NBER WORKING PAPER SERIES

ASSESSING EXTERNAL VALIDITY IN PRACTICE

Sebastian Galiani
Brian Quistorff

Working Paper 30398
http://www.nber.org/papers/w30398

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2022, Revised June 2023

The views expressed in this paper are those of the authors and do not necessarily represent the U.S. Bureau of Economic Analysis, the U.S. Department of Commerce, or the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2022 by Sebastian Galiani and Brian Quistorff. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Assessing External Validity in Practice
Sebastian Galiani and Brian Quistorff
NBER Working Paper No. 30398
August 2022, Revised June 2023
JEL No. C55

ABSTRACT

We review, from a practical standpoint, the evolving literature on assessing external validity (EV) of estimated treatment effects. We review existing EV measures, and focus on methods that permit multiple datasets (Hotz et al., 2005). We outline criteria for practical usage, evaluate the existing approaches, and identify a gap in potential methods. Our practical considerations motivate a novel method utilizing the Group Lasso (Yuan and Lin, 2006) to estimate a tractable regression-based model of the conditional average treatment effect (CATE). This approach can perform better when settings have differing covariate distributions and allows for easily extrapolating the average treatment effect to new settings. We apply these measures to a set of identical field experiments upgrading slum dwellings in three different countries (Galiani et al., 2017).

Sebastian Galiani
Department of Economics
University of Maryland
3105 Tydings Hall
College Park, MD 20742
and NBER
sgaliani@umd.edu

Brian Quistorff
Bureau of Economic Analysis
brian-work@quistorff.com

1 Introduction

For any empirical causal study, one can decompose its validity into internal and external components. Internal validity concerns whether the estimated effect is valid for the particular setting studied. External validity (EV), in contrast, looks beyond the sample studied. In evaluating the external validity of a set of experiments, one poses the question, to what other populations can this effect be generalized? (Campbell, 1957) In studies that utilize well-understood sources of variation, it is possible to assess their internal validity. External validity, however, is typically harder to assess as it is difficult to know how a treatment effect may change in different populations.

We review the measures of external validity,and focus on those that compare across settings. We use these methods to assess the external validity of estimated treatments effects from an existing study (Galiani et al., 2017) that conducted identical experiments in three countries. Since these were randomized controlled trials (RCTs), the threats to internal validity are small and addressed in the original study. Thus, we focus our attention here on external validity.

For practicality, we will focus our evaluation of methods on (a) does the method provide a simple statistical test of whether the effect generalizes across the available settings (and does this extend to more than two settings), and (b) how easily can results be extrapolated to predict the expected treatment effect in a new setting (e.g., *does one need the full original data?*, and *will extrapolating to a new domain require changing the interpretation of the estimate in a complicated way?*).

Single-setting measures of external validity were proposed by Bo and Galiani (2021). They provide a formal definition of external validity and a general theoretical treatment, as well as propose two specific measures for assessing external validity based on how estimations vary as the experimental data are reweighted. Reweightings are used to simulate

different possible populations. Reweighting "enables the researcher to compare the treatment effects in different locations" (Athey and Imbens, 2017). Bo and Galiani (2021) base their method on 1-to-1 matching. After constructing treated-control pairs, they generate reweighting vectors uniformly distributed over all possible reweighting vectors. They categorize treatment effects according to their statistical-significance category (*positive significant*, *insignificant*, and *negative significant*), and then gauge how often a reweighted sample results in an estimate that is in a different category. This measure of EV is derived from their more general definition of external validity, namely, external validity on the overarching population. They also propose a local measure that relates how their measure of EV changes with the correlation between the reweighting vector and the pair-level outcome differences (as this directly impacts the treatment effects). This measure of EV is motivated by their definition of external validity from one population onto another, letting the degree of external validity depend on how different is the parameter vector that characterizes the target population in relation to the one that characterizes the sample studied. While providing some sense of generalizability, there methods are necessarily limited by their focus on a single dataset. We do not discuss their methods as it does not allow for extrapolation to specific new settings.

The above measures consider only the observable data. If one is willing to make certain assumptions about how the role of unobservables may be different in other settings, then bounds on the estimated treatment effects can be derived (Nguyen et al., 2017; Andrews and Oster, 2019; Gechter, 2021).

While some instructive information can be gathered from a single dataset, ultimately, EV is established by replicating the same experiments in different populations (Angrist, 2004; List, 2020). Conceptually, a difference in the average treatment effect (ATE) across settings could be due to either a common, but heterogeneous conditional average treatment effect (CATE) coupled with differing covariates, or entirely differing treatment effects by

setting, e.g., due to differences in unobserved "macro-effects" across settings (Hotz et al., 2005). Formally, using the potential outcomes framework of Rubin (1974), the CATE is defined as $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$ where $Y_i(d)$ is the potential outcome for unit i receiving treatment d and X are the covariates. Via iterated expectations, we can take the expectation over the CATE to recover the ATE, $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$, but this will differ if the distribution of X differs.

Some papers approach EV informally, discussing differences in means for a subset of covariates between the two datasets (Attanasio et al., 2011; Bloom et al., 2014; Muralidharan et al., 2019). It is typically difficult, however, to use this information on its own, as one would also need to know how the treatment effect differs along those dimensions.

Hotz et al. (2005) provided a formal theory for EV across settings. The main challenge is that two settings may be different, and the observable characteristics may not be sufficient for adjustment. The concern is analogous to that in treatment effect studies when there is selection bias. The necessary assumptions required for generalizability (on top of those for internal validity in the respective settings) are therefore also analogous: overlap of the settings in terms of similar observations in both the "Sample" (i.e, the inference population) versus the "Population" (i.e, the target population) and that the setting is unconfounded conditional on observable covariates. With these in place, they then take a reweighting approach to assess external validity. They estimate of model predicting whether an observation is from the Sample or Population. Then they use inverse propensity-weights to test for statistically significant differences between control unit outcomes across settings, and then treated unit outcomes across settings. From a practical perspective, this approach does provide a simple statistical test of EV. It does not, however, make it easy to extrapolate results to new settings, as there are interpretational challenges and one would likely need access to the full original data.

Stuart et al. (2011) surveys the reweighting schemes such as those used in Hotz et al.

(2005), noting that the propensity scores can be used for unit-level weights, matching, or subclassification. They also suggest that overlap in the distribution of propensity scores between both settings alone may not be sufficient for robust inference. They suggest checking the average propensity score between the two settings and that if the difference is over 0.25 standard deviations of the propensity score distribution for the controls, the results may depend too heavily on extrapolation.

An alternative to the reweighting approach is to model the CATE directly. Here, we provide a novel solution building on two strands of the literature. The first estimates CATE models from subsets of the settings and then calculates what mean-squared-error (MSE) would result in extrapolating the estimated CATE to new settings as compared to re-estimating the CATE solely on the new setting (Kern et al., 2016; Dehejia et al., 2019; Pritchett and Sandefur, 2014).[1] While these approaches do allow for extrapolation, they do not provide a simple statistical test.

To estimate a CATE model, one must first determine its "structure" – the ways in which the model allows the CATE to vary. While this is sometimes manually specified, the second strand of CATE literature employs machine learning techniques to automate this model selection. Some of these use quite flexible methods (e.g., Wager and Athey 2018 and Nie and Wager 2020), though they do not (a) allow for a simple global test of differences in CATE across settings, and (b) do not allow for easy extrapolation to new settings. Others employ linear models, such as Chernozhukov et al. (2018) and Semenova et al. (2017), the latter using regularization to achieve a lower-dimensional model. Our method builds off of this idea as it enables an easy test of the CATE across settings as well as easy extrapolation.

Throughout this paper, we will use as an empirical example the randomized control trial of Galiani et al. (2017). That research was the first to evaluated the causal effect on the

[1]With a sizable set of replications, Vivalt (2020) and Meager (2019) use Bayesian hierarchical models to evaluate the ability of a subset of studies to extrapolate to others in the set.

extremely poor of upgrading slum dwellings and was conducted in El Salvador, Mexico, and Uruguay. Inexpensive pre-fabricated houses where provided *in situ* to treated households by the same NGO in all three countries, providing considerable consistency of the treatment. In these countries, experimental sites were chosen to be marginalized sites that typical face problems of insufficient services (water, electricity, and sanitation), contamination, and overcrowding. The location characteristics differed somewhat across the countries. In El Salvador, these sites where spread throughout the country, but excluded the capital San Salvador. In Mexico the slums were adjacent to Mexico City and in Uruguay they were located in Montevideo and Canelones. The researchers' most consistent finding was improvement in the quality of life of the treated households in all three countries. They found mixed results on safety and child-health, and no effect possession of durable goods or labor outcomes. We will focus on the the quality of life treatment effect. Improving the housing conditions of the poor is a common development project in many developing countries and so we seek to understand better the generalizability of this treatment effect.

