Evaluating Counterfactual Predictions from Competing Methods

Michael Gechter[1] Cyrus Samii[2]
Rajeev Dehejia[2] Kiki Pop-Eleches[3]

[1]Penn State

[2]NYU

[3]Columbia

GPI Workshop on New Approaches in Casual Inference and Extrapolation

Research question

- Social scientists are often asked for policy advice
- Predicting the effects of counterfactual policies in order to choose one
- How can we determine what methods are successful at this task?
    - Important for future decisions

What is a method?

For a defined class of policies, an approach to determining which should be enacted

1. A model ("reduced form", "structural")
2. An expert
3. A combination of 1 and 2

Getting concrete...

Our empirical example

- (Conditional) cash transfer programs
  - Parents are paid, often conditional on enrolling children in school
  - Widespread (more than 80 globally - Parker and Vogl (2018))
- We have experimentally evaluated a CCT program in Mexico (PROGRESA)
  - Finding a substantial positive effect on school enrollment
- A decision: should a similar program be implemented in Morocco?

There are several ways to predict the effect

1. "Reduced form"
   - Unconfounded location: Hotz, Imbens, and Mortimer (2005), Dehejia, Pop-Eleches, and Samii (2017)
   - Meta-analysis: Dehejia (2003), Meager (2016), Vivalt (2016)
2. "Structural": Todd and Wolpin (2006), Todd and Wolpin (2010), Attanasio, Meghir, and Santiago (2012)
3. Hybrid: Gechter (2016)
4. Asking experts: Banerjee, Chassang, and Snowberg (2016), DellaVigna and Pope (2017)

- Now we do an ex-post analysis of the effectiveness of the program (an experiment - Benhassine, Devoto, Duflo, Dupas, and Pouliquen (2015))
- Which method performed the best in predicting effectiveness?
  - From the perspective of helping us decide which policy to implement

A naive approach to answering the question

- Did method m correctly predict whether the Moroccan program should have been implemented?
- It will be difficult to discriminate between methods this way
  - We have few social experiments for any class of policies (usually ≤ 10)
  - So, many methods will have identical performance

Our approach

- We analyze the implementation problem
- And show each experiment carries much more decision-relevant information
- Subgroup average effects tell us who should be eligible for treatment
    - Manski (2004), Hirano and Porter (2009), Kitagawa and Tetenov (2017), Athey and Wager (2017)
- Different methods will predict different subgroup effects, generating different recommendations for who should be treated
- For each experimental treatment arm, did method m correctly predict who should be treated?
- Is the performance difference between methods l and m statistically significant?

Empirical illustration

Using Mexico and the Moroccan control group to predict Morocco

- We compare the performance of two methods
  - Extrapolation: use estimated subgroup effects from Mexico to decide who should be eligible in Morocco
  - Todd and Wolpin (2010)'s non-parametric structural approach: use the Moroccan control group outcomes to predict subgroup effects
- Speaks to Pritchett and Sandefur (2013)'s question: is local, potentially confounded information more useful than extrapolating experimtental effects?
- In our illustration extrapolation outperforms the non-parametric structural approach
- But this is a preliminary result

Existing work on this question

- No framework for formal analysis we're aware of
- Informal comparisons abound in applied micro, often under the heading of model validation
  - Using holdout samples: Todd and Wolpin (2006), Keane and Wolpin (2007), Duflo, Hanna, and Ryan (2012), Wolpin (2013)
- The closest work is in the time series forecast evaluation literature
- Our framework allows us to leverage results from this literature for inference

Related literature

- Using information on the use of prediction in judging methods: Pesaran and Skouras (2002), Granger and Machina (2006)
- Forecast evaluation theory: Diebold and Mariano (1995), White (2000), Hansen, Lunde, and Nason (2011)
- Evaluating experts: DellaVigna and Pope (2017)
- Statistical treatment assignment rules: Manski (2004), Hirano and Porter (2009), Kitagawa and Tetenov (2017), Athey and Wager (2017)
- Prediction-based model comparison: Keane and Wolpin (2007), Wolpin (2007), Wolpin (2013), Schorfheide and Wolpin (2012), Schorfheide and Wolpin (2016)
- CCTs: Banerjee, Hanna, Kreindler, and Olken (2017), Benhassine et al. (2015), De Janvry and Sadoulet (2006), Todd and Wolpin (2006), Attanasio et al. (2012)

