(Machine) Learning what Policymakers Value[*]

Daniel Björkegren[†] Joshua E. Blumenstock[‡] Samsun Knight[§]

Brown University U.C. Berkeley Brown University

September 4, 2021

Abstract

This paper develops a method to uncover the values consistent with observed allocation decisions. We use machine learning estimators for heterogeneous treatment effects to identify who benefits from an allocation. We then decompose the objective underlying the allocation into: differential (i) treatment effects, (ii) welfare weights between entities; and (iii) impact weights across outcomes. We apply this approach to Mexico's PROGRESA anti-poverty program and estimate the preferences consistent with its design. We find evidence of heterogeneous impacts by income and age; accounting for this heterogeneity, allocations imply higher welfare weights on the indigenous, poor, and for families with more children. The implied value of each missed school day and child sick day is estimated imprecisely but does not rule out conventional valuations or preferences reported by Mexican residents. Alternate eligibility criteria could have improved either average consumption, health or schooling outcomes.

[*] We thank Luk Yean and Jolie Wei for excellent research assistance. Thank you to Joseph Cummins, Brian Dillon, John Friedman, Yang Xie, and seminar audiences for helpful conversations.

[†] dan@bjorkegren.com

[‡] jblumenstock@berkeley.edu

[§] samsun_knight@brown.edu

1 Introduction

The values behind policy decisions are not always transparent. When governments decide which households receive welfare benefits, or universities select students to admit, they do not always articulate the rationale behind the criteria used. And even when policymakers do articulate rationales for allocation criteria, these claims can be difficult for constituents to verify. In particular, allocation policies may prioritize certain people either because they benefit most, or because they are favored, separately from how much they would benefit. This distinction has deep implications for understanding and designing optimal policies (Nichols and Zeckhauser, 1982; Coate and Morris, 1995). In particular, 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 the preferences that are consistent with observed policies. This method relies on recent developments in machine learning, which make it possible to estimate differential treatment effects without overfitting. We show how such methods can be followed by a second stage regression to separate heterogeneous treatment effects (who benefits the most) from implied welfare weights (who is valued) and weights on different outcomes (how different outcomes are valued). As a result, we can shift the debate from one about means — who should be eligible — to one about ends — what are the impacts we desire, and which populations are most important?

We consider a common form of decision, an allocation based on a score or ranking. These may be poverty scores in the case of welfare programs, or explicit rankings in the case of applicants for small business grants or college admission. This ranking implies a system of inequalities between the contributions of different entities to welfare. We use this system of inequalities, and a simple and general model of welfare, to estimate the implied value on different welfare outcomes (estimated using modern methods for heterogenous effects) and different entities (based on observed characteristics), using ordinal logit. Our method can also be used if an observer had only binary information on eligibility, though in that setting it will have less power.

Intuitively, if a policy allocates to one type of applicant who benefits little from the allocation over a different type who would benefit 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 from the allocation — instead of applicants whose consumption increases — that implies the policy implicitly highly values health.

We apply the method to the case of PROGRESA, one of the world's largest (and best-studied) anti-poverty programs, which provided cash transfers to eligible households in Mexico.[1] We first estimate the heterogenous treatment effects of the program, exploiting randomized variation in the program rollout. These estimates can be computed using any method for predicting heterogeneous counterfactual outcomes; we demonstrate with the causal forests machine learning method (Wager and Athey, 2018), which has obtained substantial attention. Results indicate treatment effect heterogeneity, such that younger and wealthier households benefit most from the program. This is similar to prior work based on linear models (Djebbari and Smith, 2008).

Next, we combine these estimated treatment effects with ordinal logit to identify the preferences consistent with the ranking between households. We assume that preferences are over school attendance, child health, and consumption, and include additive welfare weights based on a small number of characteristics. Our results suggest that the program's design is consistent with assigning extra value to indigenous households, poor households, and households with children. The weights imply that households are on average ranked 13.3 percentiles higher if indigenous, 7.7 percentiles lower for each standard deviation increase in household income, 21.1 percentiles higher for each additional small child (ages 5 and lower) in the household, 15.1 percentiles higher for each additional child aged 6-16, and 20.3 percentiles lower for additional adult.[2] Our estimates of weights on different impacts are imprecise, but 95% confidence intervals suggest that the Mexican government's initial allocation rule implies a value of each school day attended of less than 685 pesos of consumption, and each prevented sick day among young children less than 690 pesos.

We then compare these implied preferences with the preferences of Mexican residents, as measured by hypothetical allocation questions in a survey. These preferences are imprecisely measured, but we find similar welfare weights on income, and confidence intervals overlap for the valuation of different outcomes.

Finally, we evaluate the counterfactual allocations that would have occurred had the policymaker placed higher value on certain types of impacts (e.g., health vs. education) or certain types of households (e.g., lower-income or indigenous). If the government assigned welfare weight solely based on preferring lower income households, the rule would slightly de-prioritize indigenous households and households with small children. A technocratic ranking that weighs impacts according to external cost benefit estimates would have traded

[1] PROGRESA conditioned payouts on certain actions, but we treat the program as an unconditional transfer. For simplicity, we assume PROGRESA did not have differential spillover benefits on different households.

[2] We compute the average percentile change by first computing how each household's projected ranking would shift, given different covariates, and then taking the median change over all households.

off increased sick days to reduce missed school days and increase consumption. In practice, implemented policies may balance the needs of multiple advocates. We can use our method to assess those contributions. An education minister who valued only schooling impacts would advocate for increasing the priority on households of indigenous status. A minister valuing only health impacts would advocate for increasing the priority of smaller households. A minister valuing only consumption impacts would advocate to slightly increase the priority of indigenous households and households with lower income. Finally, we assess the impact of all of these alternative allocations on consumption, sick days, and missed days of school.

This approach makes it possible to invert the discussion about allocative 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? By how much should we value poor families versus middle class families?). The framework can be applied in numerous settings where policymakers allocate scarce resources and heterogeneous treatment effects can be estimated. Our method does not require access to all of the information that policymakers use to develop a ranking; it can be computed with the final ranking and welfare relevant characteristics of individuals.

The approach has three caveats. First, it requires defining the form of values precisely: the implied weights may depend on what outcomes and characteristics are allowed to enter the objective function. Outcomes that are unmeasured are assumed to not be valued. This definition of what may be valued is a substantive decision, and other definitions of welfare could be substituted in to our method. Second, in order to estimate the counterfactual outcomes among the entire distribution of individuals, it requires variation in treatment among both individuals who are ultimately eligible and ineligible for the program (that is, the eligibility requirements for the experiment must differ from those of the fully-implemented program). This is commonly the case with randomized controlled trials. Third, it requires a large enough dataset to both measure heterogeneous treatment effects, and the implied welfare parameters. These datasets are increasingly becoming available, particularly in settings with digital experimentation.

Related Literature

This paper contributes to literature on optimal targeting and taxation (Nichols and Zeckhauser, 1982; Barr, 2012; Fleurbaey and Maniquet, 2018), and especially work focused on targeting in developing countries versus 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. It relates to work that infers policymaker preferences from their actions (Timmins, 2003). Our empirical results also engage with research on the effects and allocation of cash transfer programs (Behrman and Todd, 1999; Gertler, 2004; John Hoddinott, 2004; Djebbari and Smith, 2008), particularly PROGRESA (Skoufias et al., 2001; Coady, 2006). We build on this work by showing how effects and allocations can be combined to audit policymaker priorities, and improve the design of future policies.

A growing literature 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 public finance literature on welfare weights. Hendren (2019) infers 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. Our paper shows how similar welfare questions can be raised across a broad set of domains where heterogeneous treatment effects can be estimated.

Our efforts relate broadly to recent work on fairness in machine learning (Dwork et al., 2012; Barocas et al., 2018). Within this subfield, several papers have studied 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 relatedly 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 a policymaker choosing how to allocate treatment among N entities, which could be, for example, individuals, households, firms, or regions. For convenience, we refer to entities as households. The policymaker determines a ranking z_i for each household i, which will ultimately be used to select a treatment status $T_i \in \{0, 1\}$. Household i has characteristics $\mathbf{x}_i$.

We assume that the ranking results from some implicit welfare function:

$$S = \sum_i S_i$$

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

where $\mu(\mathbf{x}_i)$ represents the welfare weight of a household with characteristics $\mathbf{x}_i$, with a functional form to be specified later.

u_i is the policymaker's implied valuation of household i's utility, which may be a linear combination of multiple components v_{ij} (such as consumption utility, health, and education):

$$u_i(T_i) = v_{i0}(T_i) + \sum_{j>0} \lambda_j(\mathbf{x}_i) v_{ij}(T_i) + C \cdot T_i$$

where $\lambda_j(\mathbf{x}_i)$ represents the relative value, or 'impact weight' of j relative to the numeraire or reference outcome ($j = 0$), with a functional form to be specified later. C is a constant representing the net intrinsic value of providing the program, even absent impact.[3]

Assume we can compute the impact of treatment on household i's component of utility j: $\Delta v_{ij} \coloneqq v_{ij}(1) - v_{ij}(0)$. The predicted welfare impact of treating household i is then:

$$\Delta S_i = \mu(\mathbf{x}_i) \cdot \left(\Delta v_{i0} + \sum_{j>0} \lambda_j(\mathbf{x}_i) \Delta v_{ij} + C \right)$$

The policymaker ranks each household by its contribution to welfare:

$$z_i = f(\Delta S_i + \epsilon_i) \tag{1}$$

where f is a weakly increasing transformation, which preserves the priority order of who receives treatment but not the intensity of preferences. ϵ_i is measurement error that is iid and mean zero, which could represent measurement error in estimates of welfare, or mistakes by the policymaker.

2.1 Measuring Utility Impacts

We assume that each utility component v_{ij} is a function of underlying outcome y_{ij}:

$$\Delta v_{ij} \coloneqq v_{ij}(T_i = 1) - v_{ij}(T_i = 0) = g_j(\hat{y}_{ij}^1) - g_j(\hat{y}_{ij}^0)$$

where g_j represents the utility function for j (which could be for example, $g_j(y) = \log(y)$, or $g_j(y) = y$).