This paper proceeds as follows. In Section 2 we discuss more in detail our empirical application. In Section 3 we first assess external validity using the propensity-score reweighting methods of Hotz et al. (2005). Next, in Section 4, we assess external validity by modeling the CATE, where we build off the existing literature to develop a novel algorithm. We then show that this method allows easy extrapolation for forecasting the treatment effect in other, not studied, populations. Section 5 concludes.

2 Empirical Setting

Throughout the paper, we use as an application, the housing experiment evaluated in Galiani et al. (2017) so we first briefly describe their setting and empirical results we will focus on.

Galiani et al. (2017) estimate the effect of upgrading slum housing on the living conditions of the extreme poor. The upgrades were almost identical and done by the same organization in El Salvador (ES), Uruguay (UY), and Mexico (MX). Their main finding is that better houses have a positive effect on overall housing conditions and general well-being: treated households are happier with their quality of life.

We focus on their main outcome, the "Satisfaction" Index. This is an aggregate index that summarizes several satisfaction sub-measures: Satisfaction with Floor, Wall Quality, Roof Quality, House Protection against water when it rains, and Qualify of Life. Each of those measures is turned into a Z-score, signed so that the positive direction indicates an improvement, and then added together. We focus on their specification that controls for covariates. These comprise three sets: main baseline covariates, indicators for whether the baseline controls were imputed due to being missing, and indicators for subnational geographic clusters.

We first replicate, for each country, the original estimates of Galiani et al. (2017). Results are shown in Table 1. The estimated effect is positive and statistically significant across all three countries.

Given the consistency of the results across countries, we would hope that the results would generalize well, and so we look at methods to compare across them. We show that a simple comparison that does not account for treatment effect heterogeneity fails, and then we turn to methods to address this: reweighting and modelling the CATE.

We first, though, take some preliminary steps to make the settings more comparable. This is embodied in one of the preliminary assumptions of Hotz et al. (2005), who propose that when comparing treatment effects across countries, we restrict ourselves to analyzing "overlap" households that are similar to those in the other countries. To get a sense of the difference between countries, we first compare the distributions of a few measures that

Table 1: Satisfaction

	(1)	(2)	(3)
Treatment	1.031**	0.317**	0.295**
	(0.0866)	(0.0618)	(0.0519)
	[0.861,1.201]	[0.196,0.439]	[0.193,0.397]
Observations	656	718	826
Country	ES	UY	MX

Outcome is Satisfaction Index. Stats=b/se/ci.

Models include baseline controls from Galiani et al. (2017).

* (p <0.10), ** (p <0.05), *** (p <0.01)

summarize the housing situation: the baseline measures for the main outcome[2] along with Housing quality and Housing investment, which are the two main measures related to the physical house. We show the distributions in Figure 1. El Salvador has much lower baseline levels than the other two countries for these measures, indicating some trimming for sample overlap will be required. We subsequently include these three covariates in our main set of covariates and remove observations that are outside the min-max range of the other countries for all the main baseline variables.

With just two settings, label one the Sample and the other the Population. To assess if the unadjusted average treatment effects (ATEs) are statistically different in the two settings, we first estimate an equation where we interact the standard ATE model with an

[2]The follow-up outcome constructs Z-scores by normalizing sub-measures according to each cluster's control group's mean and standard-deviation. This is helpful when analyzing follow-up data, as it can control for variation in the scale of response across locations. When using this baseline version of this measure as a control, we want to be able to compare across clusters. We therefore normalize each sub-measure by the full (three-country) control group mean and variance.

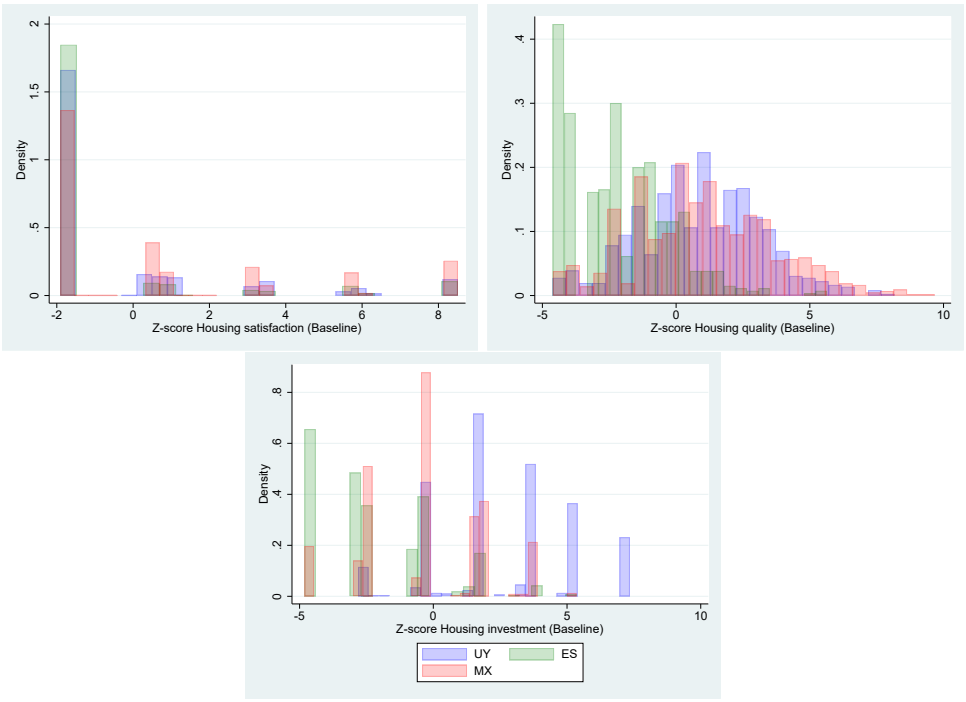

Figure 1: Comparison of Key Baseline Measures

Notes: Densities of baseline variables. El Salvador (ES) has a consistently lower distribution.

indicator for being in the Sample:

$$Y_i = D_i\beta + D_i \times \mathbb{1}_{i\in S}\delta + X_i\gamma + X_i \times \mathbb{1}_{i\in S}\gamma_d + \varepsilon_i, \tag{1}$$

where Y is the outcome (Satisfaction Index), D is the treatment assignment, X are the control variables, and $\mathbb{1}_{i\in S}$ is an indicator for whether the observation is in the Sample. We can then test the statistical significance of $\hat{\delta}$ to assess if the ATE are different in the Sample and in the Population.

With more than two settings, following Hotz et al. (2005), we rotate through them, each time considering all but one as the Sample and the other as the Population. Results for our three countries are shown in Table 2. In all three configurations, $\hat{\delta}$ is statistically significant at $p < 0.05$ and in one configuration it is also statistically significant at $p < 0.01$, indicating that the ATE is different across countries.

As mentioned before, the difference in the ATE across the countries could be due to

Table 2: ATE Sample-interacted

	(1)	(2)	(3)
Treatment × Is Sample	-0.567**	0.200**	0.199**
	(0.111)	(0.0842)	(0.0797)
Observations	1814	1814	1814
R^2	0.177	0.164	0.164
Sample	UY+MX	ES+MX	ES+UY
Population	ES	UY	MX
p-val no ATE difference	0.000000327	0.0176	0.0127

Outcome is Satisfaction Index. Stats=b/se.

Omitting non-sample-interacted coefficients.

* (p <0.10), ** (p <0.05), *** (p <0.01)

either a common, but heterogeneous CATE model coupled with differing covariates, or entirely differing treatment effects by country. We will pursue two approaches to disentangling these possibilities: a reweighting approach that attempts to make the covariate distributions similar and a regression approach that models the CATE directly.

3 Reweighting

For the reweighting approach, we follow the general path of Hotz et al. (2005), but make modifications that allow us to continue to control for covariates when estimating treatment effects.