General setup

- Let $\mathcal{M}$ define the methods under consideration
- A "method" $m \in \mathcal{M}$ is capable of producing predictions for treatments $\mathcal{T}_m$ finite
- The status quo treatment $t = 0 \in \mathcal{T}_m \; \forall m \in \mathcal{M}$
  - m could be an expert in CCTs where each $t \in \mathcal{T}_m$ represents an alternative subsidy schedule, including $t = 0$ (no subsidy)
- $\mathcal{T} = \bigcup_{m \in \mathcal{M}} \mathcal{T}_m$ defines the set of treatments covered by methods $\mathcal{M}$

Experimental treatment arms

- C different experiments, indexed by $c \in \{1, \ldots, C\}$
- Let $\mathcal{T}_c \subseteq \mathcal{T}$ denote the set of treatments active in context c
  - E.g. PROGRESA subsidies increased with age and were higher for girls
  - Moroccan transfers were smaller and not differentiated by gender
- Individual i belongs to a context c
- Context characteristics $V_c \in \mathcal{V}$ finite

Judging methods

Design parameters

- A judge would like to use the data from the C experiments to assess the methods in $\mathcal{M}$ according to their ability to assign individuals to treatments $\mathcal{T}_c$ which maximize the judge's social welfare function
- Outcome $Y_{ic} \in \mathbb{R}$, assumed to have finite first and second moments in each experiment
  - Ex. school enrollment
  - Represents individual utility from the judge's perspective
- Individual characteristics (subgroups) $W_{ic} \in \mathcal{W}$ finite
  - Subgroups where it is feasible to assign individuals to different treatments
  - Ex. gender, age

Neyman-Rubin potential outcomes framework

For simplicity we will consider $T_{ic} \in \{0, 1\}$

$$Y_{ic} = Y_{1ic} T_{ic} + Y_{0ic}(1 - T_{ic})$$

Assignment of treatment is random within context

$$(Y_{0ic}, Y_{1ic}, W_{ic}) \perp\!\!\!\perp T_{ic} | c$$
$$P(T_{ic} = 1 | c) = p_1, \text{ known}$$

The prediction problem

- The judge wants to learn methods' ability to generate welfare-maximizing treatment assignment rules
- We therefore consider two sets of contexts

$$D_c \in \{0, 1\} \in V_c$$

- Methods have access to a sample from the distribution of observable data in contexts with $D_c = 0$

$$(Y_{ic}, V_c, W_{ic}, T_{ic}) | D_c = 0$$

enabling identification of

$$(Y_{0ic}, W_{ic}) | V_c, D_c = 0 \text{ and } (Y_{1ic}, W_{ic}) | V_c, D_c = 0$$

  - Ex. Mexico

Prediction

- Methods can only access a sample from the distribution of untreated individuals in contexts with $D_c = 1$

$$(Y_{ic}, V_c, W_{ic}) | D_c = 1$$

so they can only identify

$$(Y_{0ic}, W_{ic}) | V_c, D_c = 1$$

  - Ex. Morocco

- The judge has access to sample data from the distribution of observable data in all contexts

$$(Y_{ic}, V_c, W_{ic}, T_{ic})$$

so she can also identify

$$(Y_{1ic}, W_{ic}) | V_c, D_c = 1$$

The judge's objective

Judge design parameters

- Let $p_w^c = P(W_{ic} = w|c)) \in (0,1)$.
- Manski (2004), Hirano and Porter (2009):

$$\max_{\pi \in \Pi} \sum_{w \in \mathcal{W}} p_w^c \left[\pi^c(w) \mathcal{U}(F_{Y_1|W}(\cdot), 1) + (1 - \pi^c(w)) \mathcal{U}(F_{Y_0|W}(\cdot), 0) \right] \tag{1}$$

where

- $\mathcal{U}(F_{Y|W}, t)$: judge's consequentialist social welfare function
- $\pi : \mathcal{W} \times \mathcal{V} \to [0,1], \pi^c : \mathcal{W} \to [0,1]$
- Π represents judge-determined constraints on the set of possible treatment assignment rules. I.e.,
  - Budgetary
  - Feasibility