We take as given that we have an experimental design that has recovered the predicted effect of treatment on each outcome $\Delta y_j(\mathbf{x}_i)$, which may be heterogeneous as a function of covariates $\mathbf{x}_i$. We can use this to predict the outcome for both the factual and counterfactual state.

If i is control ($T_i = 0$):

$$\begin{aligned} \hat{y}_{ij}^0 &= y_{ij} \\ \hat{y}_{ij}^1 &= y_{ij} + \Delta y_j(\mathbf{x}_i) \end{aligned}$$

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

or treated ($T_i = 1$):

$$\begin{aligned}\hat{y}^0_{ij} &= y_{ij} - \Delta y_j(\mathbf{x}_i) \\ \hat{y}^1_{ij} &= y_{ij}\end{aligned}$$

If the $g_j(\cdot)$ utility functions are incorrectly specified to be linear, then welfare weights $\mu(\mathbf{x}_i)$ and impact weights $\boldsymbol{\lambda}(\mathbf{x}_i)$ will measure the combination of the underlying welfare weights and curvature in utility to first approximation, as long as the baseline level is included as a characteristic $\mathbf{x}_i$. See Appendix Section A.1.

(Here we have assumed that treatment effects are known with certainty; in the estimation section we will consider adjustments for uncertainty.)

3 Intuition

To demonstrate the intuition behind our method, we consider a simple example in Figure 1. Consider the case of a single outcome and one dimension of heterogeneity, x, which corresponds with consumption. A policymaker allocates a program by ordering households by the function $z(x)$, prioritizing poor households. As shown in Figure 1, the same allocation rule could result from higher welfare weights on the poor, equal welfare weights, or higher welfare weights on the rich, depending on how treatment effects vary with x.

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

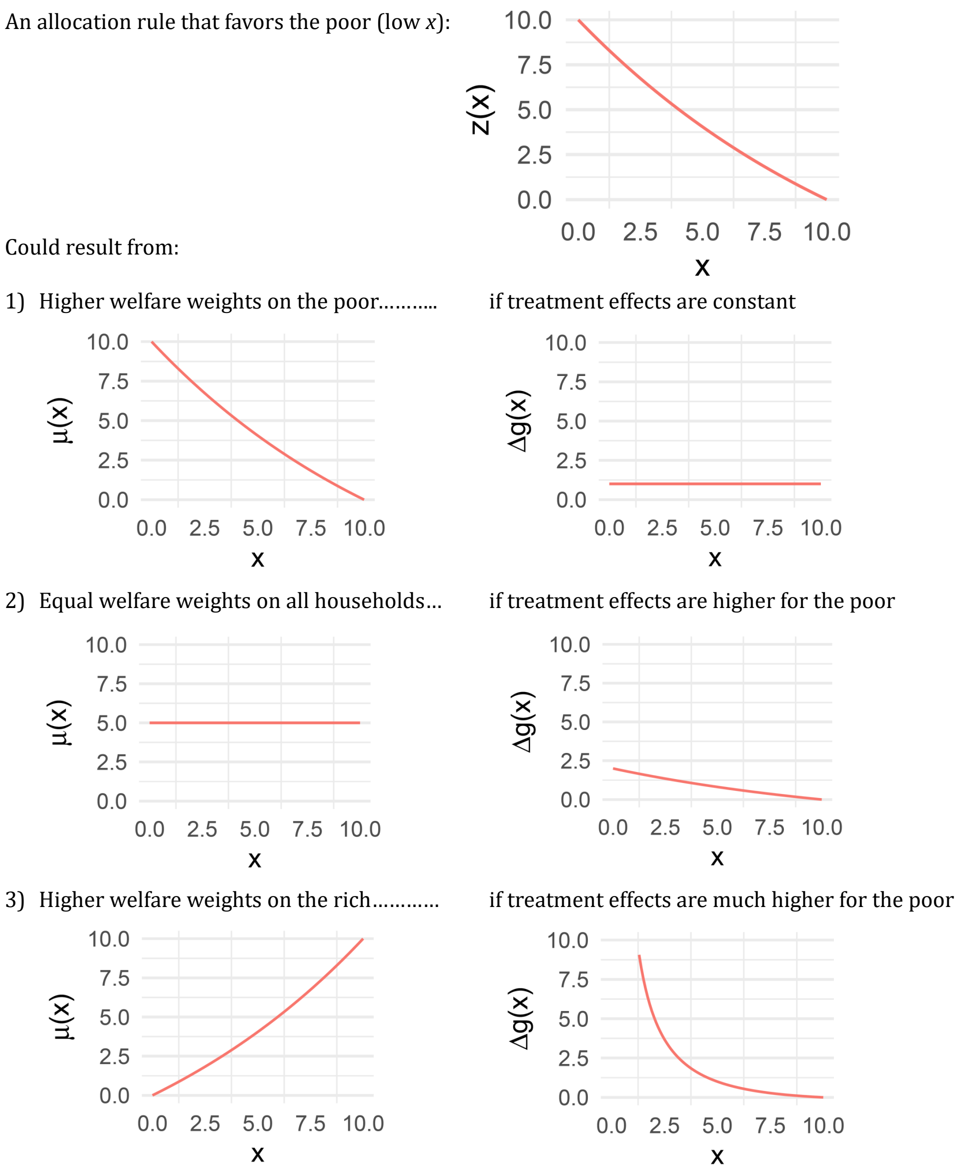


Figure 1: Intuitive Example

4 Estimation

What preferences ($\mu(\mathbf{x}_i)$, $\boldsymbol{\lambda}(\mathbf{x}_i)$, C) are consistent with a policy that allocates treatment ($\boldsymbol{z}$) and achieves effects (estimated to be $\Delta\hat{v_{ij}}$)? If the policymaker prioritizes household i over i' ($z_i > z_{i'}$), Equation 1 suggests we must have:

$$\mu(\mathbf{x}_i) \cdot \left(\Delta\hat{v}_{i0} + \textstyle\sum_{j>0} \lambda_j(\mathbf{x}_i)\Delta\hat{v}_{ij} + C\right) + \epsilon_i > \mu(\mathbf{x}_{i'}) \cdot \left(\Delta\hat{v}_{i0} + \textstyle\sum_{j>0} \lambda_j(\mathbf{x}_{i'})\Delta\hat{v}_{ij} + C\right) + \epsilon_{i'}$$

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

The 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: $\epsilon_i \sim \sigma \cdot EV(1)$. Then the logit likelihood of this particular placement of i in the ranking $\boldsymbol{z}$ is:

$$l_i = \frac{\exp\left[\frac{1}{\sigma} \cdot \mu(\mathbf{x}_i)\left(\Delta\hat{v}_{i0} + \sum_{j>0} \lambda_j(\mathbf{x}_i)\Delta\hat{v}_{ij} + C\right)\right]}{\sum_{i' \epsilon \Lambda_i} \exp\left[\frac{1}{\sigma} \cdot \mu(\mathbf{x}_{i'})\left(\Delta\hat{v}_{i'0} + \sum_{j>0} \lambda_j(\mathbf{x}_{i'})\Delta\hat{v}_{i'j} + C\right)\right]}$$

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

$$L(z, \mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}, C, \sigma) = \prod_i l_i$$

We use maximum likelihood to estimate the functions $\mu(\mathbf{x}_i)$ and $\boldsymbol{\lambda}(\mathbf{x}_i)$ and parameters C and σ that best match the observed data $\{\boldsymbol{z}, \mathbf{x}, \{\Delta\hat{v}_{ij}\}_{ij}\}$.

Unlike standard discrete choice that observes partial orderings for multiple decisionmakers, we observe one ordering of all alternatives. For this type of ranked data, we follow the exploded logit likelihood described by Train (2009).[4]

For components of utility with $g_j(\cdot)$ that is linear in its underlying outcome, we can simply compute $\Delta\hat{v}_{ij} = g_j(\hat{y}^1_{ij}) - g_j(\hat{y}^0_{ij})$ as shown in Section 2.1. However, this expression would be a biased estimate of the change in that utility component if there is curvature in $g_j(\cdot)$ and uncertainty in $\hat{y}^T_{ij}$, due to Jensen's inequality. For utility components j with curvature, we compute $\Delta\hat{v}_{ij} = \frac{1}{D}\sum_{d=1}^{D}\left[g_j(\hat{y}^1_{ijd}) - g_j(\hat{y}^0_{ijd})\right]$ for D bootstrapped estimates of treatment effects to obtain an unbiased estimate.

Standard errors are computed using a two-step bootstrap procedure that accounts for uncertainty in both treatment effects and preference parameters. Observations are drawn

[4] The likelihood considers the selection of the highest ranked household against all alternative households. Then, the highest ranked household is dropped from the comparison, and the likelihood considers selection of the second highest ranked household against all lower ranked households. It continues in this manner through to the lowest ranked household.

with replacement, and these bootstrapped samples are used to compute treatment effects, and then welfare and impact weights. Standard errors reported are the standard deviation across bootstrapped welfare and impact weight estimates.

4.1 Parameterization

Our framework will work with general functional forms for $\mu(\mathbf{x}_i)$ and $\lambda_j(\mathbf{x}_i)$. For our empirical application, we model welfare weights as additive so they can be easily interpreted:

$$\mu(\mathbf{x}_i) = 1 + \boldsymbol{\omega}\cdot\mathbf{x}_i$$

We model the relative weight on outcome j as the same for all households, due to power constraints: $\lambda_j(\mathbf{x}_i) \equiv \lambda_j$.

4.2 Identification

The primary source of identification is variation between the policy's ranking (z_i) and treatment effects on components of utility $(\hat{\Delta v_{ij}})$.

Identification requires that some households benefit more than others. Welfare weights $\boldsymbol{\omega}$ are primarily identified based on heterogeneity in impacts on the numeraire utility $\hat{\Delta v_{i0}}$. If treatment effects were homogenous, it would not possible to separately identify $\boldsymbol{\omega}$ and $\boldsymbol{\lambda}$ (their combination may be identified, in which case our method would collapse down to a standard ordinal logit that does not account for treatment effects).