For each Sample-Population configuration, we first estimate a prediction model over the pooled data of whether an observation is in the Sample:

$$\mathbb{1}_{i \in S} = X_i \cdot \zeta + e_i. \tag{2}$$

where $\mathbb{1}_{i \in S}$ is an indicator variable whether observation i is in the Sample, X_i are covariates used for prediction, and e_i is the residual. For the covariates, we use all available variables (the main covariates and missing indicators), though drop cluster dummies as these would perfectly predict being in the Sample and do not help characterize what differs across the settings. We estimate the model using a logistic regression. Results are shown in Table 3. Many of the covariates are statistically significant, suggesting that some adjustment is necessary to make the covariate distributions similar. Using this model we calculate predicted probabilities for each observation (propensity to be in the Sample), $\hat{p}_i$, that will be used to construct weights.

Stuart et al. (2011) suggests that if the covariate distributions are very dissimilar, then reweighting may rely heavily on extrapolation and may not be robust to functional form changes. They suggest calculating the difference between Sample and Population average predicted probabilities and dividing it by the standard deviation of the distribution of those predicted probabilities of the Population. They also suggest a rule-of-thumb cutoff of 0.25, arguing that a reweighting approach may not be trustworthy in situations where the normalized difference described above is higher than that. Table 3 shows that all the normalized differences are above the threshold, and so there should be some caution in terms of using a reweighting approach.

Following Hotz et al. (2005) we construct weights in each configuration according to inverse-probabilities to make the Sample and Population similar. Sample units are weighted by $1/\hat{p}_i$ and Population units are weighted by $1/(1-\hat{p}_i)$. To enable controlling for covariates, as the original analysis did, we use these weights and re-estimate Equation 1 using weighted OLS. Results are shown in Table 4. In this model, the coefficient on "Is Sample" reports whether the reweighting made the control outcomes similar. Hotz et al. (2005) view this

Table 3: Sample vs Population Prediction

	(1)	(2)	(3)
Is Sample			
Head of HH Educ.	0.167**	-0.158**	0.0181
	(0.0252)	(0.0211)	(0.0177)
Head of HH Female	-0.777**	1.253**	-0.625**
	(0.159)	(0.131)	(0.114)
Head of HH Age	-0.0220**	0.0166**	0.00333
	(0.00481)	(0.00474)	(0.00356)
HH Asset value/capita	-0.000347	0.000334	-0.000199
	(0.000458)	(0.000538)	(0.000362)
HH Income/capita	0.0150**	-0.000700	-0.00556**
	(0.00219)	(0.00122)	(0.00104)
Missing Head of HH Educ.	-0.240	-1.059**	0.936**
	(0.428)	(0.388)	(0.389)
Missing HH Asset value/capita	1.629**	-0.866**	-0.133
	(0.246)	(0.194)	(0.155)
Missing HH Income/capita	0.854**	0.0479	-0.426**
	(0.193)	(0.193)	(0.143)
Z-score Housing quality (Baseline)	0.559**	-0.0484**	-0.315**
	(0.0348)	(0.0241)	(0.0220)
Z-score Housing investment (Baseline)	0.301**	-0.522**	0.201**
	(0.0295)	(0.0288)	(0.0205)
Z-score Satisfaction (Baseline)	0.0830**	0.0824**	-0.0968**
	(0.0196)	(0.0182)	(0.0134)
Constant	1.551**	0.499*	1.122**
	(0.302)	(0.264)	(0.215)
Observations	2155	2155	2155
Sample	UY+MX	ES+MX	ES+UY
Population	ES	UY	MX
Pr. Score Diff.	1.848	1.519	0.894

Outcome is In Sample and statistics are coefficient, (standard error), and [confidence interval]
Estimation done by logit.
* (p <0.10), ** (p <0.05), *** (p <0.01)

Table 4: Reweighted Pooled Treatment Effect Estimation

	(1)	(2)	(3)
Treatment	1.080**	0.320**	0.361**
	(0.165)	(0.106)	(0.0622)
Is Sample	-0.266	-0.141	0.934**
	(0.234)	(0.254)	(0.218)
Is Sample x Treatment	-0.847**	0.173	0.192**
	(0.175)	(0.122)	(0.0912)
Observations	2155	2155	2155
Sample	UY+MX	ES+MX	ES+UY
Population	ES	UY	MX

Outcome is Satisfaction Index and statistics are coefficient and (standard error). Models include baseline controls. Propensity estimation done by logit.

* (p <0.10), ** (p <0.05), *** (p <0.01)

as a test of whether observable characteristics are sufficient to make the data comparable. In our data, this is statistically significant with $p < 0.05$ for one of the comparisons, indicating that covariates adjustments are likely not sufficient for that comparison. For those comparisons where this difference is not statistically significant, we can then check the coefficient on "Is Sample x Treatment" to test if the estimated treatment effect is similar across settings. We find that this is statistically insignificant only for the configuration comparing El Salvador and Mexico to Uruguay. Overall, the reweighting approach indicates that the ATE generalizes in only one of the three configurations.

To understand if these results are sensitive to the estimation of the propensity scores, we re-estimate the propensity (to be in the "Sample") function using a flexible Machine Learning algorithm, Random Forests (Breiman, 2001), that is quite common for these

Table 5: Reweighted Pooled Treatment Effect Estimation (Using Random Forest)

	(1)	(2)	(3)
Treatment	1.062**	0.306**	0.337**
	(0.107)	(0.0812)	(0.0599)
Is Sample	-0.451*	-1.434**	0.828**
	(0.255)	(0.234)	(0.237)
Is Sample x Treatment	-0.780**	0.263**	0.222**
	(0.117)	(0.0952)	(0.0832)
Observations	2155	2155	2155
Sample	UY+MX	ES+MX	ES+UY
Population	ES	UY	MX

Outcome is Satisfaction Index and statistics are coefficient and (standard error). Models include baseline controls. Propensity estimation done by random forest.

* (p <0.10), ** (p <0.05), *** (p <0.01)

types of tasks (Taddy, 2019).[3] Results are show in Table 5, and they are qualitatively similar to those in Table 4.

From a practical perspective, while this method provides a simple test for each Sample-Population configuration of the data, with more than two settings no aggregate test is given. One approach could be to reject generalizability if any of the pair-wise tests showed a statistically significant difference, while adjusting the tests to control for the Family-wise error rate (and accounting for the correlation among tests as they are estimated on the same data). Conceptually, we believe that this is likely to be too conservative to be useful

[3]To estimate the random forest model, we used R's `ranger` package with default settings of 500 "trees". We use out-of-sample ("out-of-bag") predictions for the propensity scores to remove the overfitting bias from Machine Learning models (Chernozhukov et al., 2018).

in practice, especially as S becomes large. We address the desire for a more practical, multi-setting aggregate test in Section 4.3.

Extrapolation to a new setting is feasible with reweighting, though somewhat difficult. One needs access to the original data to estimate a new propensity model between existing data (Sample) and new data (Population). To extrapolate an ATE for the Population, one would estimate the ATE over the Sample using weights $(1 - \hat{p}_i)/\hat{p}$.

4 CATE Modelling

In this section, we outline an alternative approach based on modelling the CATE directly. We will show that it can have a lower rate of false positives when covariate distributions are different. We will also recast the testing process as a simple combined test, rather than as a set of pair-wise tests, as the latter approach will invariably find some pairwise differences as the number of countries increases. We show that this approach allows for easy extrapolation to new settings to test for external validity.

We first outline the basic CATE method. A simple CATE model would include interactions of the treatment variable (D) with a set of variables, $\tilde{X}$, derived from the baseline covariates. For ease of notation, assume that $\tilde{X}$ includes an intercept. The model would then be:

$$Y_i = D_i \times \tilde{X}_i \beta_{\tilde{X}} + X_i \gamma_X + \varepsilon_i, \tag{3}$$

where $\beta_{\tilde{X}}$ is the vector of the CATE parameters. In order to test if the estimated $\hat{\beta}_{\tilde{X}}$ are different between two settings, we extend Equation 1 and interact the parameters of the model with an indicator of whether an observation is in the Sample:

$$Y_i = (D_i \times \tilde{X}_i \beta_{\tilde{X}} + X_i \gamma_X) + (D_i \times \tilde{X}_i \times \mathbb{1}_{i \in S} \delta_{\tilde{X}} + X_i \times \mathbb{1}_{i \in S} \delta_X) + \varepsilon_i, \tag{4}$$

where $\delta_{\tilde{X}}$ is the vector of coefficients that capture how the CATE parameters differs between subsets. We can then test for generalizability of the CATE parameters by testing whether

the $\hat{\delta}_{\tilde{X}}$ coefficients are jointly zero. Notice that even, though Equation 4 (with coefficient vector $\delta_{\tilde{X}}$) nests Equation 1 (with coefficient δ), if there is a common CATE between settings we can find that $\hat{\delta}$ is statistically significant, but $\hat{\delta}_{\tilde{X}}$, is not statistically significant.