The judge's objective

Comments

$$\max_{\pi \in \Pi} \sum_{w \in \mathcal{W}} p_w^c \left[\pi^c(w) \mathcal{U}(F_{Y_1|W}(\cdot), 1) + (1 - \pi^c(w)) \mathcal{U}(F_{Y_0|W}(\cdot), 0) \right]$$

- By definition, π^c only considers the treatment considered in c
- Not alternative treatments
- Ex. assignment of eligibility for *Morocco's CCT program* to subgroups
- *Not* assignment of Mexico's subsidy schedule (or an alternative) to Morocco
- Active research projects Coville and Vivalt (2017) and Hjort, Moreira, Santini, and Rao (2018) could provide insight into the choice of $\mathcal{U}(\cdot)$

Simplifying

From now on, implicity condition on c except for clarity and let

$$\mathcal{U}(F_{Y|W}, t) = \mu_{tw} = E[Y_{tic}|W_i = w]$$

Abstracting from

1. Inequality aversion (Dehejia (2008))
2. The judge's attitude towards uncertainty (Dehejia (2008))
3. Or her asymmetric attitude towards different treatments (i.e., status quo bias - see Tetenov (2012))

The objective is then

$$\max_{\pi \in \Pi} \sum_{w \in \mathcal{W}} p_w^c [\pi(w)(\mu_{1w} - \mu_{0w}) + \mu_{0w}] \tag{2}$$

- To incorporate judge risk aversion let

$$\mathcal{U}(F_{Y|W}, t) = \mathcal{U}(\mu_{tw})$$

Treament assignment rule m

- Method m is tasked with providing a vector of objects enabling the judge to select an method-specific treatment assignment rule π_m maximizing (1)
- With linear social welfare (2), the objects are predicted conditional average treatment effects $\hat{\tau}_{wm} = \hat{\mu}_{1wm} - \hat{\mu}_{0wm}$:

$$\begin{aligned}\pi_m &= \arg\max_{\pi \in \Pi} \sum_{w \in \mathcal{W}} p_w [\pi(w)\hat{\mu}_{1wm} + (1 - \pi(w))\hat{\mu}_{0wm}] \\ &= \arg\max_{\pi \in \Pi} \sum_{w \in \mathcal{W}} p_w [\pi(w)\hat{\tau}_{wm} + \hat{\mu}_{0wm}]\end{aligned}$$

Welfare of m's prediction

- The welfare associated with m's prediction is

$$U(\pi_m, \mu_1, \mu_0) = \sum_{w \in \mathcal{W}} p_w [\pi_m(w)\mu_{1w} + (1 - \pi_m(w))\mu_{0w}]$$

- $\mu_0 = \{\mu_{0w} \forall w \in \mathcal{W}\}$.
- $\mu_1 = \{\mu_{1w} \forall w \in \mathcal{W}\}$
- Recall: $\mu_{tw} = E[Y_{tic} | W_{ic} = w, c]$, which the judge can identify as

$$E[Y_{ic} | T_{ic} = t, W_{ic} = w, c]$$

  by random assignment of T_{ic} within context
- We rely on the Stable Unit Treatment Value (SUTVA) assumption embedded in the Neyman-Rubin causal model
- Specifically, we assume no $\pi \in \Pi$ will change $\{\mu_0, \mu_1\}$ (no change in GE effects)

Welfare contrasts for methods m and l

$$U(\pi_m, \mu_1, \mu_0) - U(\pi_l, \mu_1, \mu_0) =$$
$$\Delta_{lm} = \sum_{w \in \mathcal{W}} p_w(\pi_m(w) - \pi_l(w))(\mu_{1w} - \mu_{0w})$$

1. The welfare constrast is non-zero for values of w for which the experts disagree on treatment assignment
2. It is then the value of the conditional average treatment effect $\mu_{1w} - \mu_{0w}$ when m says to treat and l says not to (or the reverse), weighted by the fraction of context c's population belonging to subgroup w

