in the target contexts, in which case the $\mathbb{P}_c[(Y_{1ic}, W_{ic})|V_c, D_c = 1]$ distributions are revealed. This allows the planner to judge, *ex post*, the quality of the methods in terms of how their recommendations fair with respect to welfare.

Conditions C1-C3 include situations where we have a set of randomized experiments that we can feed into a set of methods to produce *ex ante* recommendations for a new target context. These conditions also admit observational studies in which conditional independence and overlap of treatment assignment holds over the covariate space, $\mathcal{W}_c$, although for the time being we focus on randomized experiments. Our specification of the potential outcomes, combined with assumption C3, rules out "interference" (Cox, 1958, p. 19), including general equilibrium effects. We consider this a starting point for our analysis. Generalizations could consider situations where interference is present, and derive criteria for judging methods by working with causal quantities that are identified under interference (Hudgens and Halloran, 2008; Aronow and Samii, 2017).

3 Planner's Objective and *Ex Ante* Recommendations

We can define the planner's objective in context c in general terms as

$$\max_{\pi} \mathcal{U}(\mathbb{P}_c(\pi))$$

where $\pi : \mathcal{V} \times \mathcal{W}_c \to \{0, 1\}$ is a treatment assignment rule that takes context c's contextual attributes and an individual's covariate values and maps them to a treatment assignment. Then, $\mathbb{P}_c(\pi)$ is the joint potential outcome-covariate distribution induced by π. The restriction to $\{0, 1\}$ for each individual in the target context follows from our focus on binary treatments. If a richer set of treatment values $\mathcal{T}$ were under consideration, the rule π could assign distribution functions over this $\mathcal{T}$ on the basis of covariate values. Sometimes the set of conditional treatment assignments that maximizes this objective *ex ante* is non-unique—i.e., there are ties. For example, multiple units may share the same covariate value. The treatment assignment that maximizes the objective may assign some fraction of such units to treatment. Then, all permutations of assignments would yield the same *ex ante* value for the objective. We assume that π encodes a tie-breaker rule that is unconditionally statistically independent and equalizes probability of treatment for such tied units.

Current approaches to this problem (as in Manski, 2004 and Tetenov, 2012) consider social welfare that is linear in expected treatment and control outcomes, what Hirano and Porter (2019) refer to as utilitarian welfare. We adopt the same approach.[2] We also incorporate treatment opportunity costs by adding a cost-effectiveness term to the planner's objective. This increases the planner's sensitivity to over- and underestimation of the treatment effects in generating treatment rules. We suppose that we are operating in context c, and therefore suppress the associated indexing except when necessary for clarification. Thus, we define

$$\mathcal{U}(\mathbb{P}(\pi)) =$$
$$\mathbb{E}\left\{\pi(W_i)\mathbb{E}\left[Y_{1i} - Y_{0i} - \frac{C(W_i, Y_{1i})}{\kappa}\middle|W_i\right] + (1-\pi(W_i))\mathbb{E}\left[\frac{C(W_i, Y_{1i})}{\kappa} - (Y_{1i} - Y_{0i})\middle|W_i\right]\right\}$$
$$= 2\mathbb{E}\left\{\pi(W_i)\mathbb{E}\left[Y_{1i} - Y_{0i} - \frac{C(W_i, Y_{1i})}{\kappa}\middle|W_i\right]\right\} - \mathbb{E}\left[Y_{1i} - Y_{0i} - \frac{C(W_i, Y_{1i})}{\kappa}\right],$$

where $\kappa > 0$ is the level of cost effectiveness at which the planner is indifferent between allocating funds to treat an individual and using the funds for some alternative purpose (Garber and Phelps, 1997; Dhaliwal et al., 2011), and $C(w, y) \geq 0$ is the

[2] For the current analysis, we put to the side considerations related to the planner's sensitivity to inequality or uncertainty (Dehejia, 2008), as well as asymmetry in the planner's preferences toward different treatments, such as status quo bias (Tetenov, 2012).

marginal cost function for treating an individual with covariate value w and treated outcome y. Choice of a treatment assignment with this specification will be equivalent to operating under a budget constraint that must be satisfied in expectation *ex ante* (cf. Bhattacharya and Dupas, 2012, versus *ex post*, as in Kitagawa and Tetenov, 2017), where the marginal value of relaxing the budget constraint is $1/\kappa$.

The prediction method m translates the planner's objective into an *ex ante* recommendation for a treatment assignment rule, π_m. Method m does not have access to $\mathbb{P}$ and rather must rely on some approximation, $\widehat{\mathbb{P}}_m$. As such, method m solves,

$$\pi_m = \arg\max_{\pi} \mathcal{U}(\widehat{\mathbb{P}}_m(\pi)),$$

where $\widehat{\mathbb{P}}_m(\pi)$ can be informed by samples from the reference contexts and status quo data from the target. Under our specification of the welfare function, we have

$$\pi_m = \arg\max_{\pi} \widehat{\mathbb{E}}_m \left\{ \pi(W_i) \widehat{\mathbb{E}}_m \left[Y_{1i} - Y_{0i} - \frac{C(w, Y_{1i})}{\kappa} \middle| W_i \right] \right\}.$$

In practice, we will use a "plug-in" rule where $\hat{E}_m[\cdot]$ is generated from model m estimated on the reference data and status quo data from the target (Hirano and Porter, 2019). The problem that method m solves generates a decision rule analogous to Manski (2004)'s conditional empirical success rule:

$$\pi_m(w) = 1 \left\{ \widehat{\mathbb{E}}_m \left[Y_{1i} - Y_{0i} - \frac{C(w, Y_{1i})}{\kappa} \middle| w \right] \geq 0 \right\}.$$

In our application to conditional cash transfer programs, the cost function takes a form that allows for further simplifications. Those details are presented below.

4 *Ex Post* Inference

The welfare contrast for two methods, l and m, is given by

$$\Delta_{lm} = \mathbb{E}\left\{(\pi_l(W_i) - \pi_m(W_i))\mathbb{E}\left[Y_{1i} - Y_{0i} - \frac{C(W_i, Y_{1i})}{\kappa}\middle| W_i\right]\right\},$$

where $2\Delta_{lm}$ would correspond to the difference in utilities as per our specification of $\mathcal{U}(\mathbb{P}(\pi))$ above. In constructing this *ex post* contrast, we condition on any data used to generate the $\hat{\mathbb{P}}$s. The welfare constrast Δ_{lm} is non-zero for values of w for which the methods disagree on treatment assignment (i.e., w for which $\pi_l(w) \neq \pi_m(w)$). It equals the marginalized value of the minimum-benefit-adjusted conditional average treatment effects when m says to treat and l says not to (or the reverse).

Ex post, we suppose that we obtain a random sample of experimental units in the target context for which conditions C1-C3 hold. We assume that in this *ex post* experiment, treatment assignment probabilities are given by $p(W_i)$, and that these probabilities are known. The experimental data in our target context allow us to estimate this welfare contrast. Given a random sample of size N in the target context, we consider an estimator for the linear welfare contrast based on inverse-probability of treatment weighting with normalized weights. This estimator is efficient among consistent estimators that avoid modeling of either the potential outcome surfaces or conditional treatment probabilities (Hirano et al., 2003; Lunceford and Davidian, 2004; Imbens and Wooldridge, 2009, 35). We define the estimator as

$$\hat{\Delta}_{lm} = \frac{\sum_{i=1}^{N} \frac{T_i}{p(W_i)} (\pi_l(W_i) - \pi_m(W_i)) \left(Y_i - \frac{C(W_i, Y_i)}{\kappa}\right)}{\sum_{i=1}^{N} \frac{T_i}{p(W_i)}} - \frac{\sum_{i=1}^{N} \frac{(1-T_i)}{1-p(W_i)} (\pi_l(W_i) - \pi_m(W_i))\, Y_i}{\sum_{i=1}^{N} \frac{(1-T_i)}{1-p(W_i)}}.$$

Inference for this estimator is based on the random sampling of (Y_i, T_i, W_i) values from $\mathbb{P}$ under conditions C0-C3:

Proposition 1. *Under conditions C0-C3, as* $N \to \infty$

$$\frac{\sqrt{N}\left(\hat{\Delta}_{lm} - \Delta_{lm}\right)}{\sqrt{\hat{V}_{\Delta_{lm}}}} \xrightarrow{d} N(0,1),$$

where

$$\hat{V}_{\Delta^g_{lm}} = \frac{\sum_{i=1}^{N} \frac{T_i}{p(W_i)^2} \left[(\pi_l(W_i) - \pi_m(W_i))(Y_i - \frac{C(W_i,Y_i)}{\kappa}) - \hat{\delta}_1\right]^2}{\sum_{i=1}^{N} \frac{T_i}{p(W_i)}} + \frac{\sum_{i=1}^{N} \frac{1-T_i}{[1-p(W_i)]^2} \left[(\pi_l(W_i) - \pi_m(W_i))Y_i - \hat{\delta}_0\right]^2}{\sum_{i=1}^{N} \frac{1-T_i}{1-p(W_i)}},$$

and

$$\hat{\delta}_1 = \frac{\sum_{i=1}^{N} \frac{T_i}{p(W_i)} \left(\pi_l(W_i) - \pi_m(W_i)\right)\left(Y_i - \frac{C(W_i,Y_i)}{\kappa}\right)}{\sum_{i=1}^{N} \frac{T_i}{p(W_i)}}$$

$$\hat{\delta}_0 = \frac{\sum_{i=1}^{N} \frac{1-T_i}{1-p(W_i)} \left(\pi_l(W_i) - \pi_m(W_i)\right) Y_i}{\sum_{i=1}^{N} \frac{1-T_i}{1-p(W_i)}}.$$

All proofs are contained in the appendix. Conditional on $W_i = w$, the recommendations, $\pi_m(w)$ and $\pi_l(w)$, are fixed. Our uncertainty about the welfare contrast is due to sampling and treatment assignment variation in the experimental data gathered in the target context that we use for the *ex post* assessment.[3]

The *ex post* situation that we consider is simple in that all randomization and sampling occurs at the unit level and there is no causal interference in the outcome

[3] Treating the treatment assignment rules as fixed means that we can have expert opinion among the methods considered in $\mathcal{M}$. Diebold (2015) makes this point in reviewing Diebold and Mariano (1995) and the literature following it, drawing a distinction between Diebold and Mariano (1995) and West (1996), which additionally considers uncertainty arising from the samples on which models are fitted to generate predictions.

data generating process. Certainly the analysis could be extended to consider cluster randomization or cluster sampling, covariate adjustment, and targeting quantities that are identified under interference. For the present analysis, we stick with this relatively simple case.