4.1 Simulation Comparison

We motivate the use of the CATE model partially due to potential concerns about high false-positive rates with the reweighting method when the covariate distributions are different. A well known problem with inverse propensity weighting methods is that when probabilities are close to 0 or 1, they can result in biased and variable estimates (Crump et al., 2009). We show how this can result in a higher level of false-positives using a simple simulation. We construct a simple DGP of two countries ("Sample" and "Population") where there is a single covariate which affects both the treatment effect and the probability of being in the sample:

$$\begin{aligned} y_i =& D_i \times X_i\beta_0 + u_i \\ \Pr(D_i = 1) =& 0.5 \\ \Pr(i \in S) =& \text{invlogit}(X_i\theta) \\ X, u \sim& N(0, 1) \\ \beta_0 =& 1 \\ \theta \in& [0, ..., 6]. \end{aligned} \tag{5}$$

We vary θ across our simulations to show how the rate of false positives using each approach changes the covariate distributions become more dissimilar. Results are shown in Figure 2. For each θ, we simulate 10,000 samples, each with $N = 10,000$. We can see that when θ is small (the samples are similar) then both methods have similar low error rates. But as θ increase, the reweighting method does worse even though the CATE method is unaffected.

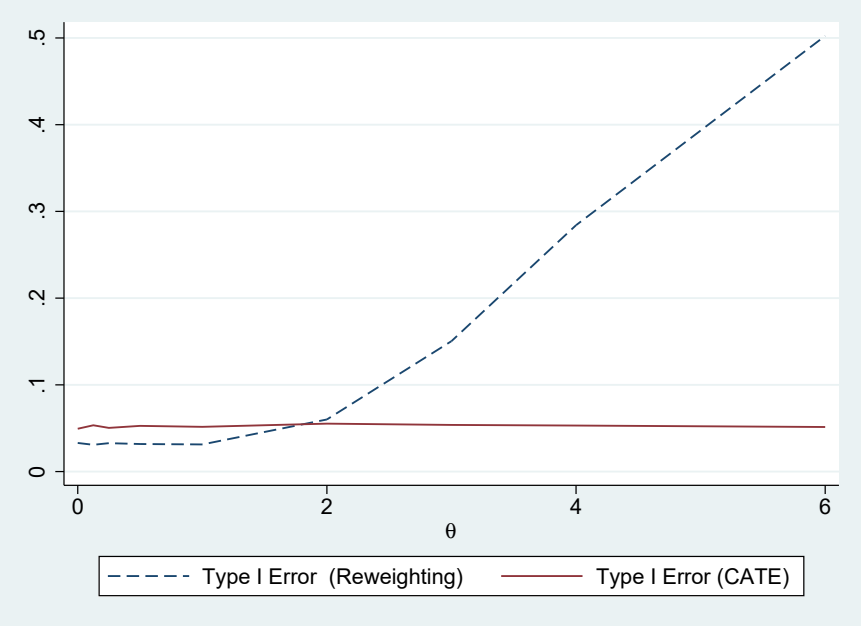

Figure 2: Simulations of False Positives by Estimation Method

Notes: Diagnostics from simulations according to Equation 5. The "Type I Error" plots are for the proportion of simulations where that method found a statistically different treatment effect between the Sample and Population. No data trimming was used for either method.

4

There is a literature that seeks to remedy the bias from extreme propensity scores (Li and Thomas, 2018), but it can be difficult to know which of the various approaches to use, and they tend to make the interpretation of the estimate more opaque. For this reason, and for the benefits for extrapolation which we detail below, we also pursue a CATE-based approach

[4]The reweighting technique's false positive rate for higher values of θ also does not improve when the true propensity scores are used in place of the estimated propensity scores (online Appendix Figure 3).

4.2 Determining the CATE Structure

The previous simulation was simple in that there was only a single covariate used. We did not additionally need to determine what dimensions the CATE parameter vector varies over. In the real world, this structure is unknown and must be estimated as well. To allow for (a) testing for global differences in CATE across settings, and (b) easy extrapolation to new settings, we will focus on selecting a low-dimensional linear model from a high-dimensional set of possibilities, similar to Semenova et al. (2017). This will have the added benefit that the final estimation will be similar to how treatment effect heterogeneity is often estimated in practice.

Though a simple idea would be to include all covariates (and possible transformations), this has the downside that the test for $\hat{\delta}_{\tilde{X}} = \mathbf{0}$ will be less powerful if $\tilde{X}$ includes extraneous variable unrelated to heterogeneity in the CATE. Those variables will tend to be insignificant and weaken the F-test. We therefore develop a machine learning approach to automatically select variables that are important dimensions of heterogeneity of the CATE.

As is common in the causal ML literature (Belloni and Chernozhukov, 2013; Belloni et al., 2014), we will use the Lasso method (Tibshirani, 1996) to select the relevant CATE variables and then estimate the CATE parameters using an OLS regression. The Lasso estimator is used to select the set of variables that together are the most predictive of the outcome variable. It augments the typical OLS objective function so that coefficients are selected to minimize the sum of squared residuals as well the sum of the coefficient sizes. Applied to the OLS objective function for CATE estimation (Equation 3), this is

$$\min_{\beta_{\tilde{X}},\gamma_X} \sum_{i=1}^{N} (Y_i - (D_i \times \tilde{X}_i \beta_{\tilde{X}} + X_i \gamma_X))^2 + \lambda \left(\sum_{\tilde{k} \in \tilde{K}} |\beta_{\tilde{X},\tilde{k}}| + \sum_{k \in K} |\gamma_{X,k}| \right), \tag{6}$$

where K and $\tilde{K}$ is the number of variables in X and $\tilde{X}$, respectively, and λ is a hyper-parameter that controls the level of penalization against complex models. The penalty on the L_1-norm of the coefficients causes some of them to be exactly zero when λ is suffi-

ciently high (unlike the Ridge Regression with an L_2 penalty, which never sets coefficients to exactly zero).[5] "Selected" variables are then those with non-zero Lasso coefficients.

A theoretical motivation for using the Lasso for variable selection is that, if most covariates are truly irrelevant and only a sparse set affects the outcome variable, the Lasso method, under certain conditions, would select the relevant set asymptotically (Zou, 2006). In finite samples, however, it is common for small perturbations in the data to result in the Lasso estimator selecting different subsets of predictors, especially when they are correlated (Mullainathan and Spiess, 2017). If our primary goal were to identify a CATE model for rigorous inspection and independent uses, such as to provide detailed policy recommendations on policy design, then the Lasso method may not be satisfactory. We, however, view the CATE parameter vector as a nuisance parameter in service of the goal of testing external validity. We therefore use the Lasso method merely as a disciplined and automated way to select a set of variables that likely matter to model treatment effect heterogeneity.

We note that the Lasso does not select variables based on statistical significance, but on predictive performance. An example of a variable that highlights this difference, is a binary variable that has a strong effect on the outcome, but is rarely non-zero. Since this variable only helps the prediction of a small number of units, even if it is statistically significant in OLS, it may not be selected by the Lasso.

One point stressed by the literature on using ML for causality (e.g., Chernozhukov et al. 2018) is that taking into account non-linearities can be particularly helpful. We therefore have as our candidate set of variables, all main covariates and their second-order interactions, which results in 72 potential CATE parameters.[6] To keep the set from being

[5]Given the penalization is on the magnitude of the coefficient, the Lasso is not invariant to covariate scaling (unlike OLS). The standard practice is to pre-normalize all covariates to have the same mean and variance.

[6]More generally one could provide a dictionary of high-order transformations. Given our data size, we include just the second-order interactions.

too large, imputation dummy variables and cluster indicators are only used as control variables in the model.

