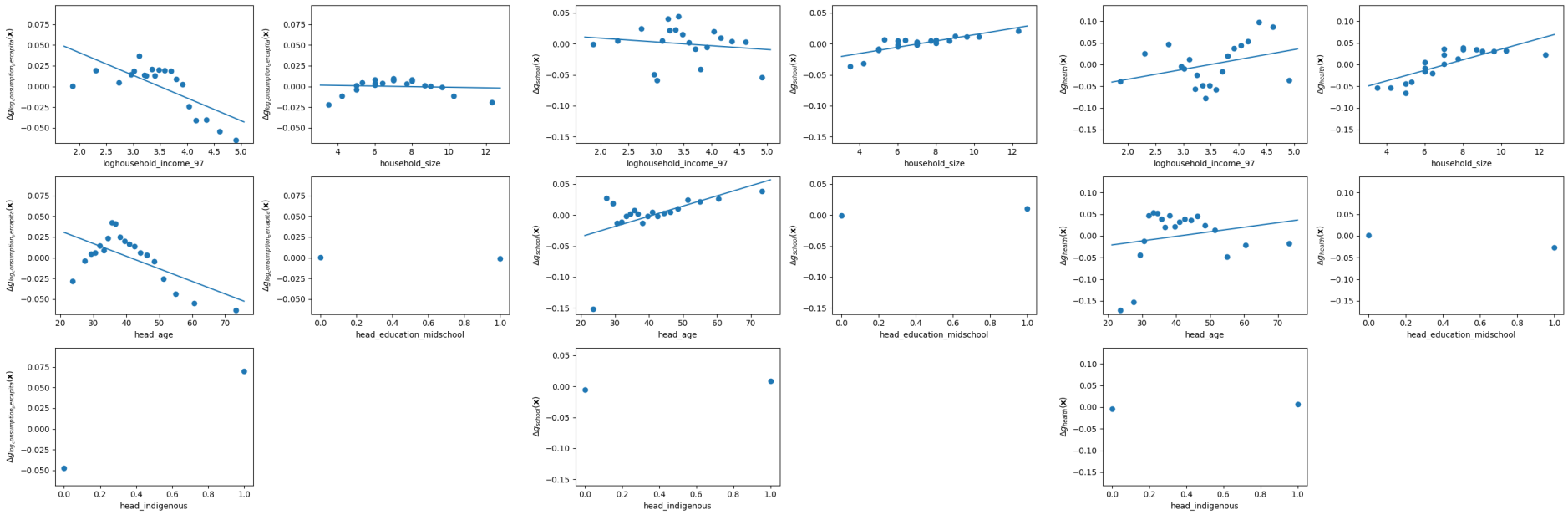
Figure S1: Binscatter Plots of Treatment Effect Heterogeneity: OLS



(a) Log Consumption Treatment Effects (b) Schooling Treatment Effects (c) Health Treatment Effects

Notes: Binscatter plots of treatment effects from OLS over five covariates: household size; household head education; household head indigenous status; household head age; and log household income in the pre-period of 1997. Figures shown for treatment effects over per-person monthly consumption, number of sick days per child, and number of missed school days per child. Treatment effects shown are residualized against remaining covariates in the regression (the other graphed covariates).