Proposition 1 is sufficient to perform inference for any pair of methods. Hansen et al. (2011) provide a sequential multiple testing algorithm for establishing a "model confidence set" (MCS) of level $1 - \alpha$, which allows one to distinguish a set of best performing algorithms with an asymptotic error rate of α.

5 Empirical Application

For an empirical illustration, we use data from Mexico and Morocco on the effects of conditional cash transfers (CCTs) on primary school enrollment. We consider a policy scenario where a planner in Morocco is seeking recommendations for implementing a conditional cash transfer (CCT) program. The planner's objective is defined as above, accounting for opportunity costs. We use data from randomized evaluations of the PROGRESA program in Mexico (Schultz, 2004; Behrman et al., 2005; De Janvry and Sadoulet, 2006; Todd and Wolpin, 2006; Attanasio et al., 2012) and the TAYSSIR program in Morocco (Benhassine et al., 2015). To construct the *ex ante* evaluations, we are limited to using the full data from Mexico and then the covariate and control group outcome data from Morocco. The *ex post* assessment is done using the full experimental data from Morocco.

5.1 General Setup

The outcome, Y_{ic}, is the school enrollment of child i. The covariate set, $\mathcal{W}_c$, is defined as the intersection pre-treatment characteristics on the PROGRESA and Moroccan questionnaires. The cost-effectiveness benchmark, κ, is based on gains from school enrollment. Montenegro and Patrinos (2014) report a 10% Mincer[4] earnings premium for each additional year of schooling in Morocco. The average market earner in the Benhassine et al. (2015) sample gets \$1578.20 per year. At a 5% discount rate, the net present value of an additional year spent in school is approximately \$1,000.[5] The treatment cost function, $C(W_i, Y_{1i})$, is based on the value of the conditional transfer. In Morocco, for 6-7 year olds, this amounts to the following:

$$60 \text{ MAD per month} \times 12 \times \frac{0.1 \text{ years of ed.}}{100USD} \times \frac{1 \text{ USD}}{8 \text{ MAD}} \times Y_{1i} = 0.09Y_{1i} = \frac{C(w, Y_{1i})}{\kappa}.$$

Correspondingly, for 8-9 year olds the transfer value is 80 MAD per month, which means $0.12Y_{1i} = \frac{C(w,Y_{1i})}{\kappa}$, and for 10+ year olds, the transfer is 100 MAD per month, and so $0.15Y_{1i} = \frac{C(w,Y_{1i})}{\kappa}$. If we define these age specific multipliers as $g(Age_i)$, then the objective for model m reduces to,

$$\pi_m = \max_{\pi} \hat{\mathbb{E}}_m \left\{ \pi(W_i) \hat{\mathbb{E}}_m \left[(1 - g(Age_i))Y_{1i} - Y_{0i} | W_i \right] \right\},$$

generating the decision rule

$$\pi_m(w) = 1 \left\{ \hat{\mathbb{E}}_m \left[(1 - g(Age_i))Y_{1i} - Y_{0i} \middle| w \right] \geq 0 \right\}.$$

This implies evaluating the signs of estimated conditional treatment effects on the adjusted outcome,

$$Y_{ic}^{adj} = T_{ic}(1 - g(\text{Age}_{ic})Y_{ic} + (1 - T_{ic})Y_{ic}.$$

[4]controlling only for potential experience and its square.

[5]Over 40 years of work.

We allow each method all observations and variables from PROGRESA ($D_c = 0$), and observations from TAYSSIR with $D_c = 1, T_{ic} = 0, U_{ic} \leq 0.5$ where we assign $U_{ic} \sim U[0,1]$ independently from all other variables once (i.e., a random 50% split of the TAYSSIR control group). The methods use these data to compute $\pi_m(w)$ as defined above and then for methods l and m, we compute $\widehat{\Delta}_{lm}$ as

$$\hat{\Delta}_{lm} = \frac{\sum_{\{i:D_c=1,T_{ic}=1\}} \frac{T_i}{p_c(W_i)} \left(\pi_l(W_i) - \pi_m(W_i)\right)(1 - g(Age_i))Y_i}{\sum_{\{i:D_c=1,T_{ic}=1\}} \frac{T_i}{p_c(W_i)}} - \frac{\sum_{\{i:D_c=1,T_{ic}=0,U_{ic}>0.5\}} \frac{(1-T_i)}{1-p_c(W_i)} \left(\pi_l(W_i) - \pi_m(W_i)\right) Y_i}{\sum_{\{i:D_c=1,T_{ic}=0,U_{ic}>0.5\}} \frac{(1-T_i)}{1-p_c(W_i)}},$$

where for the TAYSSIR experiment, $p_c(w)$ is known $= p_c$. The estimate for the variance is constructed analogously.

Methods We consider methods that are already available from the current literature and that are straightforward to implement. We highlight assumptions on the joint distribution of random variables for each method. However, we are not concerned with testing these assumptions directly, as in Allcott (2015a), Dehejia et al. (2017). Instead, we list them as part of the specification of each method. We are interested in assessing the relative empirical performance of the methods, all of which we view as likely misspecified (Wolpin, 2007).

The first two methods rely on reduced-form extrapolation of conditional treatment effects, as per, e.g., Hotz et al. (2005) and Dehejia et al. (2016). The reduced form approaches we use include a "low-tech" version that simply takes the age-sex-specific treatment effects from PROGRESA and extrapolates them using the age-sex distribution in the TAYSSIR sample. We also use a "high-tech" approach that applies the generalized random forest estimator for heterogeneous treatment effects, as proposed by Athey et al. (2019).

The next two methods include structural counterfactual predictions. This includes a static semi-parametric structural model for PROGRESA based on Todd and Wolpin (2008) as well as a dynamic, parametric structural model for PROGRESA based on Attanasio et al. (2012).

5.2 Method 1: Extrapolating Age-Sex Conditional Effects

The "low-tech" reduced form approach uses the adjusted outcomes (Y_{ic}^{adj}) from the PROGRESA data to estimate conditional treatment effects in age-sex strata for girls and boys with ages ranging from 6 to 16. These conditional treatment effects are then extrapolated to Morocco, which is justified given the following assumption.

Age-Sex 1. Unconfounded location (Hotz et al. (2005)) given age and sex.

$$\mathbb{E}\left[(1 - g(Age_i))Y_{1i} - Y_{0i}|D_c, Age_i, Sex_i\right] = \mathbb{E}\left[(1 - g(Age_i))Y_{1i} - Y_{0i}|Age_i, Sex_i\right].$$

Table 1 shows the raw treatment effects for the age-sex strata in PROGRESA. Almost all of these are positive, so any treatment assignment based on them will be very liberal. We see exactly this in the second column of Table 3 which compares the effects of treatment assignments based on age-sex extrapolation to simply treating everyone, as a researcher might be inclined to recommend after seeing the large effect of TAYSSIR on enrollment. Without adjusting for cost-effectiveness, age-sex extrapolation recommends treating 97% of the Moroccan sample.

Table 2 lists the enrollment gains by age-sex strata, after adjusting for cost-effectiveness. The cost-effectiveness threshold we've imposed in Morocco is quite stringent: only a few subgroup effects remain positive. The first column of Table 3 shows the implications. Now the age-sex extrapolation only recommends treating 13% of children in Morocco. This results in a statistically significant 3.5 percentage point decrease in enrollment gain (the third row and forth row), but a 6.5 percentage point increase in cost-effectiveness-adjusted enrollment gain (the fifth and sixth row). The increase comes from the fact that age-sex based-extrapolation avoids recommending treatment for younger children whose enrollment is already almost universal. Universal untreated enrollment has to be paid for, which is expensive, and leaves little room for enrollment gain so the cost is not worth it.

5.3 Method 2: Generalized Random Forest-Based Extrapolation

The "high-tech" reduced form approach fits generalized random forest (GRF) models to the adjusted outcomes in the PROGRESA data. At present we include as covariates the child's pre-treatment enrollment status, years of education, literacy status, age, and sex, and then for the child's household, the number of children, whether the head is male, whether it is single-parent, whether the father is alive, whether the mother is alive, whether the child lives with the father, or whether the child lives with the mother. In cases of item-level missing data, we impute a zero for the missing value and then accompany the variable with a separate indicator variable for whether the value is missing. Defining W_i the vector of variables described above, GRF-based extrapolation is justified under the following assumption.

GRF1. Unconfounded location given W_i.

$$\mathbb{E}\left[(1-g(Age_i))Y_{1i}-Y_{0i}|D_c,W_i\right]=\mathbb{E}\left[(1-g(Age_i))Y_{1i}-Y_{0i}|W_i\right].$$

Using these variables, we fit a generalized random forest (GRF) using the algorithm written by Athey et al. (2019). We set the number of trees to 2000 with a minimal leaf size of 2 units. We use 50% splits of the data to build trees on one random split of the data and then select the error-minimizing tree pruning by evaluating predictions with different candidate pruning levels on the data from the other split. These settings do not depart very much from the defaults set by Athey et al. (2019).

There are two approaches to using GRFs to characterize effect heterogeneity. The first is to fit GRFs to the treated and control outcomes separately, and then combine those fits to construct estimates of unit-level treatment effects. This approach targets loss on the level of treated and control outcomes, and then indirectly targets effect heterogeneity. The second is to fit a GRF with a loss function that is specifically targeted to effect heterogeneity. In simulations and various empirical tests on the PROGRESA sample, we found that the first approach was substantially better for our application. This was surprising, but the reason seemed to be that the approach targeting effect heterogeneity directly tended to regularize too heavily, and therefore did not discriminate strongly enough between classes of units with large or small (or even negative) effects. As such, our analysis applies the method that models the treated and control outcomes separately.

Tables 4 and 5 show variable importance summaries provided by Athey et al. (2019) for random forests fitted to enrollment without and with cost-effectiveness adjustment, respectively. The summaries essentially count instances where a variable is used in a split, where instances at the top end of the tree get higher weight than at the lower end. Age, baseline enrollment, and number of years of education completed emerge as most important but several other variables matter as well, including parental education and presence.

Tables 6 and 7 display treatment effects on enrollment, aggregated by age-sex strata to allow for comparisons to the "low-tech" reduced form approach. The GRF detects a great deal of treatment effect heterogeneity within strata, with positive and negative effects for all strata. Table 10 show the share in each stratum recommended for treatment, which increases quickly with age.

Tables 8 and 9 gives treatment effect predictions within strata, after adjusting the treated outcome for cost-effectiveness. The average treatment effect is predicted to be negative in most strata, with the exception of those including children aged 13 and up. Table 11 shows the treatment recommendations. Despite negative average effects within strata, the treatment rates are still quite high.