One novel aspect of using a selection algorithm when modelling the CATE is that for every variable we include in the CATE parameter vector, we need to include it as a control. That is, if we model heterogeneity along a particular dimension k, we need to include the pair of regressors X_{ik} and $D_i \times X_{ik}$ in the model so that we can estimate the relative difference for the treatment group. The previous Lasso technique, however, will not necessarily ensure this, as it may select X_{ik}, but drop $D_i \times X_{ik}$. While we could ex-post adjust the set of selected regressors, this is less efficient than including this constraint in the main estimation. A way to model this structure using the general Lasso approach is to use the Group Lasso (Yuan and Lin, 2006), which allows putting coefficients into groups so that entire groups to be either "selected" (all having non-zero coefficients) or "unselected" (all having zero coefficients). If each group has a single member, then this reduces to the normal Lasso. When applied to CATE estimation, each dimension of CATE heterogeneity then would have a group of two elements and covariates that are just controls would be singletons,

$$\min_{\beta_{\tilde{X}},\gamma_X} \sum_{i=1}^{N} (Y_i - (D_i \times \tilde{X}_i \beta_{\tilde{X}} + X_i \gamma_X))^2 + \lambda \left(\sum_{\tilde{k} \in \tilde{K}} \sqrt{\beta_{\tilde{X},\tilde{k}}^2 + \gamma_{\tilde{X},k}^2} + \sum_{k \in K \setminus \tilde{K}} |\gamma_{X,k}| \right). \tag{7}$$

Groups will be selected if together they are predictive. This need to be true for both covariates individually. For example, if $\tilde{X}_k$ is not very predictive, it could still be selected if its accompanying CATE predictor ($D_i \times \tilde{X}_{k,i}$) is very predictive

When using a selection technique, such as the Group Lasso, one must be careful to use the methods on separate data from that used for statistical tests so that the inference can be trusted (Leeb and Pötscher, 2008a,b). Using the ideas from Athey and Imbens (2016), we therefore split our data in "training" and "estimating" halves.[7] We will use the Group

[7]If one is willing to make stronger assumptions on the data generating process, one could use the

Lasso on the training data to select the variables that should be in the CATE and then use the estimating data to estimate the Sample-interacted CATE model and test for differences across the country groups.

The procedure will be most useful when the two halves (train and estimate) have similar distributions to the whole. If they are different, then the Lasso is more likely to select variables that are later unimportant. Splitting data while ensuring similar distributions in the splits is a common concern in RCTs where they assign treatment while often wanting to ensure balance across covariates. We will therefore employ two common methods to ensure similar distributions across the halves: blocking and rerandomization. Blocking partitions the dataset into blocks and ensures a consistent split between training and estimating halves across the blocks. By splitting an important variable into blocks, we can ensure that an even split is achieved at multiple levels of the important variable. Rerandomization conducts multiple randomizations (given constraints such as blocking) and then compares differences in means for important variables between the train and estimate halves. It then picks, as the final randomization, the one that resulted in the smallest maximum t-statistic across the compared variables. Blocking is common with discrete variables (e.g., location), where exact balance can be achieved, and rerandomization can be thought of as an approximation for continuous variables (where we just ensure balance in means). In Machine Learning, data-dependent splitting rules like these are common and improve the similarity between models estimated on a subset and on the whole data (Kohavi, 1995; Diamantidis et al., 2000; Forman and Scholz, 2010).

4.3 Combined Test For Multiple Settings

The final component of our approach is to reframe the test of generalizability to provide a simple result when there are $S > 2$ settings, where S now stands for the number of

Post-Lasso OLS using the whole data as in Belloni and Chernozhukov (2013)

settings. The approach used with reweighting provided S separate pair-wise tests rather than attempting to distill the S results to a single, combined test. A combined test would be useful, especially as S increases.

One approach, mentioned earlier, could be to reject generalizability if any of the S tests showed a statistically significant difference. This has two challenges. First, one needs to adjust for multiple hypothesis testing. This can be done by controlling for the Family-wise error rate, though this is not straight-foward with the reweighting approach, as one would have to account for the correlation among tests as they are estimated on the same data. Second, this approach treats all test equally, even if the number of observations varies across the settings. We believe, in contrast, that statistically significant differences in larger samples are more important that those in smaller samples.

These challenges motivate use to provide a single combined test of generalizability, by expanding Equation 4 to have each setting interacted with the CATE.

$$Y_i = (D_i \times \tilde{X}_i \beta_{\tilde{X}} + X_i \gamma_X) + \sum_{s>1} (D_i \times \tilde{X}_i \times \mathbb{1}_{i \in s} \delta_{s,\tilde{X}} + X_i \times \mathbb{1}_{i \in s} \delta_{s,X}) + \varepsilon_i. \tag{8}$$

We estimate this on the "Estimation" half of the data then conduct a joint test of the combined vector of coefficients $\hat{\delta}_{*,\tilde{X}} = (\hat{\delta}_{2,\tilde{X}}, ..., \hat{\delta}_{S,\tilde{X}})$. This provides a single, simple composite test that naturally adapts to the samples size differences across settings.

4.4 Full Approach

With the components of CATE selection and a combined test for multiple settings, the full approach is shown in Algorithm 1. Note that for any particular split of the data into training and estimating halves, the Group Lasso does not affect the distribution of the F-statistic as the selection is conducted on separate data.[8]

[8]In randomized experiments, as compared to here, adjustments must be made for analyses that use blocking (simply adding block dummies) and rerandomization (more complicated) because the model is

We use Algorithm 1 to estimate EV for each configuration of Sample and Population countries. We use as blocking variables the product of "treatment x cluster" (which are subnational). This ensures that both train and estimate halves of the data have treatment and control observations from every cluster, ensuring that the treatment effect estimated from each is suitably representative. We also use 100 rerandomizations comparing across the outcome and main covariates.

Algorithm 1 CATE Estimation and Test of EV

1. Split the data into training and estimating halves using tools that balance covariates. First block on any blocking variables, and then use R rerandomizations to pick the split that has the smallest maximum t-statistic over the variables to be compared.

2. Using the training portion of the data, fit a Group Lasso model of CATE (Equation 7) where the full set of CATE terms, $\tilde{X}$, includes all second-order interactions of the main covariates. We set the Lasso regularization parameter, λ, to minimize 10-fold cross-validation error. Call the subset of $\tilde{X}$ selected by the Group Lasso $\tilde{X}^*$.

3. The CATE model can be estimated by using the estimating portion of the data and the variables selected by the Group Lasso. (Used in Algorithm 2.)

4. Using the estimating portion of the data, estimate a Setting-interacted CATE model as in Equation 8 using the variables selected by the Group Lasso yielding $\hat{\delta}_{*,\tilde{X}}$.

5. Use an F-statistic to test if the $\hat{\delta}_{*,\tilde{X}}$ vector of coefficients is jointly different from zero.

We show $\hat{\delta}_{\tilde{X}}$ from the setting-interacted CATE model in Table 6. This includes all CATE variables selected by the Group Lasso. We do not reject the joint test that the estimated

estimated using both halves of the data as the split defines the treatment variable. Here, these procedures aim to improve how well the test statistic represents the result if it could have been conducted on the whole sample.

conditional average treatment effects are different across the countries ($p > 0.05$). We take this as evidence of the generalizability of the treatment effect (in the presence of covariate differences and treatment effect heterogeneity). As we treat the selected CATE models as nuisance parameters, we do not inspect them directly. We do see, though, that the size of the CATE model is much smaller than 72.

As the CATE procedure estimates treatment effect differences excluding the training data, we want to ensure that results have not changed simply because of the reduction in sample size. We therefore replicate the previous treatment effect approaches (the simple ATE comparison and the reweighting approaching) using the same subsample in online Appendix Tables 7 and 8. They are qualitatively similar. In the simple ATE comparison, two of the configurations had statistically significant differences in the ATE at $p < 0.05$. In the reweighting approach, one configuration had statistically different outcomes for control units and another had statistically different outcomes for the treated units.

4.5 Extrapolation

One benefit of constructing a regression-based CATE model is that we can now easily provide a method to assess external validity in new settings, even in the presence of treatment effect heterogeneity and differing covariate distributions. With the reweighting approach, to assess external validity on a new setting, one needs access to the original data in order to estimate the Sample-prediction model (Equation 2) to derive the weights. Our CATE-based approach avoids this; all that is needed are the estimates of the CATE model.