Identification of $\boldsymbol{\lambda}$ also requires that different households benefit by different amounts on the different components of utility. Impact weight λ_j is identified from the relative ranking of households that are impacted more or less on utility component $j > 0$ than on the numeraire $(j = 0)$. If the treatment effects were heterogeneous but colinear between different components of utility, it would be possible to identify $\boldsymbol{\omega}$ but not $\boldsymbol{\lambda}$, because the data would not reveal how different components of utility influences the ranking.

The resulting parameters $\boldsymbol{\omega}$ reveal which characteristics $\mathbf{x}$ are correlated with being prioritized. If $\mathbf{x}$ includes both a relevant variable x_{ik} as well as an irrelevant but colinear variable $x_{ik'}$, the method will have imprecise estimates of the contribution of both. In that sense, one may want to restrict oneself to using characteristics $\mathbf{x}$ that one believes the policymaker may want to use for targeting.

In all of these cases, lack of identification can be recognized from having large standard errors.

This method can also be used if one does not observe a full ranking, but instead a binary allocation ($T_i \in \{0, 1\}$) — which corresponds to a ranking with two levels.

4.3 Discussion

Our approach estimates the preferences that are consistent with the implemented policy z_i, given the estimates of impact $\hat{\Delta v_{ij}}$. This can be thought of as an ex-post audit.

If the policymaker has incorrect beliefs about these impacts at the time of the decision, these implied preferences could differ from their actual preferences. If that were the case, upon observing the results of our method, the policymaker would wish to change the policy to align it with their true preferences. Our method is 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. In that case, our method will reveal the ultimate preferences consistent with the final allocation.

5 Empirical Example

We demonstrate our method on the Mexican PROGRESA conditional cash transfer (CCT) program.

5.1 Context

First implemented by the Mexican federal government in 1997, PROGRESA served as the inspiration for a number of conditional cash transfer programs across Latin America (including the Oportunidades program it evolved into). It was designed to improve the well being of eligible poor families, by offering monthly transfers of 90 pesos to mothers conditional on regular doctor's visits and/or regular school attendance.[5] The vast majority of enrolled households, roughly 99%, met these conditions (Simone Boyce, 2003).[6]

5.1.1 Targeting

PROGRESA targeted poor communities on the basis of a 'village marginality index' (VMI), and targeted poor households within these communities on the basis of a 'household poverty score' (HPS). The VMI was based on a series of village-level variables, including the proportion

[5] In 1999, the exchange rate was 10 pesos per US dollar.

[6] For a more detailed treatment of PROGRESA and its background, see Emmanuel Skoufias (2008), John Hoddinott (2004), and Simone Boyce (2003).

of households living in poverty, population density, and health and education infrastructure. The HPS was based on a household-level proxy means test: surveyors collected data on easily observable characteristics (such as housing materials, family structure, etc.) on all households in eligible communities through a census, and for a small sample, also collected in-depth information on per-capita consumption. The coefficients from a regression of these observable characteristics on per-capita consumption for the in-depth sample then served as the weights for constructing the HPS from these more easily-collected data.

We focus on the preferences implied by household poverty scores, which were the basis for determining a household's eligibility within each community.

5.1.2 Experimental Design

During the early years of its implementation, PROGRESA administrators used a randomized experimental design as part of its staggered rollout across communities: approximately 10% (506) of the 5,000 eligible communities were selected to be part of the evaluation. Among these, 320 communities were randomly assigned to treatment, and initiated into the program in summer 1998. 185 communities were assigned to control and were not initiated into the program until 2000. Behrman and Todd (1999) show that the randomization across communities was successful in that treatment and control communities were statistically indistinguishable across a wide array of observable covariates.

5.1.3 Data

We use data from two household surveys prior to treatment (1996 and May 1998), and one household survey after treatment (November 1999). These surveys asked about household demographics, socioeconomic characteristics, health care utilization, and educational attendance. We evaluate endline outcomes reported in November 1999. These data contain information on 14,949 households over the entire experiment period. Summary statistics for the matched data sample of households present in both periods are presented in Table 1.

5.1.4 Outcomes

To provide a concrete empirical exposition, we focus our analysis on three key outcomes that are commonly prioritized in anti-poverty programs, although only the latter two outcomes were explicitly stated as policy objectives of PROGRESA.

- **Per-capita consumption**

- **Health status of young children**: average number of sick days among children 0-5 years old in the household
- **School attendance**: average number of school days missed among children 6-16 years old in the household

This focus on three outcomes represents a simplified analysis, as real-world policymakers have preferences over a larger set of observable and unobservable factors. We selected these three since previous studies have estimated significant treatment impacts for each one, using the same survey data as we use here (John Hoddinott, 2004; Emmanuel Skoufias, 2008; Simone Boyce, 2003; Djebbari and Smith, 2008). Thus, we assume that the policymaker does not value impacts on other outcomes. In Section 5.2.2, we discuss implications and extensions of this simplifying assumption.

The survey asked for these responses in the previous month. In the terminology of our theory, these outcomes are our $g_j(x_i)$, where $J = 3$; we treat per-capita consumption as our numeraire. We impute treatment effects of 0 for households without children in the relevant age bucket for schooling and health outcomes, respectively (children aged 5 or below for the health intervention, and children aged 6-16 for the schooling intervention).

5.2 Estimates

5.2.1 Heterogeneity in Treatment Effects

We first estimate heterogeneous treatment effect estimates $\Delta\hat{g}_j(\mathbf{x}_i)$. We use Wager and Athey (2018)'s causal forest method, which estimates heterogeneous treatment effects nonparametrically, and includes restrictions to limit overfitting. It allows for considerably more flexible and precise estimates of heterogeneity than possible with linear methods. (Our method can be used with any estimator of heterogeneous treatment effects; corresponding results for OLS are reported in Appendix Section A.4).

On average over our sample, PROGRESA increased household monthly consumption by 14 pesos, reduced the number of sick days per child by 0.08, and reduced the number of school days missed per child by 0.03. However, those treatment effects are heterogeneous, as summarized in Figure 2. Causal forest estimation does not translate into linear coefficients, and so is more difficult to represent in tabular form; however, we report feature importance estimates in Table 2. We find that, by this measure, income and household head age are the most important covariates in describing heterogeneity in impact.

Table 1: Descriptive Statistics

	October 1998 mean	November 1999 mean
Monthly average per capita consumption (pesos)	234.508	178.185
Assigned to treatment group	0.606	0.606
Household poverty score (1997)	695.700	695.700
Village marginality index (1997)	0.470	0.470
Household size	5.75	5.75
... Number of children less than 2 years old	0.692	0.702
... Number of children 3-5 years old	0.577	0.565
... Number of children 6-10 years old	0.948	0.928
... Number of boys 11-14 years old	0.356	0.350
... Number of girls 11-14 years old	0.338	0.332
... Number of boys 15-19 years old	0.318	0.316
... Number of girls 15-19 years old	0.310	0.308
... Number of men 20-34 years old	0.492	0.500
... Number of women 20-34 years old	0.5497	0.555
... Number of men 35-54 years old	0.444	0.445
... Number of women 35-54 years old	0.438	0.439
... Number of men at least 55 years old	0.253	0.254
... Number of women at least 55 years old	0.251	0.253
Head of household:		
... Is male	0.902	0.902
... Is an agricultural worker	0.596	0.600
... Education (in years)	2.703	2.704
... Is indigenous	0.386	0.386
... Age	45.47	45.50
Number of days a child is sick	1.310	0.857
Number of days a child misses school	0.567	0.249
N	14949	14949

To demonstrate how these treatment effects covary with covariates of interest, we present in Appendix Figures 3 and 4 binscatter plots of the final marginal treatment effect estimates against single covariates, after residualizing out variation in treatment effects that can be explained by other variables.[7]

5.2.2 Implied Policymaker Preferences

Table 3 reports estimates for the household poverty score used to allocate eligibility in our sample. The first column reports the ranking $\boldsymbol{z}$ decomposed into covariates, as described by a standard ordinal logit model. This suggests that households that are indigenous, lower income, and have more children are more highly ranked for the program. However, this standard regression does not describe why these households are ranked highly; it could be that they benefit more (higher treatment effects) or are implicitly valued more (higher welfare weights).

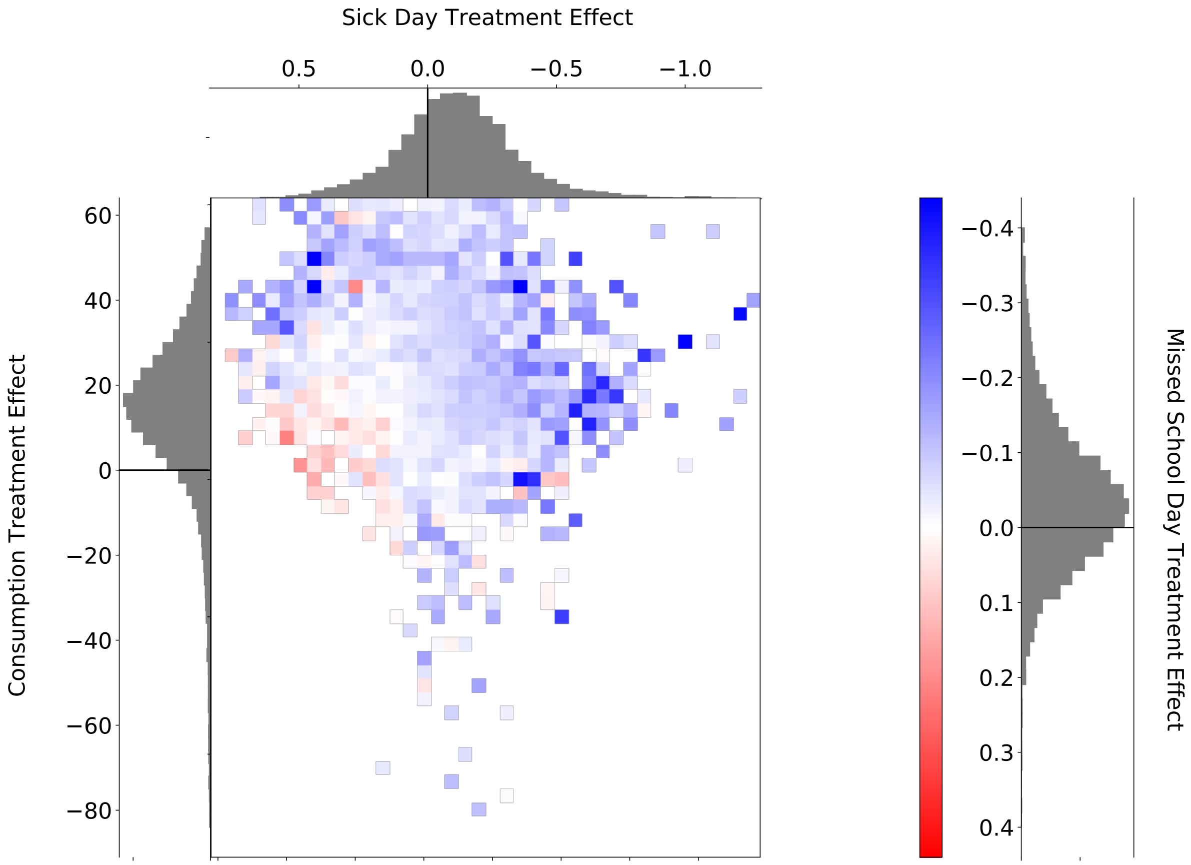
The second column estimates welfare weights and impacts weights, transforming the household poverty scores $\boldsymbol{z}$ using heterogeneous treatment effect estimates $\Delta\hat{g}_j(\mathbf{x}_i)$ and our method. We allow welfare weights $\mu(\mathbf{x}_i)$ to vary over the size of households, the indigenous status of the household head, the level of income in 1997, the number of adults aged 17 or above, and the number of children less than or equal to 5 years old, as well as the number of children 6 to 16 years old.[8] The first block of rows shows the implied welfare weights ($\boldsymbol{\omega}$), and the second block shows implied impact weights ($\boldsymbol{\lambda}$ and C) and the standard deviation of the error term (σ).