Table 12 evaluates GRF's treatment recommendations relative to simple age-sex-based extrapolation from PROGRESA. The results are perhaps surprising. GRF treats more individuals, resulting in 1 percentage point more enrollment gain. But the cost outstrips the gain and the adjusted enrollment gain is negative and statistically significant.

We believe this disappointing performance of GRF versus a much simpler method is due to a kind of "contextual overfitting". While the GRF's regularization guards against within-context overfitting, the dimensions of heterogeneity generating a large share of positive treatment effects even after adjusting for cost-effectiveness may represent quirks of PROGRESA's implementation in Mexico. It remains to be seen whether this problem persists when we add reference contexts and context characteristics to our exercise.

5.4 Method 3: Semi-Parametric Structural Approach

5.4.1 Todd and Wolpin (2008)'s Non-Parametric Structural Model

The third method takes as a starting point the non-parametric structural (NPS) model of school attendance proposed in Todd and Wolpin (2008). For parsimony in the non-parametric step of our own approach (described in the next section) we use the simplest version of the NPS model, which considers the enrollment decision for each child independently. In this version of the model, households solve the following static utility maximization problem with utility depending on household consumption c, the child's enrollment status y.

$$\max_{y \in \{0,1\}} \; U(c, y; w, \epsilon)$$
$$\text{subject to } c = n + e(1 - y). \tag{1}$$

Abusing notation slightly, W_i is here redefined as $W_i \setminus E_i, N_i$ where E_i represents the child's wage offer, N_i household income for child i excluding i's own earnings. The redefined W_i and V_c are observed shifters in preference for child schooling and ϵ_i is an unobserved shifter.

Optimal school attendance is given by

$$y_0 = \phi(n, e; w, \epsilon) = 1\{U(n, 1; w, \epsilon) > U(n + e, 0; w, \epsilon)\}.$$

Now modify the budget constraint by introducing the treatment, a grant g paid only when the child attends school:

$$\begin{aligned} c &= n + e(1 - y) + gy \\ &= (n + g) + (e - g)(1 - y). \end{aligned}$$

It is easy to see that y_1, optimal school attendance with the grant program in place, can be obtained by plugging a modified version of non-child income $(n + g)$ and modified child wage offer $(e - g)$ into the same $\phi(\cdot)$ function as before:

$$y_1 = \phi(n + g, e - g; w, \epsilon).$$

In addition to this structure, NPS imposes a key assumption on the distribution of random variables:

NPS1. Exogeneity of non-child income and child wage offers: $\epsilon_i \perp\!\!\!\perp N_i, E_i | W_i$.

The NPS CATE for children with characteristics w, n, e is identified as

$$\begin{aligned} &\mathbb{E}\left[Y_{1i}|N_i = n, E_i = e, W_i = w\right] - \mathbb{E}\left[Y_{0i}|N_i = n, E_i = e, W_i = w\right] \\ &= \mathbb{E}\left[Y_{0i}|N_i = n + g, E_i = e - g, W_i = w\right] - \mathbb{E}\left[Y_{0i}|N_i = n, E_i = e, W_i = w\right] \\ &= \mathbb{E}\left[Y_i|T_i = 0, N_i = n + g, E_i = e - g, W_i = w\right] - \mathbb{E}\left[Y_i|T_i = 0, N_i = n, E_i = e, W_i = w\right] \end{aligned} \tag{2}$$

Crucially, (2) is identified in the data provided to predictors for context c.

5.4.2 Our Implementation: the Semi-Parametric Structural (SPS) Approach

As in Todd and Wolpin (2008), we do not observe wages for most children in our contexts. In addition, in some of our contexts non-child income is not directly observable since households are smallholder farmers. We therefore make some modifications to the NPS model.

Each child's wage offer is given by exponentiating the conditional expectation of her earnings given her age, gender, industry of work[6], and locality of residence.[7]

SPS1. Conditional expectation wage offers:

$$\begin{aligned} E_i &= \exp(\mathbb{E}\left[\log(E_i)|Age_i, Sex_i, Industry_i, Locality_i\right]) \\ &= e_0(Age_i, Sex_i, Industry_i, Locality_i). \end{aligned}$$

[6]If observed, otherwise we set this to the most common industry for child workers.

[7]Todd and Wolpin (2008), in contrast set children's wage offers to the village-level wage for an agricultural worker but this variable is not available in many of our contexts, and in fact in only half of the PROGRESA evaluation villages they consider.

This is very close to how we handle wages in the dynamic parametric structural model (DPS) described in 5.5, except that in the DPS model agents optimize with respect to (1) the sample mean function $\hat{\mathbb{E}}\left[E_i|Age_i, Sex_i, Industry_i, Locality_i\right]$ rather than (2) the population expectation function $\mathbb{E}\left[E_i|Age_i, Sex_i, Industry_i, Locality_i\right]$ (use of (1) in DPS exactly follows AMS). We handle non-child income in the same way: by computing the sum of the expected earnings of all the members in i's household, excluding i.

SPS2. Conditional expectation non-child income:

$$\begin{aligned} N_i &= \sum_{j \in Household_i, j \neq i} \exp(\mathbb{E}\left[\log(E_j)|Age_j, Sex_j, Industry_j, Locality_i\right]) \\ &= n_0(Household_i). \end{aligned}$$

Additionally, assume E_i is missing at random, effectively following Todd and Wolpin (2008).

With industry and locality indicators, the set of conditioning variables is high-dimensional. We therefore estimate the conditional expectation function by LASSO, which is justified under the following assumption.

SPS3. Approximately sparse linear representation of expected wages.

$$\begin{aligned} &\mathbb{E}\left[\log(E_i)|Age_i, Sex_i, Industry_i, Locality_i\right] \\ &\approx \eta_0 + \eta_1 Age_i + \eta_2 (Age_i - 21)_+ + \eta_3 male_i + \zeta_{industry} + \lambda_{province} + \xi_{locality} + \nu_i \end{aligned}$$

where approximation is in the sense of Belloni et al. (2012). The notation $(\cdot)_+$ indicates the positive part of the expression in parentheses. $\zeta_{industry}, \lambda_{province}$, and $\xi_{locality}$ are fixed effects. Province is the top level subnational geographic unit (like a US state), which is defined for each context following IPUMS-International. Locality is the smallest geographic level available in each dataset. We do not subject the linear spline in age to the LASSO penalty because the substantial positive gradient of wage in age for youths is a key driver of the opportunity cost of enrollment. Similarly, we always include the industries employing the majority of children in each context.[8]

We select the LASSO penalty term by 5-fold least squares cross-validation. Finally, we replace assumption NPS1 with

SPS4. $\mathbb{E}\left[\varepsilon_i | e_0(Age_i, Sex_i, Industry_i, Locality_i), n_0(Household_i), Sex_i\right] = 0$ where ε_i is defined in the conditional expectation equation

$$Y_i = m_0(e_0(Age_i, Sex_i, Industry_i, Locality_i), n_0(Household_i), Sex_i) + \varepsilon_i. \quad (3)$$

Our approach is the following. We estimate $e_0(Age_i, Sex_i, Industry_i, Locality_i)$ in a first stage and use the resulting $\hat{e}$ in place of e_0 in n_0. We then estimate m_0, our analog of $\mathbb{E}\left[Y_i | T_i = 0, N_i = n, E_i = e, W_i = w\right]$ from equation (2), with $\hat{e}$ and $\hat{n}$ in place of e_0 and n_0. $\hat{e}$ and $\hat{n}$ are generated regressors in the sense of Mammen et al. (2012) who show that the second stage estimate is consistent for m_0. Our second stage implementation is by mixed datatype kernel regression with bandwidth selected by cross-validation using the `np` package in R (Hayfield and Racine (2008)). We then use the second stage estimate to generate the predicted SPS CATE just as in equation (2).

[8]We exclude education following AMS. In practice, and echoing AMS's findings, education has little effect on earnings in our rural contexts. Education is never selected by LASSO and its inclusion has almost no impact on other variables coefficients or the selected penalty term.

5.4.3 Results

Figures 1, 2, and 3 provide some intuition for how SPS works in practice. Figure 1 shows the bivariate density of child wage offers in the held out portion of the control group on the y-axis and non-child income on the x-axis. Figure 2 shows how the model predicts the effect of the TAYSSIR grant. Effective child wage offers decrease, shifting the density down, and non-child income is increased slightly, moving density slightly right.

The impact of this shift depends on the non-parametric regression of enrollment as a function of child wage offer and non-child income in the portion of the control group (non-holdout) available to predictors, which is depicted in Figure 3. The plot shows that non-child income has little association with enrollment conditional on a child's wage offer. child wage offers do matter, however, particularly in the region where probability mass is being moved.

Table 13 shows that the SPS approach tends to overstate the magnitude of treatment effects.[9] Relative to age-sex based extrapolation from PROGRESA, SPS is only slightly better in cost-adjusted enrollment gain terms than making all children in Morocco eligible for TAYSSIR grants. We think this is because the non-parametric second step is simply too volatile, despite having been regularized by cross validation based bandwidth selection.

5.5 Method 4. Dynamic Parametric Structural

Model Our Dynamic Parametric Structural (DPS) model is largely based on Attanasio et al. (2012), with a few modifications made to fit data availability and improve in-sample fit in our contexts, following standard practice (see e.g. Wolpin (2013)).

[9]This turns out to be true for both positive and negative treatment effects, with specifics to be included in later drafts.

The main dynamic features of the model are (1) a finite horizon - children can only be enrolled until age 17 and so can only accumulate subsidies up to this point and (2) persistence of education choices - the flow utility of enrollment is affected by the number of years the child is behind her age-appropriate grade level. Uniquely among our methods, the DPS model allows modeling of the entire subsidy schedule by age.