With judge risk aversion

- The relevant objects are distributions for μ_{1w} and μ_{0w}

$$\pi_m = \arg\max_{\pi \in \Pi} \sum_{w \in \mathcal{W}} p_w \left[\pi(w) \int \mathcal{U}(\mu_{1w}) dF_m(\mu_{1w}) + (1 - \pi(w)) \int \mathcal{U}(\mu_{0w}) dF_m(\mu_{0w}) \right]$$

- m is uncertain about the point predictions for μ_0, μ_1, for example due to sampling variation
- The judge wants to take this into account
- Welfare is again defined in terms of the actual expectations

$$U(\pi_m, \mu_1, \mu_0) = \sum_{w \in \mathcal{W}} p_w [\pi_m(w) \mathcal{U}(\mu_{1w}) + (1 - \pi_m(w)) \mathcal{U}(\mu_{0w})]$$

Estimation and inference

We use the following moment conditions

$$
\begin{aligned}
&E[1\{c\} - p^c] = 0 \\
&E[1\{W_i = w, c\} - p_w^c p^c] = 0 \ \forall w \in \mathcal{W} \setminus \mathcal{W}^{(|\mathcal{W}|)} \\
&E[Y_{ic} 1\{W_i = w, T_i = 1, c\} - \mu_{1w}^c p_w^c p_1^c p^c] = 0 \ \forall w \in \mathcal{W} \\
&E[Y_{ic} 1\{W_i = w, T_i = 0, c\} - \mu_{0w}^c p_w^c (1 - p_1^c) p^c] = 0 \ \forall w \in \mathcal{W}
\end{aligned}
$$

where

$$
\begin{aligned}
&p^c \in (0, 1] \\
&p_w^c \in [0, 1], \\
&\sum_{w \in \mathcal{W} \setminus \mathcal{W}^{(|\mathcal{W}|)}} p_w^c \leq 1
\end{aligned}
$$

$$\Delta(p, \mu_0, \mu_1, \pi_m, \pi_l) =$$
$$\sum_{w \in \mathcal{W} \setminus \mathcal{W}^{(|\mathcal{W}|)}} p_w^c(\pi_l(w) - \pi_m(w))(\mu_{1w} - \mu_{0w})$$
$$+ (1 - \sum_{w \in \mathcal{W} \setminus \mathcal{W}^{(|\mathcal{W}|)}} p_w^c)(\pi_l(\mathcal{W}^{(|\mathcal{W}|)}) - \pi_m(\mathcal{W}^{(|\mathcal{W}|)}))(\mu_{1\mathcal{W}^{(|\mathcal{W}|)}} - \mu_{0\mathcal{W}^{(|\mathcal{W}|)}})$$

Proposition 1.

Under the maintained assumptions

$$\begin{aligned}\hat{\Delta}_{lm} &= \Delta(\hat{p}, \hat{\mu}_0, \hat{\mu}_1, \pi_l, \pi_m) \\ &= \sum_{w \in \mathcal{W}} \hat{p}_w((\pi_l(w) - \pi_m(w))(\hat{\mu}_{1w} - \hat{\mu}_{0w})\end{aligned}$$

is consistent for $\Delta(p, \mu_0, \mu_1, \pi_l, \pi_m)$.

Note

- We are treating π_m, π_l as fixed, not as empirical objects depending on sampling variation
- Uncertainty from the evaluator's perspective concerns sampling variation in p, μ_0, and μ_1
- We want to allow for expert predictions which have the potential to improve on model-based predictions (Banerjee, Chassang, Montero, and Snowberg (2017))
- Analogous to Diebold and Mariano (1995) vs. later work by West