Although it is not very informative to directly compare magnitudes between ordinal logit and our estimated welfare weights (ordinal logit models are isomorphic under multiplication by a positive number), we can compare ratios of coefficients, and treatment effects. Indigenous households have lower treatment effects (see Appendix Figures 3), so our method finds that their prioritization can be explained with the policymaker placing higher welfare weight

[7] We present figures for OLS and causal forest, for all three dimensions of treatment effects that we inspect. With this, we can see more clearly how the causal forest method allows for more flexibility than the OLS: while the broad relationship of the estimated effects to covariates most often correspond across both estimation methods, the causal forest method captures much more non-linearity in the relationships, allowing for more precise estimation of the heterogeneity. And in some cases, the relationships are meaningfully different — whereas the OLS enforces more linearity in the relationship between household income and consumption effects, the causal forest is able to capture an undulating pattern that the linear estimator does not, leading to a zero average slope in the causal forest estimates versus a large and declining average slope in the OLS.

[8] We focus on these parameters in particular due to their estimated importance in the causal forest estimation, and due to the fact that treatment effect estimates for health and schooling are only non-zero for households with children in the appropriate categories (for schooling, 6 to 16 years old; for health, less than or equal to 5 years old).

Figure 2: Distribution of Estimated Treatment Effects



Notes: Joint and marginal distributions of estimated treatment effects of PROGRESA conditional cash transfer on schooling, health, and consumption, estimated using causal forest method of Athey and Wager (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 impute 0 for 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.

Table 2: Feature Importance Estimates: Causal Forest

	Consumption	**Health**	**Schooling**
	Monthly per capita	# Sick days	# days missed school
	(pesos)	per child	per child
household head age	0.351	0.181	0.216
log household income '97	0.204	0.192	0.329
household size	0.095	0.236	0.14
household head education	0.19	0.151	0.082
children less than 2 yrs	0.025	0.045	0.086
children 3 to 5 yrs	0.009	0.057	0.034
head agricultural worker	0.016	0.048	0.033
male head of household	0.038	0.021	0.011
num women at least 55 yrs	0.022	0.02	0.017
num women 20 to 34 yrs	0.018	0.015	0.019
num children 6 to 10 yrs	0.005	0.022	0.016
num men at least 55 yrs	0.018	0.005	0.012
num boys 15 to 19 yrs	0.008	0.008	0.005
num boys 11 to 14 yrs	0.0	0.0	0.0
num men 20 to 34 yrs	0.0	0.0	0.0
num women 35 to 54 yrs	0.0	0.0	0.0
num girls 15 to 19 yrs	0.0	0.0	0.0
head indigenous	0.0	0.0	0.0
num girls 11 to 14 yrs	0.0	0.0	0.0
num men 35 to 54 yrs	0.0	0.0	0.0
N	13438	8769	9871

Table 3: Implied Policymaker Preferences

	Household Poverty Score		
	Allocation Rule		Implied Preferences
	Ordinal Logit		*Our Method*
	Regression Coefficients		**Welfare Weights $\boldsymbol{\omega}$**
log(Income)	-0.016		-0.092 (0.043)
Indigenous	0.025		0.116 (0.099)
# Adults	0.035		-0.196 (0.086)
# Children <= 5 years old	0.095		0.209 (0.096)
# Children 6-16 years old	0.112		0.143 (0.073)
			Implied Value
Sickness (per child sick day)		λ_1	12 pesos (351)
Missed Schooling (per day)		λ_2	-17 pesos (368)
Value Regardless of Impact		C	205 pesos (5079)
		σ	0.006 (0.002)
N	13438		13438

Notes: Left column computed with standard ordinal logit.Right column computed using our method, using heterogeneous treatment effects estimated with causal forest (see Figure 2). Standard errors are computed using a two-step bootstrap procedure that accounts for uncertainty in both treatment effects and preference parameters. Observations are drawn with replacement before estimation of the treatment effects and the welfare and impact weights. Treatment effects are then estimated from these bootstrapped samples, and welfare and impact weights estimated from these bootstrapped treatment effect estimates; the standard errors reported are the standard deviation across bootstrapped welfare and impact weight estimates.

on them. Also, although the decision rule prioritizes households with more adults, those households benefit so much more that our method suggests that it the underlying welfare weights decrease with the number of adults.

Overall, we find that allocations are consistent with welfare weights that rank households 13.3 percentiles higher if indigenous, 7.7 percentiles lower for each standard deviation increase in household income, 21.1 percentiles higher for each additional small child (ages 5 and lower) in the household, 15.1 percentiles higher for each additional child aged 6-16, and 20.3 percentiles lower for additional adult. The coefficients on income, children and adults are statistically significantly different from 0 at a 5% level. The welfare weight on indigenous status is estimated slightly less precisely, with a T stat of 1.17.[9]

Most of the implied value of the program comes from simply providing the program, independent from its effect on outcomes (the constant term C). C accounts for 98% of the welfare gain for the median household. This could be consistent with the government intrinsically valuing the transfers that the program provides, even absent persistent impact on the measured outcomes. Alternately, it could be an indication that additional, unmodeled outcomes were used to guide targeting (such as food security, or political considerations). Within our same framework, it would be straightforward to allow for more multidimensional preferences by extending the components of utility j – assuming that data existed on those outcomes. In principle, as the set of modeled characteristics approach the set considered by policymakers, we would expect the share of welfare from C to decline. However, it may not be feasible to identify implied preferences over a large number of modeled outcomes without a much larger sample, given the data demands of the estimation procedure described in Section 4.

The weights on individual outcomes are measured less precisely. 95 confidence intervals suggest that the Mexican government's initial allocation rule implies a value of each missed school day among children below 685 pesos of consumption, and a value of a sick days among young children below 690.40 pesos, but confidence intervals span zero. These valuations can be compared against other estimates of the value of education and health. Based on a review of multiple studies, Psacharopoulos and Patrinos (2018) suggest a 9% average return to a year of schooling. Assuming that these gains accrue once the child is working age, with a lifetime of 40 years of work and a discount rate of 3%, this corresponds to lifetime present-discounted earnings of 2143.35 pesos or \$214.34 per missed year of school.[10] Our

[9] We compute the average percentile change by first computing how each household's projected ranking would shift, given different covariates, and then taking the median change over all households.

[10] Mean monthly consumption is 234.96 pesos. We compute $\sum_{y=16}^{16+40}(0.09 * 234.96 * 12) * (0.97)^{16-7.5+y} =$

95% confidence intervals suggest a value per missed year of school below $1933.12. WHO recommendations which consider health interventions to be a 'best buy' if they can save a DALY for $100 (Laxminarayan et al., 2006), as well as the revealed preference valuations of (Kremer et al., 2011) of $23.68 per DALY for Kenyan households, based on how far they are willing to walk for clean water. If we count each day a child is sick as a day lost, so that a full year of sickness would represent a disability adjusted life year (DALYs), then the government's allocation is consistent with a value per DALY below $1917.97.[11] Achieving more precise estimates of impact weights may require a larger sample.

5.2.3 Surveyed resident preferences

Additionally, we survey a sample of 315 Mexican residents to ask about their preferences for allocating monetary transfers to different types of households and between allocating monetary transfers and other benefits (additional schooling and child health); similar to Saez and Stantcheva (2016). We use multiple price lists to elicit indifference points. We use these survey responses to estimate resident welfare weights and impact weights.

Table 4 compares implied policymaker preferences to the preferences of residents as reported in the survey. Resident preferences are similar in some respects: welfare weights on log income are very close. They differ in others, though the survey has large standard errors. On average, survey respondents value school attendance at 420 pesos per day (standard error 329) and child sick days at 400 pesos (standard error 328). For more details about the survey see Appendix A.3.

5.3 Counterfactuals

We compare the estimates from the government allocation to counterfactual allocations in Table 5. Each column represents a different targeting policy; column 1 repeats estimates from the actual household poverty score used to allocate eligibility in our sample. Panel A presents policymaker preferences: the first block of rows shows welfare weights ($\boldsymbol{\omega}$), and the second block shows impact weights ($\boldsymbol{\lambda}$ and C) and the standard deviation of the error term (σ). Panel B presents counterfactual average outcomes in 1999 if the same number of households were selected under that targeting policy.