Flow utility The in-period utility for child i[10] at age a in school and work are u_{ia}^{S} and u_{ia}^{W}, respectively:

$$u_{ia}^{S} = \gamma\delta g_{ia} + \mu_i + \psi' z_{ia} + b' \cdot \text{yrs\_behind}_{ia} + 1\left(p_{ia}=1\right)\beta^p x_{ia}^p + 1\left(s_{ia}=1\right)\beta^s x_{ia}^s + \varepsilon_{ia}$$
$$u_{ia}^{W} = \delta w_{ia}$$

g_{ia} represents the grant i is entitled to given her completed years of schooling and other characteristics (for example gender). μ_i is a child-specific shifter to the preference for enrollment, drawn from a discrete distribution with K points of support. We will refer to each point of support as an unobserved child "type". In practice, we estimate the model with three types. z_{ia} includes other observed covariates affecting preference for enrollment. Specifically it includes a and a dummy variable for whether i's father received any formal schooling. yrs_behind_{ia} is a three-element vector of dummy variables with indicators for being behind grade level by 1 year, 2 years, or ≥ 3 years. p_{ia} is a dummy variable measuring whether i's years of schooling at age a make her eligible for primary school and x_{ia}^p measures distance to school, proxying for cost of attendance. s_{ia} is a dummy variable equal to one if i's years of schooling make her eligible for secondary school and x_{ia}^s is a constant. e_{ia} is i's wage offer. ε_{ia} is an IID idiosyncratic shock to the utility of i's attending school at age a, which follows the logistic distribution.[11]

[10] Again, schooling decisions are made by the household for each child independently.
[11] Note that this ε_i has no relation to the ε_i from equation (3).

We allow some of the coefficients to vary by i's sex. In particular, β^s and the component of ψ multiplying age vary by gender. The unobserved type probability also depends on the sex of the child.

Wage offer Log wage offers are computed according to the same sparse linear regression representation equation described in assumption SPS3 of the SPS model. The only difference is that in the SPS model predicted values from the LASSO regression are treated as estimates of the true wage offer. In the DPS model, agents are assumed to use the same predicted values as the econometrician.

Terminal value The value of having accumulated ed_{i18} years of school in the terminal period (when the child is 18) is given below.

$$V\left(ed_{i,18}\right)=\frac{\alpha_1}{1+\exp\left(-\alpha_2 ed_{i,18}\right)}+\alpha_3\cdot 1\{ed_{i,18}\geq sec\}.$$

α_3 measures the additional value of having completed secondary school, measured by $ed_{i,18}$ being greater than the last year of secondary school, sec.

Value functions The value of choosing to have i attend school after having completed ed_{ia} years of education by age a is:

$$\begin{aligned}V_{ia}^{S}\left(ed_{ia}\right)=u_{ia}^{S}+\beta\{p_{a}^{S}\left(ed_{ia}+1\right)\mathbb{E}\max\left[V_{i,a+1}^{S}\left(ed_{ia}+1\right),V_{i,a+1}^{\mathrm{W}}\left(ed_{ia}+1\right)\right]\\+\left(1-p_{a}^{S}\left(ed_{it}+1\right)\right)\mathbb{E}\max\left[V_{i,a+1}^{S}\left(ed_{ia}\right),V_{i,a+1}^{W}\left(ed_{it}\right)\right]\}.\quad(4)\end{aligned}$$

We implicitly condition on covariates W_{ia} redefined as $W_{ia}\setminus ed_{ia}$ and μ_i. $p_a^s(ed)$ is the probability of successfully passing grade ed at age a conditional on enrolling. We estimate it non-parametrically, outside the model (like the wage offer function). If i successfully passes the grade, she expects to receive the maximum of the value of enrolling or choosing to work in the next year with her education equal to $ed_{ia}+1$. The expectation is taken over possible realizations of the $\varepsilon_{i,a+1}$ shock. If i does not pass, she expects to get the maximum of the value of enrolling/working being one year older and with education still equal to ed_{ia}. The term in braces in equation 4 is thus the next-period expected value. From the point of view of this period, the expected value is discounted by β which we set equal to 0.95, following AMS. We add the flow utility of being enrolled to complete the definition of the value function when enrolling. The value of having i work this period is simpler since ed_{ia} stays fixed:

$$V_{it}^{W}\left(ed_{ia}\right)=u_{it}^{W}+\beta\mathbb{E}\max\left\{V_{i,a+1}^{S}\left(ed_{ia}\right),V_{i,a+1}^{W}\left(ed_{ia}\right)\right\}.$$

Given a set of parameters, we solve for the outputs each value function (enroll, work) for of all possible combinations of age and years of education completed by backward induction, beginning by calculating the terminal value for each set of candidate parameters.[12]

Likelihood With the value function in hand and the logistic error distribution assumption, it is straightforward to compute the likelihood of each child's being enrolled given observed characteristics W_i and unobserved type μ_i: $\Pr(Y_i=1|W_i,\mu_i=\mu_k)$.[13]

[12]Note that since the only error term ε_{it} follows an IID logistic distribution, then $\mathbb{E}\max\left\{V_{it+1}^{s}\left(ed_{it}\right),V_{it+1}^{\mathrm{w}}\left(ed_{it}\right)\right\}$ has a closed form (see Keane et al. (2011)). Our closed form is slightly different from theirs because they use two Type 1 Extreme Value random variables instead of one logistic draw. We simply subtract ε_2 in both equations in their function to derive our closed form:

$$\mathbb{E}\max\left\{V_{ia}^{S}\left(ed_{ia}\right),V_{ia}^{W}\left(ed_{ia}\right)\right\}=\rho\{\log[(\exp(V_{ia}^{S})/\rho+\exp(V_{it}^{W})/\rho]\}.$$

We normalize the scale ρ of the error term to 1 in estimation.

[13]We drop the a subscript because the model is estimated on a single cross-section so age does not vary with i.

The full likelihood associated with $Y_i = 1$ is given by:[14]

$$\Pr(Y_i = 1|W_i) = \sum_{k=1}^{K} \frac{1}{1 + \exp(V_{it}^{W}(ed_i|\mu_k, W_i) - V_{it}^{S}(ed_i|\mu_k, W_i))} \Pr(\mu_i = \mu_k|W_i, \mu_k). \tag{5}$$

Since the utility shock ε_i is IID across time and individuals, we could in principle derive the conditional type probability as a function of the history of characteristics and education decisions of i starting at the age when i's enrollment was first considered (min_a): $(W_{i,a-1}, \ldots, W_{i,min\_a})$ and $(Y_{i,a-1}, \ldots, Y_{i,min\_a})$ respectively:

$$\begin{aligned}&\Pr(\mu_i = \mu_k|(Y_{i,a-1}, \ldots, Y_{i,min\_a}), (W_{i,a-1}, \ldots, W_{i,min\_a}))\\ &= \frac{\Pr(Y_{ia}|W_{ia}, \mu_i = \mu_k) \cdots \Pr(Y_{i,min\_a}|W_{i,\min\_a}, \mu_i = \mu_k) \Pr(\mu_k)}{\sum_{k=1}^{K} \Pr(Y_{ia}|W_{ia}, \mu_i = \mu_k) \cdots \Pr(Y_{i,min\_a}|W_{i,\min\_a}, \mu_i = \mu_k) \Pr(\mu_k)}.\end{aligned} \tag{6}$$

Actually estimating (6) is infeasible, however, since it requires knowledge of the full histories $(W_{i,a-1}, \ldots, W_{i,min\_a})$ and $(Y_{i,a-1}, \ldots, Y_{i,min\_a})$[15] and, furthermore, would be very high-dimensional.

Following Todd and Wolpin (2006), we instead use a multinomial logit approximation:

$$\begin{aligned}&\Pr(\mu_i = \mu_k|(Y_{i,a-1}, \ldots, Y_{i,min\_a}), (W_{i,a-1}, \ldots, W_{i,min\_a}))\\ &\approx \Pr(\mu_i = \mu_k|a, ed_{ia}, gender, father\_ed)\\ &\quad \approx \frac{\exp(\beta_k^{\mu} \cdot (1, a, ed_{ia}, gender, father\_ed))}{1 + \sum_k \exp(\beta_k^{\mu} \cdot (1, a, ed_{ia}, gender, father\_ed)}, k \in 1 \ldots, K-1\end{aligned}$$

Age proxies for the length of the history, ed_{ia} for $(Y_{i,a-1}, \ldots, Y_{i,min\_a})$ and *gender* and $father\_ed$ for $(W_{i,a-1}, \ldots, W_{i,min\_a})$ (specifically W_{i,min_a}). Importantly, ed_{ia} is

[14]If $Y_i = 0$ the likelihood contribution is 1 - (5).
[15]These would be included in W_i, but unfortunately we do not have them.

excluded from flow utility so there is independent variation to identify the conditional type probabilities. We estimate the parameters described above, along with the support points $(\mu_1, \cdots, \mu_K)$ by maximum likelihood.

5.5.1 Results

Figures 4 and 5 show that the model fits well in the portion of the Moroccan control group made available to predictors. Figure 4 shows the fit to enrollment rates by age. The size of each point on the graph representing the sample size in that age-sex stratum. The model captures delayed entry into school, near-universal enrollment at young ages, and the sharp drop in enrollment for teenagers. Figure 5 illustrates the fit by number of years of education completed. It captures the drop in enrollment at the transition to secondary school (year 7 in Morocco).

Table 14 shows the results from using the DPS model to predict the effect of the TAYSSIR program. For all but a tiny fraction of the Moroccan holdout sample, predicted enrollment gains - while reasonable - are too small to exceed the cost-effectiveness threshold. We show this visually in Figure 6, plotting the CDF of predicted treatment effects due to observables. A key point from Table 14 is how close TAYSSIR is to being non-cost-effective for *any* child. Age-sex based extrapolation from PROGRESA only provides a small, statistically insignificant increase in welfare relative to DPS's no-treatment recommendation.

6 Conclusion

We develop a decision-based approach to comparing the relative performance of methods for generating counterfactual predictions that are then used to make policy recommendations. We consider a social planner who is operating in a target context and is seeking recommendations on what policy to choose from a set of feasible options.

The richness of the space of policy options determines the nature of the recommendations being sought — e.g., whether a simple up-or-down recommendation to treat everyone or no one, or a more refined recommendation about who should be treated and who not. Recommendations could be based on econometric estimates, whether reduced form or structural, or expert opinions. Our leading application is one where the planner maximizes a linear welfare objective in assigning treatments on the basis of available covariate information. In this case, the success of a method for generating recommendations depends on how accurately it can predict conditional treatment effects in the target context.

We define a welfare contrast to use for conducting an *ex post* analysis of how well different methods performed with respect to the planner's goals. We estimate this welfare contrast by using experimental data that reveals how a treatment affects the outcome distribution in the target population. The welfare contrast is straightforward to compute, and it allows us to judge whether one method outperforms another in a manner that is statistically significant.

We provide an empirical illustration that considers a planner seeking a recommendation on how to implement program using conditional cash transfers (CCTs) to boost school enrollment in Morocco. The data available for generating recommendations include a randomized evaluation of CCTs in Mexico as well as data from Moroccan households under the status quo *ex ante*, in which no CCTs have been applied. We generate recommendations from reduced form methods and structural models. We then perform an *ex post* evaluation of these methods using data from a randomized evaluation of CCTs in Morocco. We view this toy example as helping build intuition for how to specify methods to evaluate in a full-featured empirical portion of the paper including the contexts from Banerjee et al. (2017) which we will pre-specify.