Consistency with judge risk aversion

An analogous result applies for the uncertainty-averse evaluator with a different $\Delta(\cdot)$ function if $\mathcal{U}(\cdot)$ is continuously differentiable. Then

$$\sum_{w \in \mathcal{W} \setminus \mathcal{W}^{(|\mathcal{W}|)}} p_w(\pi_l(w) - \pi_m(w))(\mathcal{U}(\hat{\mu}_{1w}) - \mathcal{U}(\hat{\mu}_{0w}))$$
$$+ (1 - \sum_{w \in \mathcal{W} \setminus \mathcal{W}^{(|\mathcal{W}|)}} p_w)(\pi_l(\mathcal{W}^{(|\mathcal{W}|)}) - \pi_m(\mathcal{W}^{(|\mathcal{W}|)}))(\mathcal{U}(\hat{\mu}_{1\mathcal{W}^{(|\mathcal{W}|)}}) - \mathcal{U}(\hat{\mu}_{0\mathcal{W}^{(|\mathcal{W}|)}}))$$

is consistent for $\Delta(p, \mu_0, \mu_1, \pi_l, \pi_m)$.

Inference

Let

$$\Sigma = E\left[\begin{bmatrix} \hat{p}^c - p^c \\ \hat{p} - p \\ \hat{\mu}_1 - \mu_1 \\ \hat{\mu}_0 - \mu_0 \end{bmatrix} \left[\hat{p}^c - p^c, (\hat{p} - p)', (\hat{\mu}_1^c - \mu_1^c)', (\hat{\mu}_0^c - \mu_0^c)'\right]\right]$$

Proposition 2.

Under the maintained assumptions

$$\sqrt{N}\left(\begin{bmatrix} \Delta(\hat{p}, \hat{\mu}_0, \hat{\mu}_1, \pi_1, \pi_2) - \Delta(p, \mu_0, \mu_1, \pi_1, \pi_2) \\ \vdots \\ \Delta(\hat{p}, \hat{\mu}_0, \hat{\mu}_1, \pi_{M-1}, \pi_M) - \Delta(p, \mu_0, \mu_1, \pi_{M-1}, \pi_M) \end{bmatrix}\right) \to \mathcal{N}(0, \Delta_\theta \Sigma \Delta_\theta')$$

where

$$\Delta_\theta = \begin{bmatrix} \frac{\partial \Delta(\theta, \pi_1, \pi_2)}{\partial \theta'} \\ \vdots \\ \frac{\partial \Delta(\theta, \pi_{M-1}, \pi_M)}{\partial \theta'} \end{bmatrix}$$

$$\theta = (p', \mu_1', \mu_0')'.$$

Inference

- This result is sufficient to do inference on the performance difference between a pair of experts
- To form a "method confidence set" of level $1 - \alpha$ apply Hansen et al. (2011)'s sequential algorithm
    - A set of methods including the best performing with probability $1 - \alpha$

Empirical illustration

- We return to the CCT setting: using Mexico and the Moroccan control group data to predict Morocco
- And compare the performance of two different methods
  1. Direct extrapolation: use experimental treatment effect estimates to estimate $\tau_{w,e}$ and compute π_e
  2. Non-parametric structural (Todd and Wolpin (2010)): π_{nps}
     - Wolpin (2013) reports similar or better performance relative to Todd and Wolpin (2006)
- Is local observational analysis preferred to extrapolation of experimental effects?
- We can do simple two-way inference on the welfare differential

Evaluator design parameters

- Y_{ic}: enrollment of child i
- $\mathcal{W}$: gender $\times$ age
- Π: cost per student of 50 Moroccan dirhams (MAD) per month
  - Transfers for 10-16 year-olds are 100 MAD per month
  - 50 is arbitrary, but shrinking the budget induces tradeoffs between methods

Extrapolation

$$\pi_e = \arg\max_{\pi \in \Pi} \sum_{w \in \mathcal{W}} \hat{p}_w [\pi(w) \hat{\tau}_{w,e} + \hat{\mu}_{0w,e}] \tag{3}$$

- where $\hat{\tau}_{w,e}$ is an estimated conditional average effect for subgroup w in Mexico
- and $\hat{p}_w$ is computed from the Moroccan data
- As in Allcott (2015)
- (3) can be solved through linear programming

Subgroup fractions in Morocco

age	male	female
10	0.071	0.067
11	0.080	0.078
12	0.089	0.081
13	0.085	0.082
14	0.069	0.065
15	0.062	0.060
16	0.053	0.059