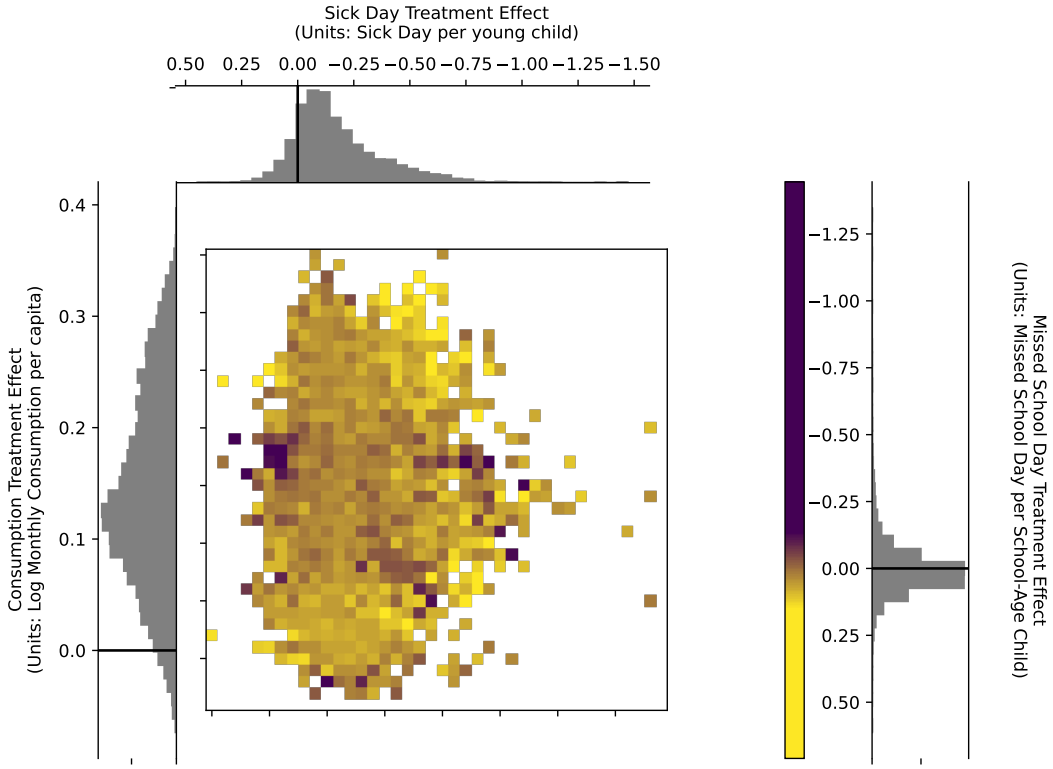
Figure S2: Binscatter Plots of Treatment Effect Heterogeneity: Causal Forest



(a) Log Consumption Treatment Effects (b) Schooling Treatment Effects (c) Health Treatment Effects

Notes: Binscatter plots of treatment effects from causal forest over a selected group of five covariates: household size; household head education; household head indigenous status; household head age; and log household income in the pre-period of 1997. Figures shown for treatment effects over per-person monthly consumption, number of sick days per child, and number of missed school days per child. Treatment effects shown are residualized against remaining covariates in the regression (the other graphed covariates).

Figure S3: Distribution of Estimated Treatment Effects (Casual Forest)