2143.35, where 7.5 is the average child age in our sample.

[11] As a rough conversion between sick days and DALYs, we multiply the value of each sick day by the 28 days asked over the survey to convert the valuation of sick days to the valuation of a 'sick year.'

Table 4: Implied Policymaker vs. Resident Preferences

	Household Poverty Score	Resident Preferences
	Implied Preferences	Implied Preferences
	Our Method	*Survey*
Welfare Weights $\boldsymbol{\omega}$		
log(Income)	-0.092 (0.043)	-0.099 (0.159)
Indigenous	0.116 (0.099)	-0.554 (2.068)
# Members	-	-0.104 (0.518)
# Adults	-0.196 (0.086)	-
# Children <= 5 years old	0.209 (0.096)	-
# Children 6-16 years old	0.143 (0.073)	-
Implied Value		
Sickness (per child sick day) λ_1	12 pesos (351)	400 pesos (328)
Missed Schooling (per day) λ_2	-17 pesos (368)	420 pesos (329)
Value Regardless of Impact C	205 pesos (5079)	
σ	0.006 (0.002)	
N	13438	315

Notes: Left column computed using our method, using heterogeneous treatment effects estimated with causal forest (see Figure 2). Right column reports revealed preference estimates from an online survey of Mexican residents. In left column, standard errors are computed using a two-step bootstrap procedure that accounts for uncertainty in both treatment effects and preference parameters. Observations are drawn with replacement before estimation of the treatment effects and the welfare and impact weights. Treatment effects are then estimated from these bootstrapped samples, and welfare and impact weights estimated from these bootstrapped treatment effect estimates; the standard errors reported are the standard deviation across bootstrapped welfare and impact weight estimates. Some respondents in survey did not respond to all questions.

Table 5: Alternate Allocation Rules

	Actual	**Counterfactual**							
	HH Poverty Score	Empirical Impact Weights		Technocratic Impact Weights		Policymaker only values impact on:			2003 Poverty Score
						Education	Health	Consumption	
		Equal Weights	Prefer Poor	Empirical	Equal	Empirical	Empirical	Empirical	
Panel A: Preferences	*Implied*	*Specified*		*Specified*		*Specified*			*Implied*
Welfare Weights $\boldsymbol{\omega}$									
log(Income)	-0.092 (0.043)	0	-1	-0.092	0	-0.092	-0.092	-0.092	-0.01 (0.018)
Indigenous	0.116 (0.099)	0	0	0.116	0	0.116	0.116	0.116	0.191 (0.535)
Number of Adults	-0.196 (0.086)	0	0	-0.196	0	-0.196	-0.196	-0.196	-0.09 (0.286)
Number of Children $\leq$ 5 yrs	0.209 (0.096)	0	0	0.209	0	0.209	0.209	0.209	0.146 (0.314)
Number of Children 6-16 yrs	0.143 (0.073)	0	0	0.143	0	0.143	0.143	0.143	0.101 (0.245)
Impact Weights $\boldsymbol{\lambda}$									
Consumption (pesos)	1	1	1	1	1	1	0	0	1
Health (pesos/child sick day)	12 (351)	12	12	-36	-36	0	12	0	253 (387)
Education (pesos/missed day)	-17 (368)	-17	-17	-97	-97	0	0	-17	207 (729)
Value (pesos)	205 (5079)	205	205	205	205	0	0	0	1947 (153094)
σ	0.006 (0.002)	.	.	.	.	.	.	.	0.0015 (0.0006)
Panel B: Counterfactual outcomes (monthly)									
Consumption (pesos)	225.74	235.35	225.25	225.75	234.60	225.77	226.76	226.74	227.53
Sickness (sick days/child)	1.11	1.11	1.11	1.11	1.107	1.11	1.11	1.11	1.10
Missed school (days/child)	0.27	0.26	0.27	0.27	0.25	0.27	0.27	0.27	0.27
N	13438	13438	13438	13438	13438	13438	13438	13438	13438

5.3.1 Alternate welfare weights

When welfare weights are set equal across households (column 2), the resulting score puts positive weight on log income and puts relatively less weight on households with more small children, and positive weight on households with more adults. When welfare weights rank households solely by log income (column 3), the resulting score deprioritizes households with more small children and also deprioritizes indigenous households.

5.3.2 Technocratic impact weights

In columns 4-5 of Table 5, we keep the original welfare weights but assume technocratic impact weights, as might be input from external valuations. We do not intend to take a stand on these valuations, so these results should be viewed as speculative. We demonstrate results assuming valuations of 1000 pesos ($100) per DALY and 2143.35 pesos per missed school day. The $\boldsymbol{z}'$ ranking implied by our assumed weights is quite similar to the original. By contrast, changing the household covariate welfare weights to an equal weighting across households so that $\mu(\mathbf{x}_i) \equiv 1$ leads to positive weight on income and on household number of adults.

5.3.3 Focus on different outcomes

In practice, implemented policies may balance the desires of multiple advocates. In columns 6-8 of Table 5, we present alternative allocations that value only a single outcome. A hypothetical health minister who values only health effects would advocate for a $\boldsymbol{z}'$ that prioritizes non-indigenous households and also households with fewer children. A hypothetical economics minister who values only consumption outcomes would advocate for a $\boldsymbol{z}'$ that prioritizes households where the head is indigenous and households with more children. A hypothetical education minister who values only schooling outcomes would advocate for a $\boldsymbol{z}'$ which prioritizes households with higher income and fewer small children. In all three of these settings, smaller households (in terms of members of all ages) are given higher priority.

5.3.4 An alternative government scoring rule

In 2003, the Mexican government expanded PROGRESA, and updated their poverty score. In column 9, we find that the welfare weights implied by this rule are largely similar to the welfare weights from the 1997 original ranking, but with less negative weight on income and on households with fewer children, and higher weight on indigenous status. On average, the government would rank households 36.8 percentiles higher if indigenous, 1.4 percentiles lower for each standard deviation increase in log household income, 16.5 percentiles lower for each additional adult, 18.4 percentiles higher for each additional child aged 6-16, and 25.2 percentiles higher for additional child aged 5 and lower. These covariates are estimated with considerably less precision, however, and none are statistically significantly different from 0. The impact weights are imprecisely estimated; we find the valuation of a missed day of school has a 95% confidence interval below 1607 pesos, and the valuation of a young child sick day below 489.73 pesos.

5.4 Extensions

5.4.1 Testing models of policymakers

The method can also be used to test whether policymakers make choices that are internally consistent with a postulated framework. If there is more than one potential treatment or policy, one could estimate different sets of these kind of welfare weights for each one. If there is no way for these weights to be reconciled, you can rule out that policymaking is utilitarian.

5.4.2 Bounding

In settings where heterogeneous treatment effects are unavailable or difficult to measure, one could alternately use this framework to bound the treatment effects that would be consistent with stated preferences.

6 Conclusion

While economists reason about primitives of utility and welfare weights, policy discussions instead commonly revolve around the mechanics of implementation. This paper demonstrates how these primitives can be recovered from observed policies, using a model of preferences and new methods for estimating heterogeneous treatment effects.

We develop this approach and apply it to a large anti-poverty program in Mexico, to estimate the preferences consistent with the program's implementation. Our analysis shows that the program's effects were heterogeneous along multiple dimensions. After accounting for these differential effects, observed allocations are consistent with placing higher weight on the welfare of indigenous, poor, and large families. The implied value of each missed school day and child sick day is estimated imprecisely but is consistent with valuations estimated in prior work, and is also consistent with the stated preferences reported by Mexican residents in a choice survey.

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 an existing program, to help hold policymakers accountable for past decisions – and in particular, to evaluate whether the implemented allocation reflects the stated goals of the policy. Perhaps most importantly, this approach can be used to adjust existing policies to better align with those goals.

References

Alatas, V., A. Banerjee, R. Hanna, B. A. Olken, and J. Tobias (2012): "Targeting the Poor: Evidence from a Field Experiment in Indonesia," *American Economic Review*, 102, 1206–1240.

Athey, S. and S. Wager (2020): "Policy Learning with Observational Data," *arXiv:1702.02896 [cs, econ, math, stat]*, arXiv: 1702.02896.

Barocas, S., M. Hardt, and A. Narayanan (2018): *Fairness and Machine Learning*, fairmlbook.org.

Barr, N. (2012): *Economics of the welfare state*, Oxford university press.

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

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

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

Djebbari, H. and J. Smith (2008): "Heterogeneous impacts in PROGRESA," *Journal of Econometrics*, 145, 64–80.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel (2012): "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, ACM, 214–226.

Emmanuel Skoufias, V. D. M. (2008): "Conditional Cash Transfers, Adult Work Incentives, and Poverty," *Journal of Development Studies*, 44, 935–960.

Ensign, D., S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian (2017): "Runaway Feedback Loops in Predictive Policing," *arXiv:1706.09847 [cs, stat]*, arXiv: 1706.09847.

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

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

Gechter, M., C. Samii, R. Dehejia, and C. Pop-Eleches (2019): "Evaluating Ex Ante Counterfactual Predictions Using Ex Post Causal Inference," *arXiv:1806.07016 [stat]*, arXiv: 1806.07016.

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

Greco, S., A. Ishizaka, M. Tasiou, and G. Torrisi (2019): "On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness," *Social Indicators Research*, 141, 61–94.

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

Hendren, N. (2019): "Efficient Welfare Weights," Working Paper 20351, National Bureau of Economic Research.

Hu, L. and Y. Chen (2018): "Welfare and Distributional Impacts of Fair Classification," *arXiv:1807.01134 [cs, stat]*, arXiv: 1807.01134.

Jayachandran, S., M. Biradavolu, and J. Cooper (2021): "Using Machine Learning and Qualitative Interviews to Design a Five-Question Women's Agency Index," Tech. Rep. w28626, National Bureau of Economic Research.

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

Kasy, M. and R. Abebe (2020): "Fairness, equality, and power in algorithmic decision making," in *ICML Workshop on Participatory Approaches to Machine Learning*.

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

Kremer, M., J. Leino, E. Miguel, and A. P. Zwane (2011): "Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions," *The Quarterly Journal of Economics*, 126, 145–205.