PROGRESA conditional average treatment effects

age	male	female
10	0.023	0.039
11	0.021	-0.017
12	0.045	0.086
13	0.019	0.052
14	0.135	0.096
15	0.015	0.187
16	-0.020	-0.063

- Based on Attanasio et al. (2012)'s difference-in-difference approach

Linear program

$$\pi_m = \arg \max_{\pi \in [0,1]^{14}} \sum_{w \in \mathcal{W}} \hat{p}_w \pi(w) \hat{\tau}_{wm}$$

$$\text{subject to}$$

$$\sum_{w \in \mathcal{W}} 100 \hat{p}_w (\hat{\tau}_{wm} + \hat{\mu}_{0wm}) \leq 50$$

Treatment assignment based on PROGRESA CATEs

age	male	female
10	0.791	1.0
11	0.000	0.0
12	1.000	1.0
13	0.000	1.0
14	1.000	1.0
15	1.000	1.0
16	0.000	0.0

Todd and Wolpin (2010)'s "non-parametric structural model" of school attendance

One child

Households solve

$$U(c, y; w, \epsilon)$$
$$s.t.\ c = n + e(1 - y)$$

where

- c: consumption
- n: household income excluding child earnings
- e : child's wage offer

Optimal school attendance

$$s^* = \phi(n, e; w, \epsilon) = 1\{U(n, 1; w, \epsilon) > U(n + e, 0; w, \epsilon)\}$$

Add a conditional subsidy s

$$c = n + e(1 - y) + sy$$
$$c = (n + s) + (e - s)(1 - y)$$

So

$$s^{**} = \phi(n + s, e - s; w, \epsilon)$$

$$\tilde{n} = n + s$$
$$\tilde{e} = e - s$$

Assume

$$f(\epsilon|n, e, w) = f(\epsilon|\tilde{n}, \tilde{e}, w) = f(\epsilon|w)$$

i.e. n, e are exogenous.

Then

- We can predict the effect of the subsidy for any child in the the Moroccan control group by plugging his $\tilde{n}, \tilde{e}$ into a non-parametric regression of enrollment on n and e
- We assume wage offers are observed at random
- After producing estimates $\hat{\tau}_{w,nps}$, use (3) to produce π_{nps}

Conditional average treatment effects according to non-parametric structural approach

age	male	female
10	-0.042	-0.012
11	-0.049	-0.018
12	-0.068	0.085
13	-0.044	0.074
14	-0.115	0.180
15	-0.094	0.154
16	-0.088	0.135

Treatment assignment based on non-parametric structural approach

age	male	female
10	0	0
11	0	0
12	0	1
13	0	1
14	0	1
15	0	1
16	0	1

Weighted difference in treatment assignment rules

age	male	female
10	0.056	0.067
11	0	0
12	0.089	0
13	0	0
14	0.069	0
15	0.062	0
16	0	-0.059

Results

- Based on estimated CATEs for Morocco, PROGRESA-based extrapolation outperforms NPS by 0.018 (0.006)
- Compare this to the overall ATE in Morocco: 0.085 (0.011)

Extensions and next steps

- Adding contexts and methods
- On-line learning: does model selection/averaging based on our welfare criterion produce better forecasts?
- Addressing experimental site selection by embedding our problem in Gechter and Meager (2018)
- More realistic preferences for the judge

Conclusion

- We developed a decision-based method for comparing the relative performance of different methods for generating counterfactual predictions
- Using data from an ex-post treatment effect analysis
- We are soliciting suggestions for extensions to our illustration!

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

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, S. Montero, and E. Snowberg (2017). A THEORY OF EXPERIMENTERS.

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.

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.

Coville, A. and E. Vivalt (2017). How Do Policymakers Update ? pp. 1.

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 (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. (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*.

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.

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.

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

Hjort, J., D. Moreira, J. Santini, and G. Rao (2018). The Effect of Research Information on Policy Making: Evidence from Brazil.

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.

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*.

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

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

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.

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*.

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 (2010). Ex ante evaluation of social programs. *Annales d'Economie et de Statistique* (91).

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

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.