For our data, we now consider all three countries as the Sample (any new setting would be the Population) and conduct steps 1-3 of Algorithm 1. This yields the selected CATE variables, $\tilde{X}^*$ and Sample CATE estimates $\hat{\beta}_{S,\tilde{X}^*}$ and the associate sub-matrix $\hat{V}_{S,\beta}$ of the overall estimator variance-covariance matrix. These are show in online Appendix Tables 10 and 11.

Table 6: Setting-interacted CATE

	(1) UY	(2) ES offset	(3) MX offset
Treatment	-0.546	0.612	0.820
	(0.419)	(0.598)	(0.570)
Treatment x Head of HH Educ.	0.0874**	-0.00557	-0.0580
	(0.0352)	(0.0525)	(0.0464)
Treatment x Head of HH Female	0.0540	-0.182	-0.100
	(0.273)	(0.445)	(0.346)
Treatment x Head of HH Age	0.0111	0.00521	-0.0109
	(0.00831)	(0.0110)	(0.0110)
Treatment x Z-score Satisfaction (Baseline)	-0.0839	-0.0198	0.0503
	(0.0615)	(0.0936)	(0.0724)
Treatment x HH Asset value/capita Sq.	0.00000420	-0.00000361	-0.00000281
	(0.00000481)	(0.00000499)	(0.00000519)
Treatment x Head of HH Educ. x Z-score Satisfaction (Baseline)	0.0116	-0.0109	-0.0105
	(0.00852)	(0.0118)	(0.0106)
Treatment x Head of HH Female x HH Asset value/capita	0.0000221	-0.000210	0.00112
	(0.00242)	(0.00278)	(0.00296)
Treatment x Head of HH Age x Z-score Housing quality (Baseline)	0.0000367	0.000394	0.000221
	(0.00166)	(0.00242)	(0.00202)
Treatment x HH Asset value/capita x Z-score Housing quality (Baseline)	-0.000708	0.000991	0.000123
	(0.000454)	(0.000623)	(0.000591)
Treatment x HH Asset value/capita x Z-score Housing investment (Baseline)	-0.000545	-0.000126	-0.0000767
	(0.000383)	(0.000625)	(0.000554)
Treatment x HH Income/capita x Z-score Housing quality (Baseline)	0.0000724	-0.00218	-0.00110
	(0.000579)	(0.00134)	(0.000910)
Treatment x HH Income/capita x Z-score Satisfaction (Baseline)	-0.0000880	0.00335**	0.000297
	(0.000438)	(0.00166)	(0.000543)
Treatment x Z-score Housing quality (Baseline) x Z-score Housing investment (Bas	0.00749	0.0207	0.0165
	(0.0216)	(0.0386)	(0.0311)
Treatment x Z-score Housing investment (Baseline) x Head of HH Educ.	-0.0198**	0.0302	0.0200
	(0.0100)	(0.0203)	(0.0177)
Treatment x Z-score Housing investment (Baseline) x Head of HH Female	0.104	-0.0216	-0.0236
	(0.0807)	(0.137)	(0.114)
Treatment x Z-score Housing investment (Baseline) x Z-score Satisfaction (Baseli	-0.00205	0.00833	-0.0139
	(0.0149)	(0.0241)	(0.0183)
Observations	905	905	905
R^2	0.339	0.339	0.339
p-val no CATE difference	0.243		0.243

Outcome is Satisfaction Index and statistics are coefficient and (standard error). Test-sample only.

Omitting non-sample-interacted coefficients.

* (p <0.10), ** (p <0.05), *** (p <0.01)

Notes: The three columns report coefficients from a single (setting-interacted CATE) model. Only CATE coefficients are shown. The UY column contains the base coefficients and the ES and MX offset columns report the coefficient for those same variables interacted with dummy variables for whether the observation was in that country.

On a new setting (Population) one can then use Algorithm 2 to extrapolate what the ATE would be expected and test whether the original treatment effect generalizes:

Algorithm 2 ATE Extrapolation and Test of Generalization

1. Calculate the Population's average values for $\tilde{X}_P^*$. Call this r_P.
2. The estimate of the ATE using extrapolation in the Population is then $\hat{\beta}_{P,Ext} = r'_P \hat{\beta}_{S,\tilde{X}^*}$.
3. Construct confidence intervals for this new ATE using the standard Wald test for linear combinations of coefficients, $Var(\hat{\beta}_{P,Ext}) = r'_P \hat{V}_{S,\beta} r_P$.
4. Estimate the ATE in the Population directly, $\hat{\beta}_{P,Dir}$.
5. If $\hat{\beta}_{P,Dir}$ is outside the confidence interval $\hat{\beta}_{P,Ext}$, then this implies a failure of generalization in this case.

4.6 Effect of Trimming

We note that the preceding treatment effect estimations were conditional on trimming observations that had values of key covariates outside the bounds of the countries. We check if our results are robust to inclusion of these observations in two ways. For both checks, we will need to compare estimated coefficient vectors across sample trimming methods. We therefore hold fixed the selected CATE variables and consider the initially trimmed observations as part of the "estimation" subset of the data. First, we check if the estimated CATE coefficients $\hat{\beta}_{\tilde{X}}$ pooling all three countries changes with the inclusion of the initially trimmed observations. A joint test of the difference in coefficients yields a p-value of 0.89. Second, we check if Algorithm 1 still yields an insignificant result from the joint test of the sample-interacted CATE coefficients. Results are shown in online Appendix Table 9, where

we see that we still do not reject that overall country-specific CATE changes across countries are zero. Given this, we conclude that the effects in Galiani et al. (2017) generalizes, regardless of sample trimming.

5 Conclusion

In this paper, we evaluate various strategies for practically assessing external validity (EV) of treatment effects estimates. We evaluate if the various methods provide (a) a simple statistical test of whether the effect generalizes across the available settings (and does this extend to more than two settings), and (b) how easily can results be extrapolated to predict the expected treatment effect in a new setting (e.g., *does one need the full original data?*, and *will extrapolating to a new domain require changing the interpretation of the estimate in a complicated way?*). We then apply the various methods to data from an RCT that was conducted across three countries (Galiani et al., 2017). This study found a strong and statistically significant effect across all three countries of housing upgrades on a summary index of respondent's satisfaction with their housing situation.

We evaluate two ways of assessing if there is a common conditional average treatment effect (CATE) coupled with changes in the covariate distribution, or if the treatment effects fundamentally differ. Results from the reweighting procedure of Hotz et al. (2005) suggests that the treatment effects do differ across countries. We show that this procedure can yield false-positives in the presence of covariate differences, which we have in this data.

To address this short-coming, we provide a method that allows for modelling the CATE directly. To allow for a tractable regression-based model that can be used for statistical tests, we develop a new machine-learning (ML) based method that uses the Group Lasso (Yuan and Lin, 2006) to select a possibly non-linear the CATE model. We note that, while we model the CATE directly, we view it as a nuisance parameter in the service of testing for

external validity. We do not, therefore, need to estimate the true CATE, just a reasonable approximation, which is what the ML algorithm allows us to do.

When we apply our procedure to the data and test for differences in the CATE across countries, the results are no longer statistically different, indicating that the procedure was able to find a common treatment effect in the presences of covariate differences. We then show that this regression-based CATE model allows researchers in new settings to predict the treatment effect and confidence intervals in a new setting without access to the original data.

The authors report there are no competing interests to declare.

References

Andrews, I. and E. Oster (2019, May). A simple approximation for evaluating external validity bias. *Economics Letters 178*, 58–62.

Angrist, J. D. (2004, March). Treatment effect heterogeneity in theory and practice. *The Economic Journal 114* (494), C52–C83.

Athey, S. and G. Imbens (2016, July). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113* (27), 7353–7360.

Athey, S. and G. W. Imbens (2017). The econometrics of randomized experiments. In *Handbook of Field Experiments*, pp. 73–140. Elsevier.

Attanasio, O., A. Kugler, and C. Meghir (2011, July). Subsidizing vocational training for disadvantaged youth in Colombia: Evidence from a randomized trial. *American Economic Journal: Applied Economics 3* (3), 188–220.

Belloni, A. and V. Chernozhukov (2013, May). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2).

Belloni, A., V. Chernozhukov, and C. Hansen (2014, April). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Bloom, N., J. Liang, J. Roberts, and Z. J. Ying (2014, November). Does working from home work? Evidence from a Chinese experiment. *The Quarterly Journal of Economics 130*(1), 165–218.