Laxminarayan, R., J. Chow, and S. A. Shahid-Salles (2006): *Intervention Cost-Effectiveness: Overview of Main Messages*, The International Bank for Reconstruction and Development / The World Bank.

Liu, L. T., S. Dean, E. Rolf, M. Simchowitz, and M. Hardt (2018): "Delayed Impact of Fair Machine Learning," in *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, vol. 80 of *Proceedings of Machine Learning Research*, 3156–3164.

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

Mouzannar, H., M. I. Ohannessian, and N. Srebro (2018): "From Fair Decision Making to Social Equality," *arXiv:1812.02952 [cs, stat]*, arXiv: 1812.02952.

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

Noriega, A., B. Garcia-Bulle, L. Tejerina, and A. Pentland (2018): "Algorithmic Fairness and Efficiency in Targeting Social Welfare Programs at Scale," *Bloomberg Data for Good Exchange Conference*.

Psacharopoulos, G. and H. A. Patrinos (2018): "Returns to investment in education," .

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

Rolf, E., M. Simchowitz, S. Dean, L. T. Liu, D. Björkegren, M. Hardt, and J. Blumenstock (2020): "Balancing Competing Objectives with Noisy Data: Score-Based Classifiers for Welfare-Aware Machine Learning," .

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

Simone Boyce, P. G. (2003): "An Experiment in Incentive-Based Welfare: The Impact of PROGRESA on Health in Mexico," Royal Economic Society, vol. 85.

Skoufias, E., B. Davis, and S. de la Vega (2001): "Targeting the Poor in Mexico: An Evaluation of the Selection of Households into PROGRESA," *World Development*, 29, 1769–1784.

Timmins, C. (2003): "Measuring the Dynamic Efficiency Costs of Regulators' Preferences: Municipal Water Utilities in the Arid West," *Econometrica*, 70, 603–629.

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

UNDP (1990): "Human Development Report 1990: Concept and Measurement of Human Development," Tech. rep.

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

Wang, F. (2020): "The Optimal Allocation of Resources Among Heterogeneous Individuals," *Available at SSRN*.

A Appendix

A.1 Generalized curvature in utility components

This section considers what will be measured if the utility functions are assumed to be linear ($\tilde{g}_j(y) = y$) but in fact the true utility functions $g_j(y)$ have curvature. The true impact of the program on component of utility 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 $g_j(y_{ij} + \Delta) \approx g_j(y_{ij}) + \Delta g_j'(y_{ij})$. Thus for any $g_j(\cdot)$ we have:

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

We can then express the utility benefit of treating i as:

$$\Delta S_i \approx \underbrace{\mu(\mathbf{x}_i) g_0'(y_{i0})}_{\tilde{\mu}(\mathbf{x}_i,\{y_{i0}\})} \left[\widehat{\Delta y}_0(\mathbf{x}_i) + \sum_j \underbrace{\lambda_j(\mathbf{x}_i) g_j'(y_{ij})}_{\tilde{\lambda}_j(\mathbf{x}_i,\{y_{ij}\})} \widehat{\Delta y}_j(\mathbf{x}_i) \right]$$

This implies that if we do not specifically account for curvature and estimate a linear model, the welfare and impact weights we estimate ($\tilde{\mu}$ and $\tilde{\boldsymbol{\lambda}}$) are approximately a combination of the underlying welfare and impact weights (μ and $\boldsymbol{\lambda}$) 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 $\tilde{\mu}$ coincides with μ and $\tilde{\boldsymbol{\lambda}}$ with $\boldsymbol{\lambda}$. Otherwise, utility curvature multiplies the weights.

A.2 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 across multiple years. Three different survey years are used: a baseline survey in October 1997, and two follow-up surveys in October 1998 and November 1999. Villages were randomly assigned to treatment and control groups, with the latter joining the program two years later than the former. Within villages in the treatment group, a poverty index score is computed based on household income and assets, and all households meeting the score requirement are eligible to receive the program's conditional transfers. In our specified timeframe, there were two rounds of eligibility classification: the first during the 1997 baseline survey and the second during a July 1999 survey. The second round increased the scope of the program to include more households, and we use this criteria to classify eligible households.

We compute a measure of average household monthly consumption per member based on the survey responses. The October 1998 and November 1999 surveys ask households about the quantity consumed, quantity purchased and amount of money expended 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 of money expended to construct household-specific prices which are then multiplied by 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 smallest level available).

A.3 Preference survey

We additionally survey 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.

A.3.1 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 translated in Mexican Spanish. First, respondents were asked to select which attributes the government should consider when prioritizing which households receive cash transfers, from a list (age, income, household size, education, agricultural, indigenous, and gender). Second, subjects were asked to make monetary allocation decisions between different households using multiple price lists (see Figure 5 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.[12] Each subject filled in one price list for each focal attribute. Third, for a particular household, subjects were asked to make allocation decisions between money and education and child health using multiple price lists (see Figure 6). The description of the household included three randomly selected control attributes. Finally, subjects were asked for basic demographics.

[12] Each participant saw all of the tradeoff numbers multiplied by 1x, 2x, or 3x, selected at random.

A.3.2 Estimation

We use the survey responses to estimate $\boldsymbol{\omega}$ and $\boldsymbol{\lambda}$:

To identify $\boldsymbol{\omega}$, compare impacts in dollars of consumption (where other impacts $\Delta g_j(x_i) = 0$). Then if individual i differs from i' only in attribute j and the crossover point is $\Delta g_0(x_i) = a$ and $\Delta g_0(x_{i'}) = b$, then we must have:

For additive welfare weights:

$$\begin{aligned}[1 + \omega_j x_{i,j}]a &= [1 + \omega_j x_{i',j}]b \\ \omega_j &= -\frac{b-a}{x_{i',j}b - x_{i,j}a}\end{aligned}$$

For multiplicative welfare weights:

$$\begin{aligned}\omega_{-j}^{x_{i,-j}} \omega_j^{x_{i,j}} a &= \omega_{-j}^{x_{i',-j}} \omega_j^{x_{i',j}} b \\ \omega_j &= \left(\frac{b}{a}\right)^{\frac{1}{x_{i,j}-x_{i',j}}}\end{aligned}$$

To identify $\boldsymbol{\lambda}$, now instead hold fixed individual attributes, and consider impacts on different outcomes. If the crossover point is $\Delta g_0(x_i) = a$ and $\Delta g_j(x_i) = b$ then $\lambda_j = \frac{a}{b}$.

A.3.3 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 understanding of the study. 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.[13] 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.

[13] 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.

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.[14] 91% of respondents agreed with the first and last statement, and disagreed with the middle; 58% did so strongly.

There was an optional comment box at the conclusion of the survey; 49% of respondents filled in a comment, suggesting a high level of engagement with the survey. Although some respondents used the box to indicate some confusion with the selector interface, several respondent affirmatively to the approach of basing policy on resident preferences, such as (translated to English):

- ‘Excellent that they do these surveys to assess the policies of support to families’
- ‘I think this survey was very important since the benefits that sometimes come are the same for all people and the situations of each person are not considered. For some it may be enough but for others it is too little.’
- ‘excellent survey, hopefully and we could society decide these support, because that is how we would eradicate poverty’

A.4 OLS Treatment Effect Estimates

A.4.1 Estimation: OLS

To estimate the potentially heterogeneous impacts of PROGRESA on our set of outcome variables, we first follow Djebbari and Smith (2008) and estimate a linear regression equation. We allow treatment effects to vary by age and gender composition of the household, total household size, and several characteristics of the household head: education level, indigenous status, gender, working in the agricultural sector, and age.[15] Formally, we estimate:

$$g_{ij} = \beta_0 + \boldsymbol{\beta}_{\mathbf{x}}\mathbf{x}_i + (\beta_T + \boldsymbol{\beta}_{T\mathbf{x}}\mathbf{x}_i)T_i + e_i \tag{2}$$

[14] Apart from two pilot respondents.

[15] We depart from Djebbari and Smith (2008) in that we omit poverty scores and village marginality index and their respective interactions in the list of covariates, to avoid potential correlated errors from using these rankings in both the treatment effect estimates and in the preference-learning method.

where g_{ij} is the endline outcome, $\mathbf{x}_i$ is the vector of baseline covariates, and $T_i \in \{0, 1\}$ is a dummy variable for treatment status of household i. This model allows endline outcomes to differ systematically according to household covariates, and additionally allows the treatment effect of PROGRESA to differ across households according to their covariates.

We construct our variables for treatment effects from the predicted values from our estimated Equation 2, as

$$\Delta\hat{g}_j(\mathbf{x}_i) = \hat{\beta}_T + \hat{\boldsymbol{\beta}}_{T\mathbf{x}}\mathbf{x}_i$$

A.4.2 Results

On average over our sample, using OLS PROGRESA increased household monthly consumption by 12.87, to have reduced the number of sick days per child by 0.111, and reduced the number of school days missed per child by 0.039.

However, the effects of the program differ across households. The overall distributions of treatment effects by outcome for OLS, are presented in Figure 7. The distribution of estimated effects estimated under causal forest is tighter, in particular for the schooling and health outcomes. With OLS, we see a fairly strong correlation between health treatment effect estimates and schooling treatment effect estimates, but with causal forest this correlation is much less apparent.