We see this exercise as making three contributions. First, as our application attempts to show, it provides a clear framework to assess internal validity versus external validity trade-offs. In particular, our application allows us to assess how robust and internally valid estimates from external contexts fare relative to within-context estimates that may be biased due to model misspecification (Pritchett and Sandefur, 2013). Second, it provides a principled basis for assessing the performance of different methods by tying the assessment to welfare considerations. This is important, because different objective functions can imply different rank orderings of methods. Our approach thus forces one to first consider the welfare objective so as to be clear about the relevant objective. Third, we show that each experiment or observational study may contain much more decision-relevant information than would be contained in a single treatment effect estimate.

We are undertaking a number of extensions to what we have done here. Model selection or model averaging approaches based on our welfare criteria may lead to better predictions. We also plan to work with evidence bases that include more external contexts. In doing so, we would want to account for site selection, as per Allcott (2015b) and Gechter and Meager (2018).

A Proofs

Proof of Proposition 1. By the weak law of large numbers, Slutsky's theorem, and conditions C2 and C3, $\hat{\Delta}^g_{lm}$ has the same limit as

$$\begin{aligned}\tilde{\Delta}^g_{lm} =& \frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{p(W_i)}\left(\pi_l(W_i)-\pi_m(W_i)\right)g(Y_i^P(1)) \\ &-\frac{1}{N}\sum_{i=1}^{N}\frac{1-T_i}{1-p(W_i)}\left(\pi_l(W_i)-\pi_m(W_i)\right)g(Y_i^P(0)).\end{aligned}$$

Take the first term on the right-hand side. By the weak law of large numbers, iterated expectations, and condition C1,

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{p(W_i)}\left(\pi_l(W_i)-\pi_m(W_i)\right)g(Y_i^P(1)) \\ &\qquad \xrightarrow{p} \mathbb{E}\left[\mathbb{E}\left[T_i|W\right]\frac{1}{p(W)}\mathbb{E}\left[\left(\pi_l(W_i)-\pi_m(W_i)\right)g(Y_i^P(1))|W\right]\right] \\ &\qquad = \mathbb{E}\left[\left(\pi_l(W_i)-\pi_m(W_i)\right)g(Y_i^P(1))\right],\end{aligned}$$

and similar for the second term. Thus as $N \to \infty$, $\mathbb{E}\,[\hat{\Delta}^g_{lm} - \Delta^g_{lm}] \xrightarrow{p} 0$. Having established that $\hat{\Delta}^g_{lm}$ is asymptotically unbiased for Δ^g_{lm}, inference follows from the usual generalized method of moments results (Newey and McFadden, 1994; Lunceford and Davidian, 2004). To see this, first note that $\hat{\Delta}^g_{lm} = \hat{\delta}_1 - \hat{\delta}_0$ for $(\hat{\delta}_1, \hat{\delta}_0)$ that solve the score equations

$$\sum_{i=1}^{N}\psi_1(\hat{\delta}_1) = 0 \text{ and } \sum_{i=1}^{N}\psi_0(\hat{\delta}_0) = 0,$$

$$\text{where } \psi_1(\hat{\delta}_1) = \frac{T_i\left[(\pi_l(W_i)-\pi_m(W_i))g(Y_i)-\hat{\delta}_1\right]}{p(W_i)},$$

$$\text{and } \psi_0(\hat{\delta}_0) = \frac{(1-T_i)\left[(\pi_l(W_i)-\pi_m(W_i))g(Y_i)-\hat{\delta}_0\right]}{1-p(W_i)}.$$

Then, given random sampling, bounded first and second moments, and conditions C1-C2, we have

$$\frac{\sqrt{N}\left(\hat{\Delta}^g_{lm} - \Delta^g_{lm}\right)}{\sqrt{V_{\Delta^g_{lm}}}} \xrightarrow{d} N(0,1), \tag{7}$$

where

$$\begin{aligned} V_{\Delta^g_{lm}} &= \mathbb{E}\left[\psi_1(\delta_1)^2 + \psi_0(\delta_0)^2\right] \\ &= \mathbb{E}\left\{\frac{\left[(\pi_l(W_i) - \pi_m(W_i))g(Y_i^P(1)) - \delta_1\right]^2}{p(W_i)} + \frac{\left[(\pi_l(W_i) - \pi_m(W_i))g(Y_i^P(0)) - \delta_0\right]^2}{1 - p(W_i)}\right\}, \end{aligned}$$

with $\delta_t = \mathbb{E}\left[(\pi_l(W_i) - \pi_m(W_i))g(Y_i^P(t))\right]$. Then, by the same conditions for which $\hat{\Delta}^g_{lm}$ is consistent for Δ^g_{lm}, $\hat{V}_{\Delta^g_{lm}}$ is consistent for $V_{\Delta^g_{lm}}$. □

References

Allcott, H. (2015a). Site Selection Bias in Program Evaluation. *Quarterly Journal of Economics 130*(3), 1117–1165.

Allcott, H. (2015b). Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics 130*(3), 1117–1165.

Andrews, I. and E. Oster (2017). WEIGHTING FOR EXTERNAL VALIDITY.

Aronow, P. M. and C. Samii (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics 11*(4), 1912–1947.

Arrow, K. J. (1972). The Value of and Demand for Information. In C. McGuire and R. Radner (Eds.), *Decision and Organization. A Volume in Honor of Jacob Marschak* (Amsterdam ed.)., Chapter 6, pp. 131–140. North-Holland.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Athey, S. and S. Wager (2017). Efficient Policy Learning.

Attanasio, O., C. Meghir, and A. Santiago (2012). Education Choices in Mexico: Using a Structural Model and a Randomised Experiment to Evaluate PROGRESA. *Review of Economic Studies 79*(1), 37–66.

Banerjee, A., S. Chassang, and E. Snowberg (2016). Decision Theoretic Approaches to Experiment Design and External Validity. In *Handbook of Field Experiments, forthcoming.*

Banerjee, A., R. Hanna, G. Kreindler, and B. A. Olken (2017). Debunking the Stereotype of the Lazy Welfare Recipient: Evidence from Cash Transfer Programs Worldwide. *The World Bank Research Observer 32*(2), 155–184.

Behrman, J. R., P. Sengupta, and P. E. Todd (2005, oct). Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico. *Economic Development and Cultural Change 54*(1), 237–275.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. *Econometrica 80*(6), 2369–2429.

Benhassine, N., F. Devoto, E. Duflo, P. Dupas, and V. Pouliquen (2015). Turning a shove into a nudge? A "labeled cash transfer" for education. *American Economic Journal: Economic Policy 7*(3), 1–48.

Bhattacharya, D. and P. Dupas (2012). Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics 167*(1), 168–196.

Cox, D. R. (1958). *Planning of Experiments*. Wiley.

De Janvry, A. and E. Sadoulet (2006, mar). Making conditional cash transfer programs more efficient: designing for maximum effect of the conditionality. *The World Bank Economic Review 20*(1), 1.

Dehejia, R., C. Pop-Eleches, and C. Samii (2016). From local to global: External validity in a fertility natural experiment.

Dehejia, R., C. Pop-Eleches, and C. Samii (2017). From Local to Global: External Validity in a Fertility Natural Experiment. *NBER Working Paper 21459*.

Dehejia, R. H. (2003, jan). Was There a Riverside Miracle? A Hierarchical Framework for Evaluating Programs With Grouped Data. *Journal of Business and Economic Statistics 21*(1), 1–11.

Dehejia, R. H. (2005, apr). Program evaluation as a decision problem. *Journal of Econometrics 125*(1-2), 141–173.

Dehejia, R. H. (2008). When is ATE enough? Risk aversion and inequality aversion in evaluating training programs. *Advances in Econometrics 21*, 263–287.

DellaVigna, S. and D. Pope (2017). Predicting Experimental Results: Who Knows What? *Journal of Political Economy, forthcoming*.

Dhaliwal, I., E. Duflo, and R. Glennerster (2011). Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education.

Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. *Journal of Business & Economic Statistics 33*(1), 1–1.

Diebold, F. X. and R. S. Mariano (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics 13*(3), 253.

Duflo, E., R. Hanna, and S. Ryan (2012). Incentives work: Getting teachers to come to school. *The American Economic Review 102*(4), 1241–1278.

Garber, A. M. and C. E. Phelps (1997). Economic foundations of cost-effectiveness analysis. *Journal of Health Economics 16*(1), 1–31.

Garcia, S. and J. Saavedra (2017). Educational impacts and cost-effectiveness of conditional cash transfer programs in developing countries: A meta-analysis. NBER Working Paper No. 23594.

Gechter, M. (2016). Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India. *Working Paper*.

Gechter, M. and R. Meager (2018). Incorporating Experimental and Observational Studies in Meta-Analysis. *Working Paper*.

Granger, C. and M. Machina (2006). Forecasting and decision theory. *Handbook of economic forecasting 1*(05), 81–98.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The Model Confidence Set. *Econometrica 79*(2), 453–497.

Hayfield, T. and J. S. Racine (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software 27*(5).

Hirano, K., G. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Hirano, K. and J. R. Porter (2009). Asymptotics for Statistical Treatment Rules. *Econometrica 77*(5), 1683–1701.

Hirano, K. and J. R. Porter (2019). Statistical Decision Rules in Econometrics. In *Handbook of Econometrics, Volume 7*.

Hotz, V. J., G. Imbens, and J. Mortimer (2005). Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations. *Journal of Econometrics 125*, 241–270.

Hudgens, M. G. and M. E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association 103*(482), 832–842.

Imai, K. and A. Strauss (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis 19*(1), 1–19.

Imbens, G. W. G. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 5–86.

Keane, M. P., P. E. Todd, and K. I. Wolpin (2011). The Structural Estimation of Behavioral Models: Discrete Choice Dynamic Programming Methods and Applications. *Handbook of Labor Economics, Volume 4 4*(2), 331–461.

Keane, M. P. and K. I. Wolpin (2007, dec). Exploring the usefulness of a nonrandom holdout sample for model validation: welfare effects on female behavior. *International Economic Review 48*(4), 1351–1378.

Kitagawa, T. and A. Tetenov (2017). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica, Forthcoming*.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine 23*(19), 2937–2960.

Mammen, E., C. Rothe, and M. Schienle (2012). Nonparametric regression with nonparametrically generated covariates. *Annals of Statistics 40*(2), 1132–1170.