Bo, H. and S. Galiani (2021, September). Assessing external validity. *Research in Economics 75*(3), 274–285.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin 54*(4), 297–312.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, January). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2018, June). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009, January). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*(1), 187–199.

Dehejia, R., C. Pop-Eleches, and C. Samii (2019, August). From local to global: External validity in a fertility natural experiment. *Journal of Business & Economic Statistics 39*(1), 217–243.

Diamantidis, N., D. Karlis, and E. Giakoumakis (2000, January). Unsupervised stratification of cross-validation for accuracy estimation. *Artificial Intelligence 116*(1-2), 1–16.

Forman, G. and M. Scholz (2010, November). Apples-to-apples in cross-validation studies. *ACM SIGKDD Explorations Newsletter 12*(1), 49–57.

Galiani, S., P. J. Gertler, R. Undurraga, R. Cooper, S. Martínez, and A. Ross (2017, March). Shelter from the storm: Upgrading housing infrastructure in Latin American slums. *Journal of Urban Economics 98*, 187–213.

Gechter, M. (2021, October). Generalizing the results from social experiments: Theory and evidence from Mexico and India. Mimeo.

Hotz, V. J., G. W. Imbens, and J. H. Mortimer (2005, March). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics 125*(1-2), 241–270.

Kern, H. L., E. A. Stuart, J. Hill, and D. P. Green (2016, January). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness 9*(1), 103–127.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, San Francisco, CA, USA, pp. 1137–1143. Morgan Kaufmann Publishers Inc.

Leeb, H. and B. M. Pötscher (2008a, April). Can one estimate the unconditional distribution of post–model–selection estimators? *Econometric Theory 24*(02).

Leeb, H. and B. M. Pötscher (2008b, April). Recent developments in model selection and related areas. *Econometric Theory 24*(02).

Li, F. and L. E. Thomas (2018, September). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*.

List, J. (2020, July). Non est disputandum de generalizability? A glimpse into the external validity trial. Technical report, National Bureau of Economic Research.

Meager, R. (2019, January). Understanding the average impact of microcredit expansions: A Bayesian Hierarchical Analysis of seven randomized experiments. *American Economic Journal: Applied Economics 11*(1), 57–91.

Mullainathan, S. and J. Spiess (2017, May). Machine learning: An applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106.

Muralidharan, K., A. Singh, and A. J. Ganimian (2019, April). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review 109*(4), 1426–1460.

Nguyen, T. Q., C. Ebnesajjad, S. R. Cole, and E. A. Stuart (2017, March). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics 11*(1).

Nie, X. and S. Wager (2020, September). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika 108*(2), 299–319.

Pritchett, L. and J. Sandefur (2014, January). Context matters for size: Why external

validity claims and development practice do not mix. *Journal of Globalization and Development 4*(2).

Rubin, D. B. (1974, October). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Semenova, V., M. Goldman, V. Chernozhukov, and M. Taddy (2017). Estimation and inference on heterogeneous treatment effects in high-dimensional dynamic panels.

Stuart, E. A., S. R. Cole, C. P. Bradshaw, and P. J. Leaf (2011, April). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 174*(2), 369–386.

Taddy, M. (2019). *Business Data Science*. McGraw-Hill Education Ltd.

Tibshirani, R. (1996, January). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Vivalt, E. (2020, September). How much can we generalize from impact evaluations? *Journal of the European Economic Association 18*(6), 3045–3089.

Wager, S. and S. Athey (2018, June). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Yuan, M. and Y. Lin (2006, February). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.

Zou, H. (2006, December). The Adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

A Online Appendix

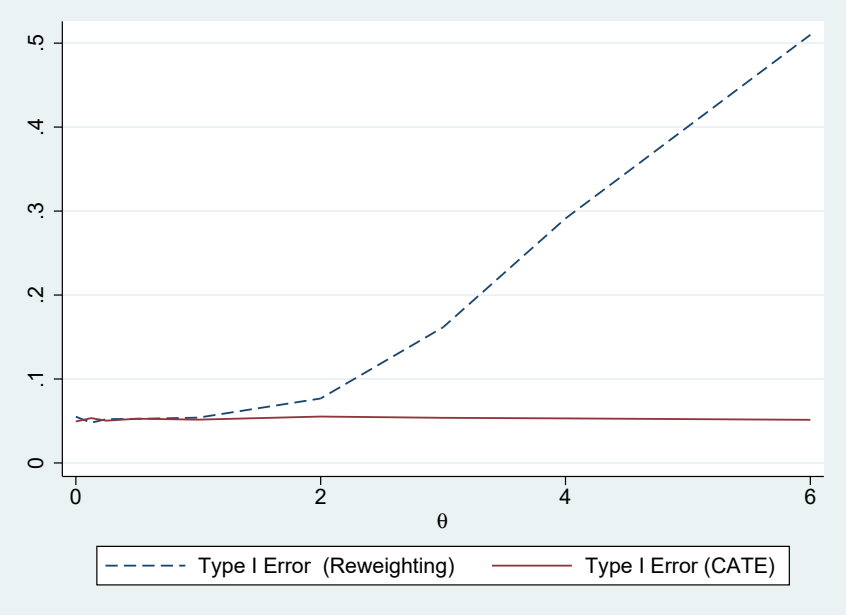

Figure 3: Simulations of False Positives by Estimation Method (Using True Propensity)

Notes: Diagnostics from simulations according to Equation 5. The "Type I Error" plots are for the proportion of simulations where that method found a statistically different treatment effect between the Sample and Population. No sample trimming was used for either method. In this figure the reweighting method uses the estimated propensity score rather than the true one.

Table 7: ATE Sample-interacted (Excluding Training Data)

	(1)
Treatment × Is ES	0.689**
	(0.178)
Treatment × Is MX	0.116
	(0.128)
Observations	905
R^2	0.251
p-val no ATE difference	0.000361

Outcome is Satisfaction Index and and statistics are coefficient and (standard error). Excludes training data.

Omitting non-sample-interacted coefficients.

* (p <0.10), ** (p <0.05), *** (p <0.01)

Table 8: Reweighted Pooled Treatment Effect Estimation (Excluding Training Data)

	(1)	(2)	(3)
Treatment	1.111**	0.326**	0.354**
	(0.160)	(0.105)	(0.0625)
Is Sample	0.154	-0.00806	0.956**
	(0.271)	(0.782)	(0.386)
Is Sample x Treatment	-0.764**	0.260**	0.158
	(0.174)	(0.129)	(0.108)
Observations	1312	1295	1358
Sample	UY+MX	ES+MX	ES+UY
Population	ES	UY	MX

Outcome is Satisfaction Index and statistics are coefficient and (standard error). Models include baseline controls.

Propensity estimation done by logit.

Exludes Training data.

* (p <0.10), ** (p <0.05), *** (p <0.01)

Table 9: Sample-interacted CATE (Including Initially Trimmed Observations)

	(1) UY	(2) ES offset	(3) MX offset
Treatment	-0.228	0.576	0.382
	(0.333)	(0.505)	(0.455)
Treatment x Head of HH Educ.	0.0557*	0.00230	-0.0261
	(0.0311)	(0.0492)	(0.0400)
Treatment x Head of HH Female	0.154	0.0152	-0.107
	(0.247)	(0.379)	(0.302)
Treatment x Head of HH Age	0.00878	0.00328	-0.00896
	(0.00662)	(0.00940)	(0.00856)
Treatment x Z-score Satisfaction (Baseline)	-0.0809	0.00661	0.0754
	(0.0563)	(0.0835)	(0.0637)
Treatment x HH Asset value/capita Sq.	0.00000289	-0.00000261	-0.00000316
	(0.00000375)	(0.00000377)	(0.00000378)
Treatment x Head of HH Educ. x Z-score Satisfaction (Baseline)	0.0122	-0.0114	-0.0125
	(0.00768)	(0.0114)	(0.00930)
Treatment x Head of HH Female x HH Asset value/capita	-0.000905	0.000425	0.00251
	(0.00210)	(0.00223)	(0.00228)
Treatment x Head of HH Age x Z-score Housing quality (Baseline)	-0.000216	0.00169	0.000490
	(0.00131)	(0.00187)	(0.00150)
Treatment x HH Asset value/capita x Z-score Housing quality (Baseline)	-0.000593	0.000717	0.000213
	(0.000370)	(0.000475)	(0.000410)
Treatment x HH Asset value/capita x Z-score Housing investment (Baseline)	0.0000260	-0.000362	0.0000214
	(0.000260)	(0.000409)	(0.000354)
Treatment x HH Income/capita x Z-score Housing quality (Baseline)	-0.0000121	-0.00140*	-0.000462*
	(0.000146)	(0.000783)	(0.000260)
Treatment x HH Income/capita x Z-score Satisfaction (Baseline)	-0.0000985	0.00181	0.000254
	(0.000133)	(0.00112)	(0.000235)
Treatment x Z-score Housing quality (Baseline) x Z-score Housing investment (Bas	0.0113	0.0273	-0.0177
	(0.0117)	(0.0193)	(0.0167)
Treatment x Z-score Housing investment (Baseline) x Head of HH Educ.	-0.0141**	0.0265**	0.0114
	(0.00631)	(0.0127)	(0.0118)
Treatment x Z-score Housing investment (Baseline) x Head of HH Female	0.0176	0.0710	0.0632
	(0.0559)	(0.0911)	(0.0762)
Treatment x Z-score Housing investment (Baseline) x Z-score Satisfaction (Baseli	-0.00736	0.00663	-0.00246
	(0.00677)	(0.0135)	(0.0107)
Observations	1291	1291	1291
R^2	0.321	0.321	0.321
p-val no CATE difference	0.243		0.243