OLS coefficient estimates are presented in Table 6, with standard errors in parentheses. Similar to Djebbari and Smith (2008), our OLS point estimates show that consumption treatment impacts are higher for households with indigenous status and male heads of households, and lower for larger households.[16]

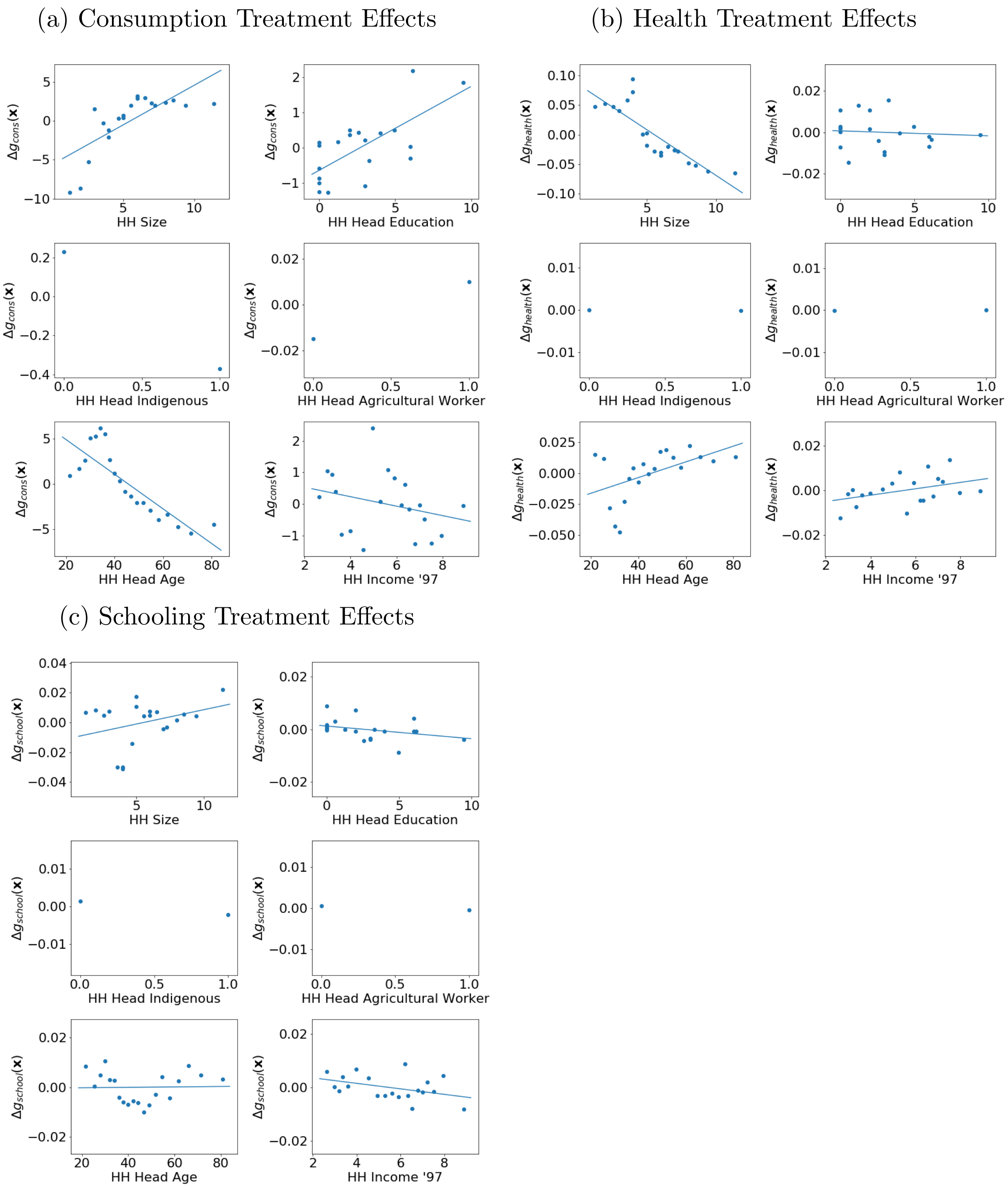
[16] Note that our specification differs meaningfully from Djebbari and Smith (2008) in that we exclude the ranking metrics from the list of covariates, to avoid correlated errors with our subsequent estimation step. Therefore, the estimates are not directly comparable.

Table 6: Treatment Effect Coefficient Estimates: OLS

	Consumption (Monthly avg. per person, in pesos)	Health (Avg. sick days per child)	Schooling (Avg. days of missed school per child)
Treatment	16.9658 (63.022)	0.0507 (0.2)	0.1796 (0.271)
Treatment X num child less than 2 yrs	-5.1289 (7.082)	0.0157 (0.075)	-0.0487 (0.034)
Treatment X num child 3 to 5 yrs	-1.0651 (8.529)	-0.0992 (0.091)	0.0286 (0.04)
Treatment X num child 6 to 10 yrs	2.5588 (7.148)	-0.0548 (0.078)	-0.0071 (0.033)
Treatment X num boys 11 to 14 yrs	7.49 (9.802)	-0.0889 (0.111)	-0.0105 (0.042)
Treatment X num girls 11 to 14 yrs	-0.7555 (230.95)	-1.0672 (1.548)	0.0557 (0.967)
Treatment X num boys 15 to 19 yrs	-8.4626 (10.547)	-0.0561 (0.123)	0.0719 (0.047)
Treatment X num girls 15 to 19 yrs	42.6459 (238.282)	-0.0 (0.0)	0.2002 (0.998)
Treatment X num men 20 to 34 yrs	3.6628 (11.498)	-0.0945 (0.137)	0.0279 (0.056)
Treatment X num women 20 to 34 yrs	21.3229 (119.141)	-0.0 (0.0)	0.1001 (0.499)
Treatment X num men 35 to 54 yrs	-14.1408 (15.313)	-0.1827 (0.188)	0.1056 (0.075)
Treatment X num women 35 to 54 yrs	125.7494 (224.113)	-0.6767 (1.549)	-0.1632 (0.938)
Treatment X num men at least 55 yrs	2.8881 (17.454)	-0.2877 (0.205)	0.1518 (0.085)
Treatment X num women at least 55 yrs	-45.4099 (413.329)	-0.8655 (2.084)	0.0573 (1.73)
Treatment X household size	2.4514 (4.429)	0.0654 (0.051)	-0.0079 (0.02)
Treatment X male head of household	-4.3571 (61.282)	0.0507 (0.2)	0.0796 (0.261)
Treatment X head agricultural worker	42.422 (31.097)	-0.2727 (0.381)	0.0424 (0.148)
Treatment X head education	-3.5337 (2.23)	-0.0511 (0.024)	-0.0248 (0.011)
Treatment X head indigenous	23.6333 (10.833)	0.1268 (0.123)	-0.0467 (0.053)
Treatment X head age	-1.0169 (0.58)	0.0056 (0.007)	-0.0039 (0.003)
Treatment X log household income ’97	9.7069 (7.241)	-0.1135 (0.084)	-0.0357 (0.036)
Baseline Covariates	X	X	X
R^2	0.045	0.020	0.010
N	13438	8769	9871

Notes: Baseline covariates here includes the covariates without interaction with treatment effects, e.g. head age, as well as a constant term.*p<0.1; **p<0.05; ***p<0.01.

Figure 3: Binscatter of Treatment Effects and Selected Covariates: Causal Forest



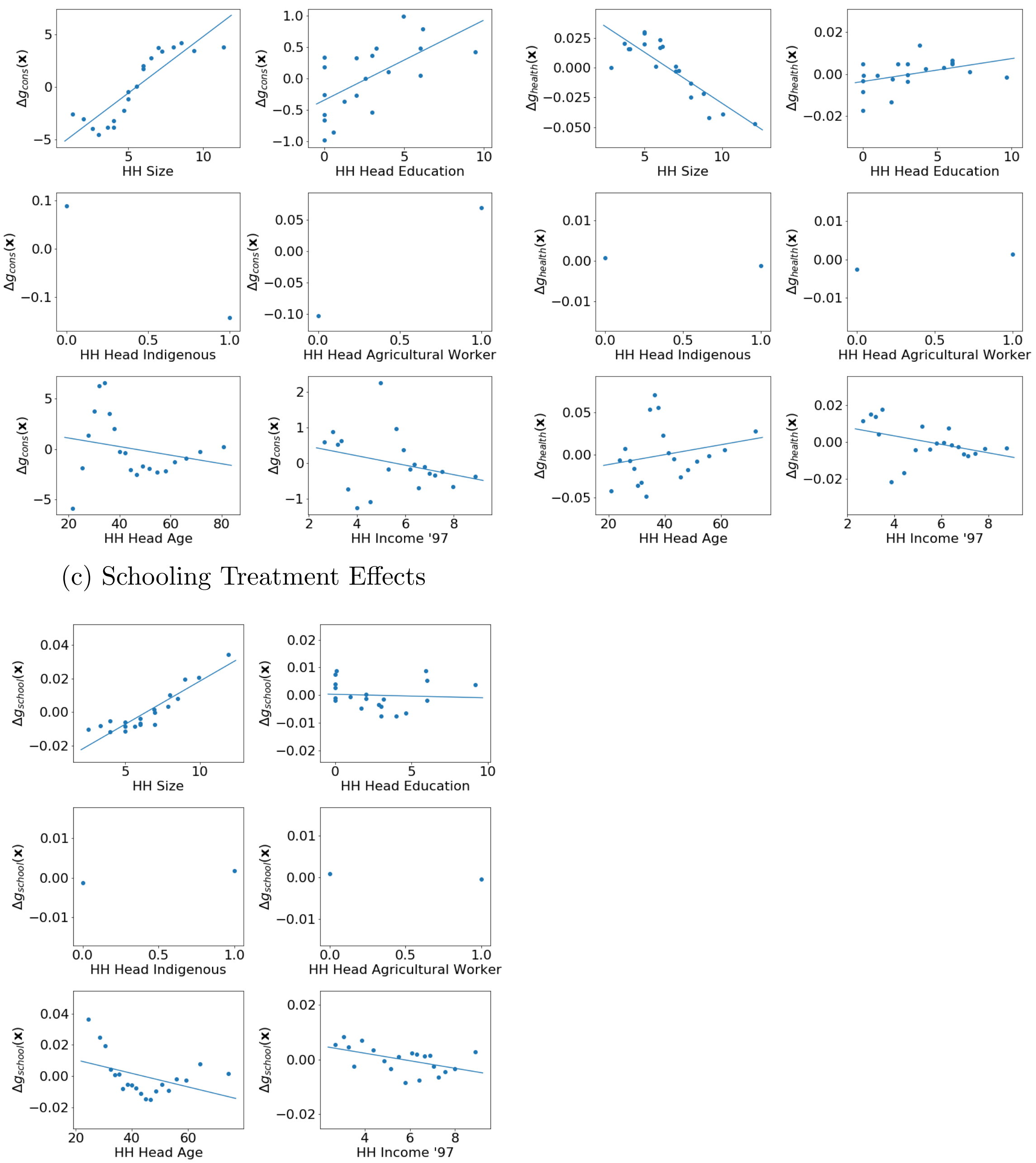
(a) Consumption Treatment Effects

(b) Health Treatment Effects

(c) Schooling Treatment Effects

Notes: Binscatter plots of treatment effects from causal forest over a selected group of six 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, including some not shown.