Manski, C. F. (1996). Learning about Treatment Effects from Experiments with Random Assignment of Treatments. *The Journal of Human Resources 31*(4), 709.

Manski, C. F. (2004, jul). Statistical Treatment Rules for Heterogeneous Populations. *Econometrica 72*(4), 1221–1246.

Manski, C. F. (2011). Choosing Treatment Policies Under Ambiguity. *Annual Review of Economics 3*(1), 25–49.

Marschak, J. and K. Miyasawa (1968). Economic Comparability of Information Systems. *Internal Economic Review 9*(2), 137–174.

Meager, R. (2016). Understanding the Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of 7 Randomised Experiments.

Montenegro, C. E. and H. A. Patrinos (2014). Comparable Estimates of Returns to Schooling Around the World. *World Bank Policy Research Working Paper 7020*.

Newey, W. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

Parker, S. W. and T. Vogl (2018). Do conditional cash transfers improve economic outcomes in the next generation? Evidence from mexico.

Pesaran, M. and S. Skouras (2002). Decision-Based Methods for Forecast Evaluation. In M. Clements and D. Hendry (Eds.), *A Companion to Economic Forecasting*. Oxford: Blackwell.

Pritchett, L. and J. Sandefur (2013). Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix. *Journal of Globalization and Development 4*(2), 161–197.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association 75*(371), 591–593.

Schorfheide, F. and K. I. Wolpin (2012, may). On the Use of Holdout Samples for Model Selection. *American Economic Review 102*(3), 477–481.

Schorfheide, F. and K. I. Wolpin (2016). To Hold Out or Not To Hold Out. *Research in Economics, forthcoming*.

Schultz, T. P. (2004, jun). School subsidies for the poor: evaluating the Mexican Progresa poverty program. *Journal of Development Economics 74*(1), 199–250.

Tetenov, A. (2012). Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics 166*(1), 157–165.

Todd, P. E. and K. I. Wolpin (2006). Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *The American Economic Review 96*(5), 1384–1417.

Todd, P. E. and K. I. Wolpin (2008). Ex ante evaluation of social programs. *Annales d'Economie et de Statistique* (91/92), 263–291.

Vivalt, E. (2016). How Much Can We Generalize From Impact Evaluations? *Mimeo*.

Vuong, Q. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica 57*(2), 307–333.

West, K. (1996). Asymptotic inference about predictive ability. *Econometrica 64*(5), 1067–1084.

White, H. (2000). A reality check for data snooping. *Econometrica 68*(5), 1097–1126.

Wolpin, K. (2007, may). Ex Ante Policy Evaluation, Structural Estimation, and Model Selection. *The American economic review 97*(2), 48–52.

Wolpin, K. I. (2013). *The Limits of Inference Without Theory.* The MIT Press.