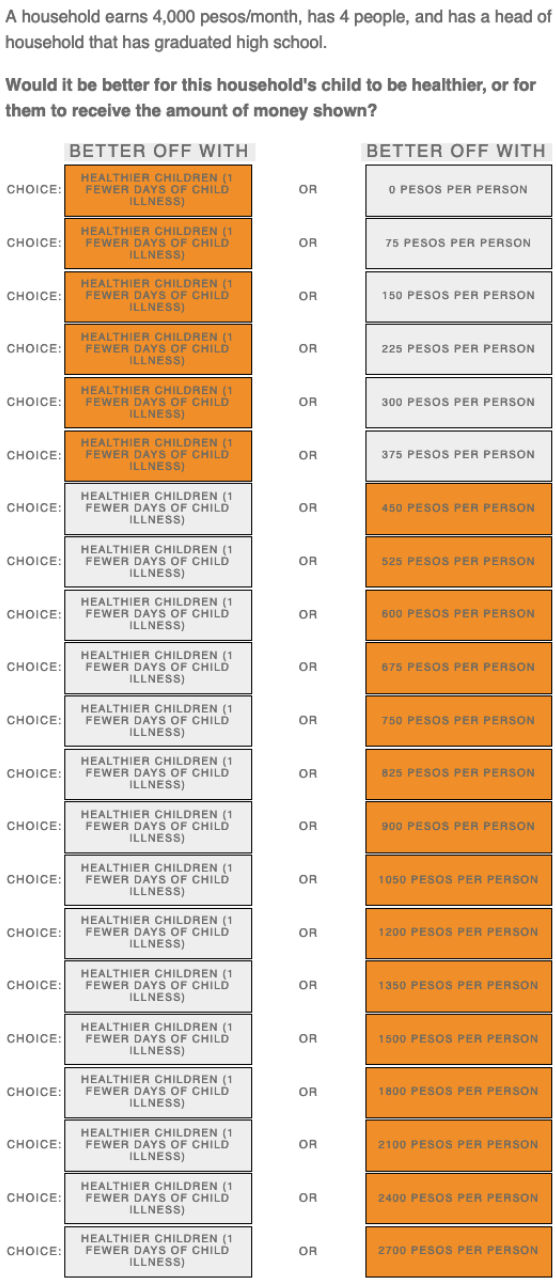


Notes: Joint and marginal distributions of estimated treatment effects of PROGRESA conditional cash transfer on schooling, health, and consumption, estimated using causal forest (Wager and Athey, 2018). Schooling treatment effects are measured over the number of missed school days per school-age child in a given household. Health treatment effects are measured over the number of sick days per young (0-5 years old) child in a given household. Consumption treatment effects are measured over per-person consumption in pesos in a given household. Marginal distributions for consumption and health treatment effects are shown over the y and x axes, respectively, and are binned together in the center figure. Average schooling treatment effects in each consumption-health-treatment-effect bin is shown by the fill color of the bin, according to the index of the legend on the right. The marginal distribution of schooling treatment effects is shown in parallel to this legend. Note that missed school days and sick days are inferred to be "bads", according to our estimated weights, and so higher negative values for these treatment effects are associated with higher social utility. Note also that we drop households without children in the relevant age range for health and schooling treatment effects; the above graphs show only TEs for households for which these TEs are defined.

Figure S4: Welfare Weight Survey Question Example

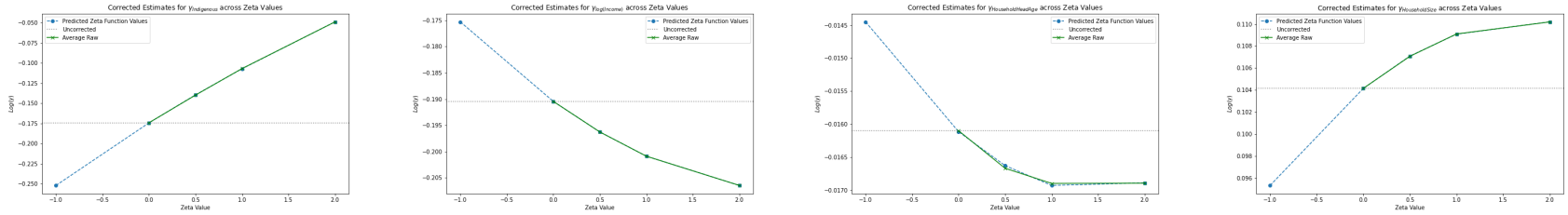
Notes: Respondents saw a version of this question translated into Spanish.

Figure S5: Impact Weight Survey Question Example



Notes: Respondents saw a version of this question translated into Spanish.

Figure S6: SIMEX Adjustment Illustration: Application to PROGRESA Estimates



(a) Indigenous (b) Log(Income) (c) Household Head Age (d) Household Size

(e) Education (f) Log(Consumption) (g) Health (h) School

Notes: Plots of SIMEX extrapolation function over different ζ values. Solid green line represents averages of estimated parameters at given ζ, dotted blue line represents projected $f_\theta(\zeta)$ function values. Dotted horizontal line represents baseline estimates using OLS treatment effects. Y axis scaled to $log(\gamma)$ for welfare weights and β for impact weights. See Section S5.2 for more details on SIMEX estimation procedure.