Outcome is Satisfaction Index and statistics are coefficient and (standard error). Test-sample only.

Omitting non-sample-interacted coefficients.

* (p <0.10), ** (p <0.05), *** (p <0.01)

The three columns report coefficients from a single (setting-interacted CATE) model. Only CATE coefficients are shown. The UY column contains the base coefficients and the ES and MX offset columns report the coefficient for those same variables interacted with dummy variables for whether the observation was in that country.

Table 10: CATE Coefficients

	(1)
Treatment	-0.1055
Treatment x Head of HH Educ.	0.03306
Treatment x Head of HH Female	0.1399
Treatment x Head of HH Age	0.008237
Treatment x Z-score Satisfaction (Baseline)	-0.02852
Treatment x HH Asset value/capita Sq.	2.759e-07
Treatment x Head of HH Educ. x Z-score Satisfaction (Baseline)	-0.0007141
Treatment x Head of HH Female x HH Asset value/capita	0.0006955
Treatment x Head of HH Age x Z-score Housing quality (Baseline)	-0.001679
Treatment x HH Asset value/capita x Z-score Housing quality (Baseline)	-0.0003103
Treatment x HH Asset value/capita x Z-score Housing investment (Baseline)	-0.0002562
Treatment x HH Income/capita x Z-score Housing quality (Baseline)	0.0001504
Treatment x HH Income/capita x Z-score Satisfaction (Baseline)	0.0001566
Treatment x Z-score Housing quality (Baseline) x Z-score Housing investment (Bas	0.01394
Treatment x Z-score Housing investment (Baseline) x Head of HH Educ.	-0.01247
Treatment x Z-score Housing investment (Baseline) x Head of HH Female	0.07701
Treatment x Z-score Housing investment (Baseline) x Z-score Satisfaction (Baseli	0.001989
Observations	905

Outcome is Satisfaction Index. Stats=b. Test-sample only.

Variables align with those in Table 11.

Table 11: CATE Variance-Covariance Sub-Matrix

V_complex_lmt	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	.0527297	-.0023474	-.0125147	-.0007591	.0001889	-1.30e-08	-.0000424	.0000143	-1.95e-06	-1.20e-07	-2.76e-06	-3.08e-06	-5.15e-06	-.0002755	.0000327	.0003476	.0000456
2	-.0023474	.000321	-.0001757	.0000268	-.0000413	9.78e-10	.0000101	-4.19e-07	-4.27e-07	-7.07e-08	1.83e-07	2.74e-07	5.39e-07	.0000182	-.000032	.0000669	-6.26e-06
3	-.0125147	-.0001757	.0168428	.0000519	.0001028	3.20e-08	-.0000253	-.0000385	-5.15e-06	1.03e-06	-1.67e-06	2.39e-06	4.34e-07	.0000476	.0001501	-.0007106	-.0000375
4	-.0007591	.0000268	.0000519	.0000155	-4.53e-06	6.94e-11	1.05e-06	-1.93e-07	1.14e-07	-1.55e-08	4.50e-08	1.97e-08	3.38e-08	9.45e-07	-1.52e-06	6.36e-06	2.99e-07
5	.0001889	-.0000413	.0001028	-4.53e-06	.0007343	1.05e-09	-.0000783	-7.83e-07	1.35e-06	2.44e-07	-4.74e-07	-1.24e-06	-4.05e-06	.000018	7.62e-06	.000027	.0000346
6	-1.30e-08	9.78e-10	3.20e-08	6.94e-11	1.05e-09	8.01e-13	-1.98e-10	-7.26e-10	-4.88e-11	2.39e-11	-3.77e-11	2.11e-11	4.07e-12	-1.41e-09	2.55e-10	2.61e-09	-6.99e-10
7	-.0000424	.0000101	-.0000253	1.05e-06	-.0000783	-1.98e-10	.0000178	1.86e-07	-9.16e-08	-4.69e-08	1.03e-07	2.03e-08	1.56e-07	-1.21e-06	-1.78e-06	4.19e-06	-6.06e-06
8	.0000143	-4.19e-07	-.0000385	-1.93e-07	-7.83e-07	-7.26e-10	1.86e-07	8.05e-07	3.37e-08	-8.83e-09	4.20e-08	-2.23e-08	-3.88e-09	1.17e-06	-3.28e-07	-2.64e-06	5.31e-07
9	-1.95e-06	-4.27e-07	-5.15e-06	1.14e-07	1.35e-06	-4.88e-11	-9.16e-08	3.37e-08	4.25e-07	-3.79e-08	8.16e-09	-1.04e-07	-1.42e-08	2.11e-07	-1.92e-07	-4.17e-07	3.72e-07
10	-1.20e-07	-7.07e-08	1.03e-06	-1.55e-08	2.44e-07	2.39e-11	-4.69e-08	-8.83e-09	-3.79e-08	4.60e-08	-8.26e-09	-9.49e-09	-1.46e-09	-8.25e-08	1.28e-07	1.70e-07	-4.97e-08
11	-2.76e-06	1.83e-07	-1.67e-06	4.50e-08	-4.74e-07	-3.77e-11	1.03e-07	4.20e-08	8.16e-09	-8.26e-09	3.79e-08	-2.51e-09	-2.40e-10	2.01e-07	-2.00e-07	-1.15e-06	-1.96e-07
12	-3.08e-06	2.74e-07	2.39e-06	1.97e-08	-1.24e-06	2.11e-11	2.03e-08	-2.23e-08	-1.04e-07	-9.49e-09	-2.51e-09	1.24e-07	1.48e-08	-6.13e-07	8.47e-09	4.65e-09	-8.62e-08
13	-5.15e-06	5.39e-07	4.34e-07	3.38e-08	-4.05e-06	4.07e-12	1.56e-07	-3.88e-09	-1.42e-08	-1.46e-09	-2.40e-10	1.48e-08	6.33e-08	-8.29e-08	4.04e-09	-8.12e-07	-3.64e-07
14	-.0002755	.0000182	.0000476	9.45e-07	.000018	-1.41e-09	-1.21e-06	1.17e-06	2.11e-07	-8.25e-08	2.01e-07	-6.13e-07	-8.29e-08	.0001199	-.0000121	-6.59e-06	-1.27e-06
15	.0000327	-.000032	.0001501	-1.52e-06	7.62e-06	2.55e-10	-1.78e-06	-3.28e-07	-1.92e-07	1.28e-07	-2.00e-07	8.47e-09	4.04e-09	-.0000121	.0000343	-.0001466	3.10e-06
16	.0003476	.0000669	-.0007106	6.36e-06	.000027	2.61e-09	4.19e-06	-2.64e-06	-4.17e-07	1.70e-07	-1.15e-06	4.65e-09	-8.12e-07	-6.59e-06	-.0001466	.0016716	2.12e-06
17	.0000456	-6.26e-06	-.0000375	2.99e-07	.0000346	-6.99e-10	-6.06e-06	5.31e-07	3.72e-07	-4.97e-08	-1.96e-07	-8.62e-08	-3.64e-07	-1.27e-06	3.10e-06	2.12e-06	.0000555

Notes: Variables align with those in Table 10.