Figure 1: SPS model: original density for boys

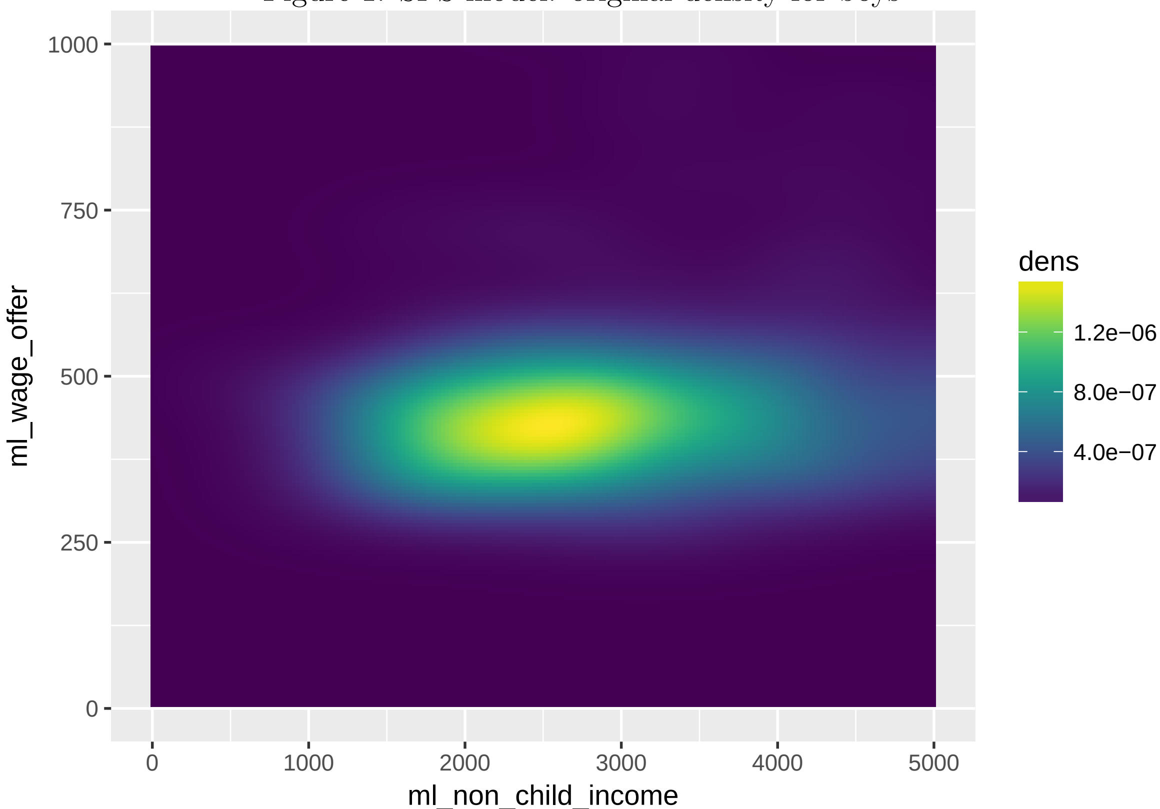


Figure 2: SPS model: counterfactual density for boys

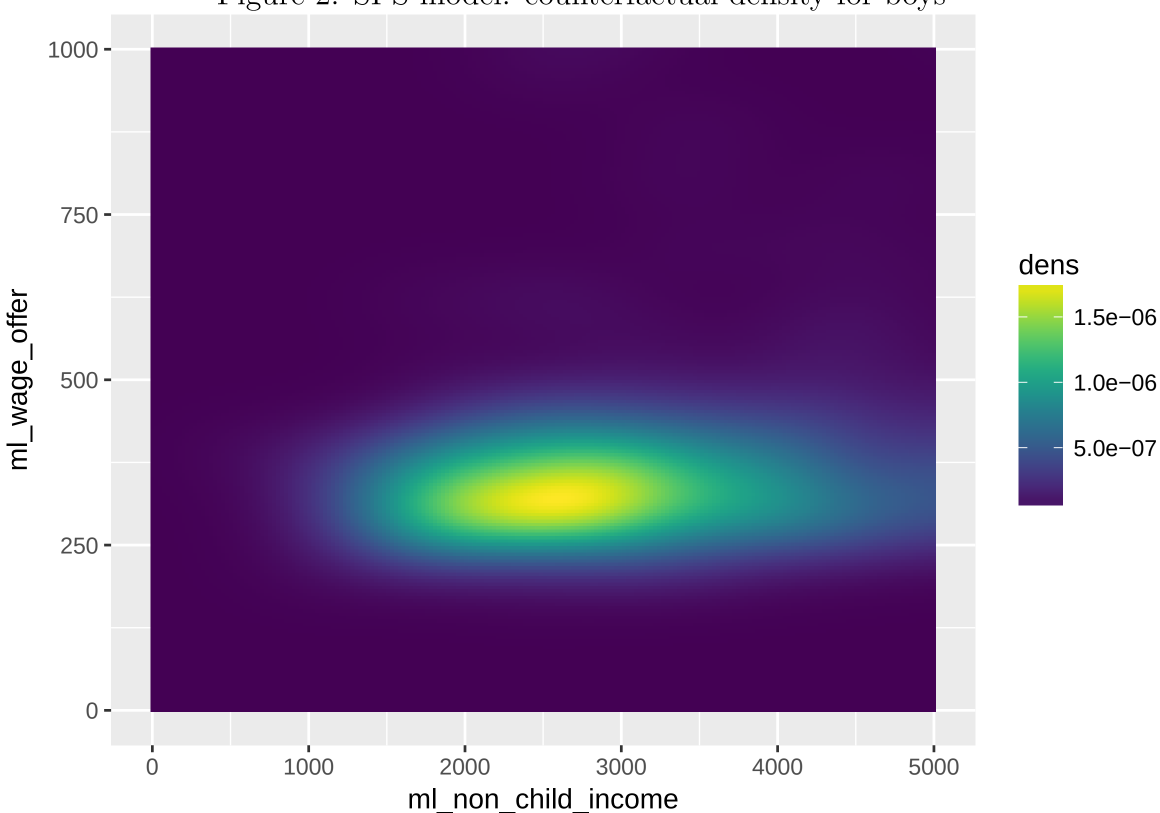


Figure 3: SPS: regression function

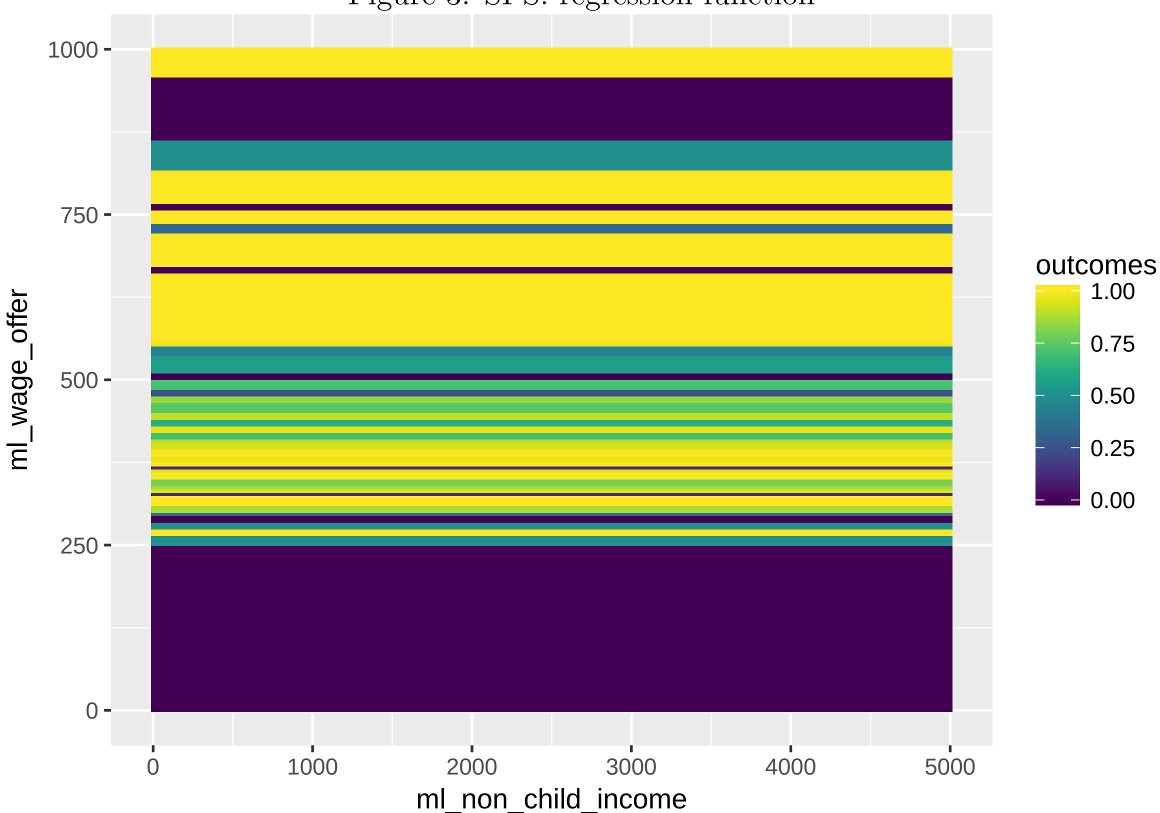


Figure 4: DPS: in-sample fit by age

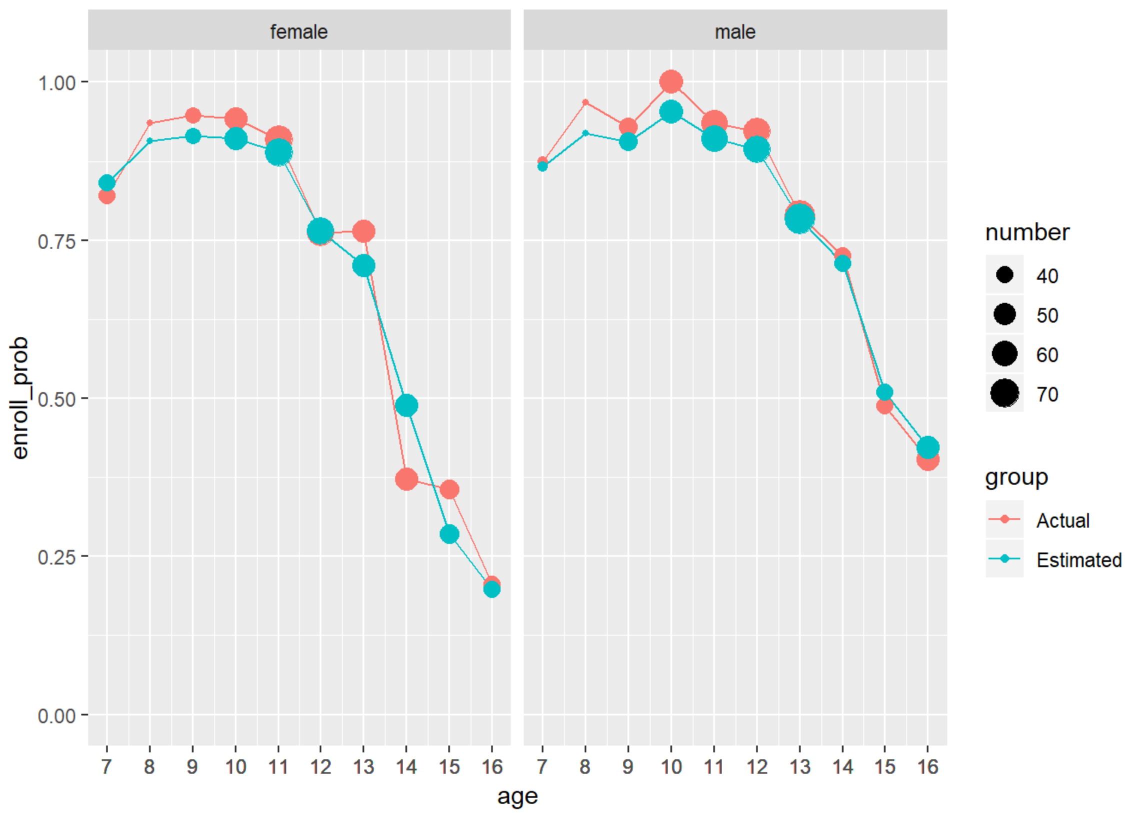


Figure 5: DPS: in-sample fit by years of education completed

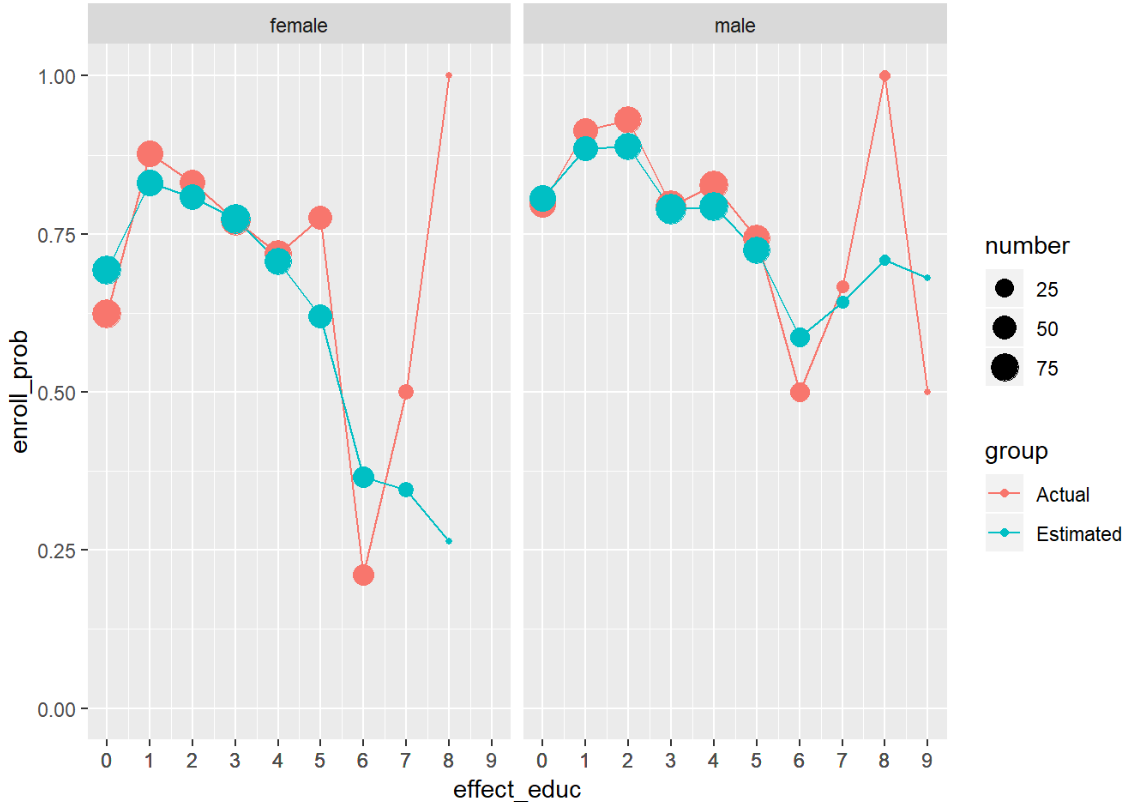
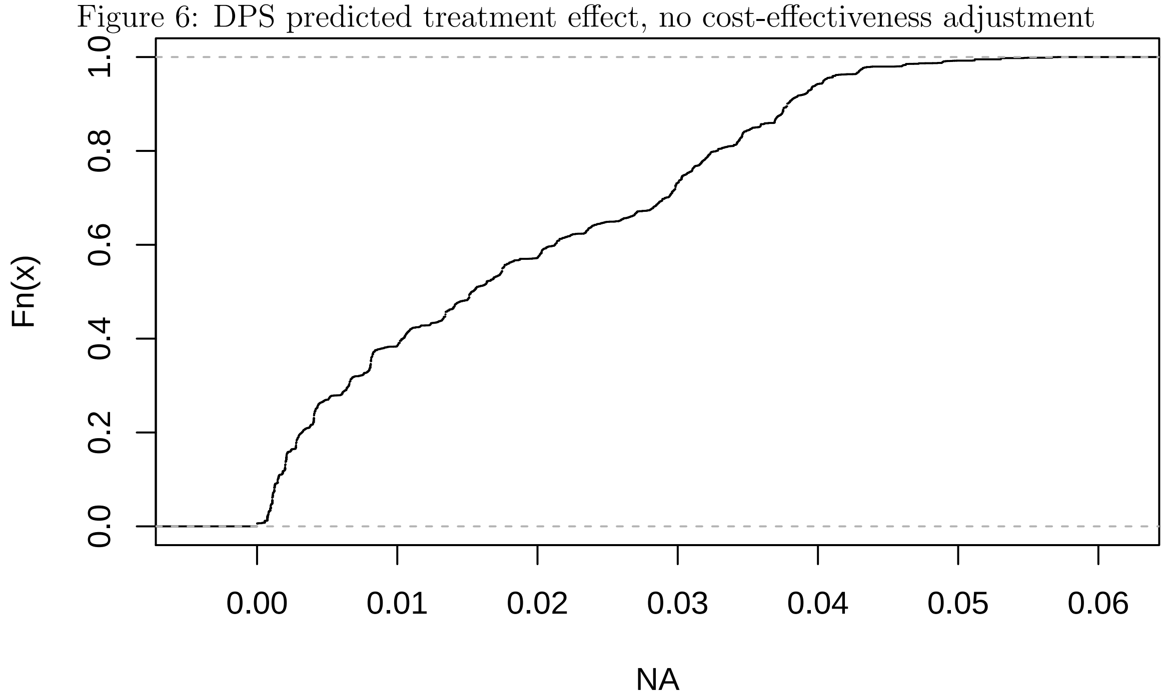