Figure 4: Binscatter of Treatment Effects and Selected Covariates: OLS



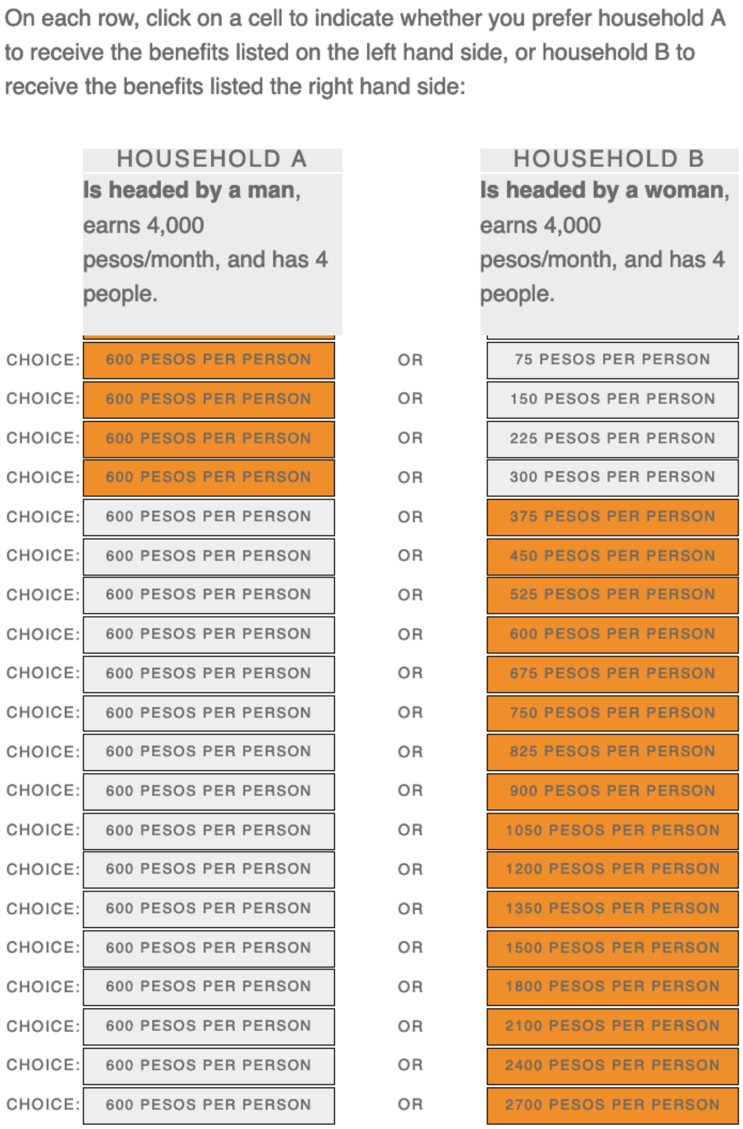
(a) Consumption Treatment Effects

(b) Health Treatment Effects

(c) Schooling Treatment Effects

Notes: Binscatter plots of treatment effects from OLS over a selected group of six 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, including some not shown.

Figure 5: Welfare Weight Survey Question Example

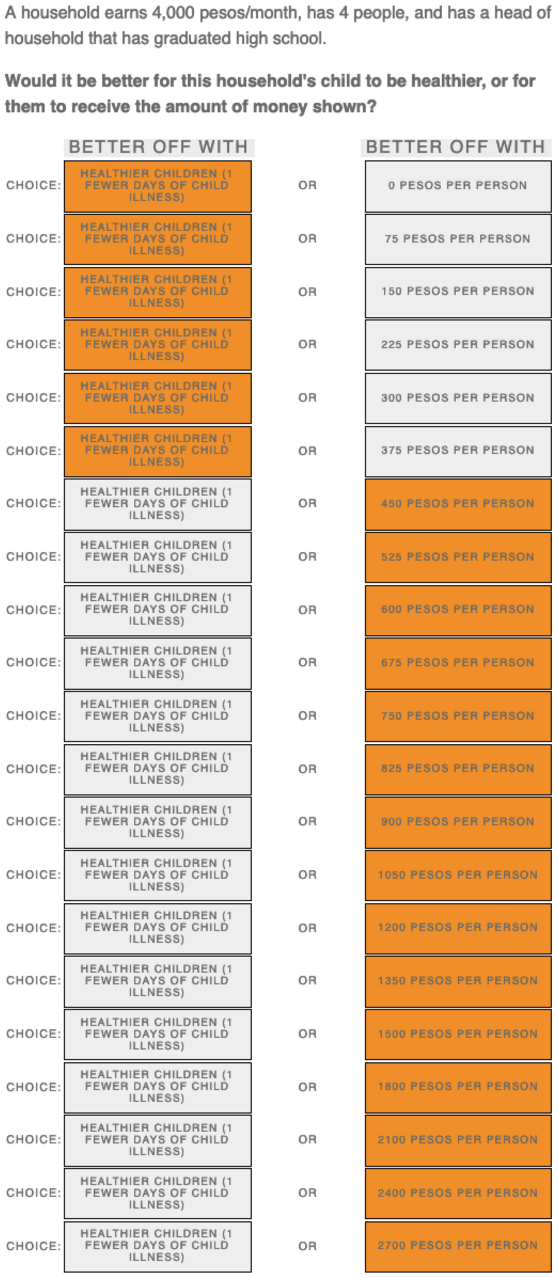


On each row, click on a cell to indicate whether you prefer household A to receive the benefits listed on the left hand side, or household B to receive the benefits listed the right hand side:

	HOUSEHOLD A		HOUSEHOLD B
	Is headed by a man, earns 4,000 pesos/month, and has 4 people.		**Is headed by a woman,** earns 4,000 pesos/month, and has 4 people.
CHOICE:	[✓ 600 PESOS PER PERSON]	OR	75 PESOS PER PERSON
CHOICE:	[✓ 600 PESOS PER PERSON]	OR	150 PESOS PER PERSON
CHOICE:	[✓ 600 PESOS PER PERSON]	OR	225 PESOS PER PERSON
CHOICE:	[✓ 600 PESOS PER PERSON]	OR	300 PESOS PER PERSON
CHOICE:	600 PESOS PER PERSON	OR	[✓ 375 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 450 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 525 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 600 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 675 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 750 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 825 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 900 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 1050 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 1200 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 1350 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 1500 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 1800 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 2100 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 2400 PESOS PER PERSON]
CHOICE:	600 PESOS PER PERSON	OR	[✓ 2700 PESOS PER PERSON]

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

Figure 6: Impact Weight Survey Question Example



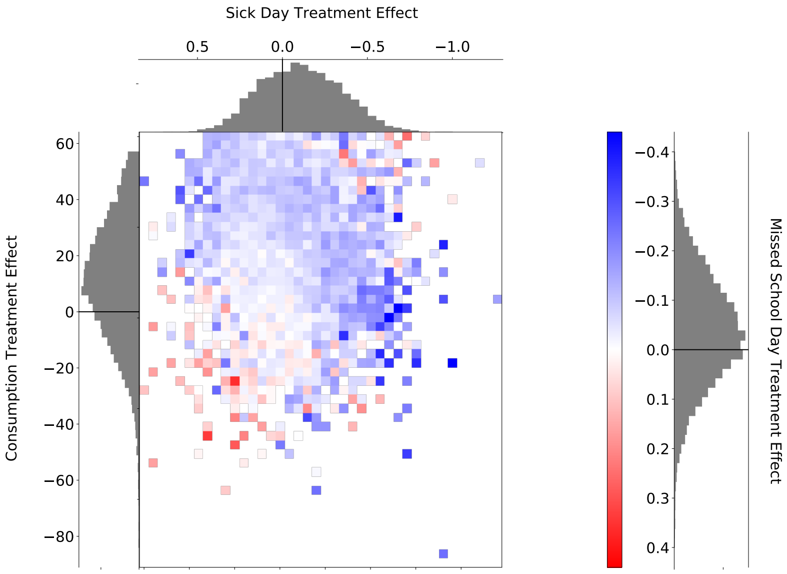
A household earns 4,000 pesos/month, has 4 people, and has a head of household that has graduated high school.

Would it be better for this household's child to be healthier, or for them to receive the amount of money shown?

	BETTER OFF WITH		BETTER OFF WITH
CHOICE:	[✓ HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)]	OR	0 PESOS PER PERSON
CHOICE:	[✓ HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)]	OR	75 PESOS PER PERSON
CHOICE:	[✓ HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)]	OR	150 PESOS PER PERSON
CHOICE:	[✓ HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)]	OR	225 PESOS PER PERSON
CHOICE:	[✓ HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)]	OR	300 PESOS PER PERSON
CHOICE:	[✓ HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)]	OR	375 PESOS PER PERSON
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 450 PESOS PER PERSON]
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 525 PESOS PER PERSON]
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 600 PESOS PER PERSON]
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 675 PESOS PER PERSON]
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 750 PESOS PER PERSON]
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 825 PESOS PER PERSON]
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 900 PESOS PER PERSON]
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 1050 PESOS PER PERSON]
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 1200 PESOS PER PERSON]
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 1350 PESOS PER PERSON]
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 1500 PESOS PER PERSON]
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 1800 PESOS PER PERSON]
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 2100 PESOS PER PERSON]
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 2400 PESOS PER PERSON]
CHOICE:	HEALTHIER CHILDREN (1 FEWER DAYS OF CHILD ILLNESS)	OR	[✓ 2700 PESOS PER PERSON]

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

Figure 7: Distribution of Estimated Treatment Effects (OLS)



Notes: Joint and marginal distributions of estimated treatment effects of PROGRESA conditional cash transfer on schooling, health, and consumption, estimated using OLS. 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 impute 0 for 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.