Figure 6: DPS predicted treatment effect, no cost-effectiveness adjustment



Age	Male	Treated	Control	CATE	PEN_ub	PNE_ub
6.00	1.00	0.93	0.91	0.02	0.09	0.07
7.00	1.00	0.96	0.96	-0.00	0.04	0.04
8.00	1.00	0.97	0.96	0.01	0.04	0.03
9.00	1.00	0.98	0.96	0.02	0.04	0.02
10.00	1.00	0.97	0.96	0.01	0.04	0.03
11.00	1.00	0.96	0.93	0.03	0.07	0.04
12.00	1.00	0.91	0.86	0.05	0.14	0.09
13.00	1.00	0.83	0.76	0.07	0.24	0.17
14.00	1.00	0.74	0.60	0.14	0.40	0.26
15.00	1.00	0.53	0.46	0.07	0.53	0.46
16.00	1.00	0.35	0.32	0.03	0.35	0.32
6.00	0.00	0.90	0.89	0.01	0.11	0.10
7.00	0.00	0.95	0.95	0.00	0.05	0.05
8.00	0.00	0.96	0.96	0.00	0.04	0.04
9.00	0.00	0.97	0.95	0.01	0.05	0.03
10.00	0.00	0.97	0.96	0.00	0.04	0.03
11.00	0.00	0.94	0.92	0.03	0.08	0.06
12.00	0.00	0.88	0.79	0.09	0.21	0.12
13.00	0.00	0.74	0.68	0.05	0.32	0.26
14.00	0.00	0.65	0.52	0.13	0.48	0.35
15.00	0.00	0.41	0.35	0.06	0.41	0.35
16.00	0.00	0.33	0.25	0.07	0.33	0.25

Table 1: Age-sex Outcomes for Mexico

Age	Boys treated	Boys control	Girls treated	Girls control
6.00	0.84	0.91	0.82	0.89
7.00	0.87	0.96	0.87	0.95
8.00	0.85	0.96	0.85	0.96
9.00	0.86	0.96	0.85	0.95
10.00	0.83	0.96	0.82	0.96
11.00	0.82	0.93	0.80	0.92
12.00	0.77	0.86	0.75	0.79
13.00	0.70	0.76	0.63	0.68
14.00	0.63	0.60	0.56	0.52
15.00	0.45	0.46	0.35	0.35
16.00	0.30	0.32	0.28	0.25

Table 2: Adjusted Age-sex Outcomes for Mexico

	w/ CE	w/o CE
Share treated (age-sex)	0.126	0.966
Share treated (all)	1.000	1.000
Enrollment difference	-0.035	-0.005
SE enroll. diff.	0.011	0.004
Welfare difference	0.065	-0.002
SE welfare diff.	0.010	0.004

Table 3: Welfare comparison for age-sex extrapolation vs. treat all

Variable	Importance Y1	Importance Y0
ml_base_enrolled	0.21	0.19
ml_age	0.33	0.32
ml_n_child	0.02	0.01
ml_male	0.01	0.01
ml_hh_head_male	0.00	0.00
ml_single_parent	0.00	0.00
ml_al_father	0.00	0.01
ml_al_mother	0.00	0.00
ml_lw_father	0.00	0.00
ml_lw_mother	0.00	0.00
ml_literacy	0.03	0.04
ml_n_total	0.01	0.01
ml_yrs_educ	0.16	0.18
ml_hh_head_edu	0.07	0.07
ml_father_educ	0.05	0.04
ml_mother_educ	0.07	0.08
ml_base_enrolled_mi	0.01	0.01
ml_male_mi	0.00	0.00
ml_al_father_mi	0.00	0.00
ml_al_mother_mi	0.00	0.00
ml_literacy_mi	0.01	0.01
ml_yrs_educ_mi	0.01	0.01
ml_father_educ_mi	0.01	0.00
ml_mother_educ_mi	0.00	0.00
ml_hh_monthly_consump	0.00	0.00
ml_n_child_mi	0.00	0.00
ml_lw_father_mi	0.00	0.00
ml_hh_monthly_consump_mi	0.00	0.00

Table 4: GRF variable importance: no cost-effectiveness adjustment

Variable	Importance Y1	Importance Y0
ml_base_enrolled	0.19	0.19
ml_age	0.28	0.27
ml_n_child	0.02	0.01
ml_male	0.01	0.01
ml_hh_head_male	0.01	0.00
ml_single_parent	0.00	0.00
ml_al_father	0.01	0.01
ml_al_mother	0.03	0.03
ml_lw_father	0.00	0.00
ml_lw_mother	0.01	0.01
ml_literacy	0.03	0.04
ml_n_total	0.01	0.01
ml_yrs_educ	0.16	0.15
ml_hh_head_edu	0.08	0.08
ml_father_educ	0.05	0.04
ml_mother_educ	0.06	0.08
ml_base_enrolled_mi	0.01	0.01
ml_male_mi	0.01	0.00
ml_al_father_mi	0.00	0.00
ml_al_mother_mi	0.00	0.00
ml_literacy_mi	0.01	0.02
ml_yrs_educ_mi	0.01	0.02
ml_father_educ_mi	0.00	0.00
ml_mother_educ_mi	0.01	0.00

Table 5: GRF variable importance: Y1 adjusted for cost-effectiveness

Age	Male	Avg. TE	Min. TE	Max TE	SD TE
6.00	1.00	-0.06	-0.20	0.08	0.06
7.00	1.00	-0.05	-0.18	0.12	0.06
8.00	1.00	0.03	-0.07	0.12	0.04
9.00	1.00	0.06	-0.04	0.16	0.04
10.00	1.00	0.06	-0.10	0.16	0.03
11.00	1.00	0.04	-0.07	0.15	0.03
12.00	1.00	0.05	-0.14	0.18	0.05
13.00	1.00	0.07	-0.19	0.19	0.09
14.00	1.00	0.15	-0.11	0.41	0.08
15.00	1.00	0.20	-0.16	0.45	0.11
16.00	1.00	0.09	-0.26	0.32	0.14

Table 6: GRF CATE predictions for Morocco: Boys, no cost-effectiveness adjustment

Age	Male	Avg. TE	Min. TE	Max TE	SD TE
6.00	1.00	-0.04	-0.21	0.14	0.05
7.00	1.00	-0.02	-0.14	0.12	0.05
8.00	1.00	0.01	-0.09	0.17	0.05
9.00	1.00	0.03	-0.12	0.17	0.06
10.00	1.00	0.03	-0.09	0.14	0.04
11.00	1.00	0.02	-0.08	0.15	0.03
12.00	1.00	0.04	-0.14	0.16	0.04
13.00	1.00	0.05	-0.18	0.19	0.09
14.00	1.00	0.14	-0.08	0.41	0.09
15.00	1.00	0.14	-0.12	0.39	0.13
16.00	1.00	0.06	-0.18	0.38	0.11

Table 7: GRF CATE predictions for Morocco: Girls, no cost-effectiveness adjustment

Age	Male	Avg. Adj. TE	Min. Adj. TE	Max Adj. TE	SD Adj. TE
6.00	1.00	-0.10	-0.22	0.05	0.06
7.00	1.00	-0.08	-0.18	0.06	0.05
8.00	1.00	-0.03	-0.11	0.09	0.03
9.00	1.00	-0.00	-0.07	0.12	0.04
10.00	1.00	-0.02	-0.13	0.09	0.03
11.00	1.00	-0.03	-0.12	0.08	0.03
12.00	1.00	-0.01	-0.16	0.11	0.04
13.00	1.00	0.01	-0.24	0.11	0.07
14.00	1.00	0.07	-0.11	0.26	0.06
15.00	1.00	0.14	-0.17	0.33	0.10
16.00	1.00	0.05	-0.23	0.23	0.11

Table 8: GRF CATE predictions for Morocco: Boys, adjusted for cost-effectiveness

Age	Male	Avg. Adj. TE	Min. Adj. TE	Max Adj. TE	SD Adj. TE
6.00	1.00	-0.08	-0.24	0.06	0.04
7.00	1.00	-0.06	-0.16	0.05	0.04
8.00	1.00	-0.05	-0.17	0.09	0.05
9.00	1.00	-0.04	-0.14	0.11	0.05
10.00	1.00	-0.05	-0.14	0.04	0.03
11.00	1.00	-0.05	-0.13	0.12	0.02
12.00	1.00	-0.03	-0.16	0.05	0.04
13.00	1.00	-0.01	-0.22	0.10	0.07
14.00	1.00	0.07	-0.08	0.28	0.06
15.00	1.00	0.09	-0.12	0.31	0.11
16.00	1.00	0.02	-0.21	0.27	0.09

Table 9: GRF CATE predictions for Morocco: Girls, adjusted for cost-effectiveness

Age	Girls share	Girls treat rate	Boys share	Boys treat rate
6.00	0.05	0.18	0.06	0.17
7.00	0.07	0.28	0.07	0.23
8.00	0.09	0.55	0.08	0.81
9.00	0.08	0.73	0.08	0.94
10.00	0.10	0.86	0.12	0.94
11.00	0.12	0.78	0.12	0.88
12.00	0.13	0.86	0.13	0.91
13.00	0.12	0.71	0.12	0.83
14.00	0.09	0.95	0.10	0.98
15.00	0.09	0.91	0.08	0.94
16.00	0.06	0.80	0.05	0.79

Table 10: GRF treatment rates for Morocco, no cost-effectiveness adjustment

Age	Girls share	Girls treat rate	Boys share	Boys treat rate
6.00	0.05	0.05	0.06	0.04
7.00	0.07	0.09	0.07	0.04
8.00	0.09	0.14	0.08	0.20
9.00	0.08	0.27	0.08	0.41
10.00	0.10	0.09	0.12	0.27
11.00	0.12	0.03	0.12	0.12
12.00	0.13	0.19	0.13	0.39
13.00	0.12	0.51	0.12	0.62
14.00	0.09	0.89	0.10	0.93
15.00	0.09	0.79	0.08	0.91
16.00	0.06	0.58	0.05	0.73

Table 11: GRF treatment rates for Morocco, adjusted for cost-effectiveness

	w/ CE	w/o CE
Share treated (GRF)	0.377	0.772
Share treated (age-sex)	0.126	0.966
Enrollment difference	0.011	-0.002
SE enroll. diff.	0.010	0.008
Welfare difference	-0.018	0.014
SE welfare diff.	0.009	0.008

Table 12: Welfare comparison for GRF vs. age-sex extrapolation

	w/ CE
Share treated (SPS)	0.447
Share treated (age-sex)	0.126
Enrollment difference	-0.018
SE enroll. diff.	0.010
Welfare difference	-0.046
SE welfare diff.	0.009

Table 13: Welfare comparison for SPS vs. age-sex extrapolation

	w/ CE
Share treated (DPS)	0.001
Share treated (age-sex)	0.151
Enrollment difference	-0.017
SE enroll. diff.	0.008
Welfare difference	-0.004
SE welfare diff.	0.008

Table 14: Welfare comparison for DPS vs. age-sex extrapolation