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

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

ABSTRACT

We review, from a practical standpoint, the evolving literature on assessing external validity (EV) of estimated treatment effects. We provide an implementation and real-world assessment of the general EV measures developed in Bo and Galiani (2021). In the context of estimating conditional average treatment effect models for assessing external validity, we provide a novel method utilizing the Group Lasso (Yuan and Lin, 2006) to estimate a tractable regression-based model. 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 conducted 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

Assessing External Validity in Practice

Sebastian Galiani
University of Maryland and NBER

Brian Quistorff*
Bureau of Economic Analysis

August 18, 2022

Abstract

We review, from a practical standpoint, the evolving literature on assessing external validity (EV) of estimated treatment effects. We provide an implementation and real-world assessment of the general EV measures developed in Bo and Galiani (2021). In the context of estimating conditional average treatment effect models for assessing external validity, we provide a novel method utilizing the Group Lasso (Yuan and Lin, 2006) to estimate a tractable regression-based model. 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 conducted in three different countries (Galiani et al., 2017).

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, both those that focus solely on a single setting and those that compare across settings. We use both types of 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.

Single-setting measures of external validity were proposed by Bo and Galiani (2021). They provide a theoretical treatment of external validity 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

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

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 the degree of EV changes with the correlation between the reweighting vector and the pair-level outcome differences. This measure of EV is motivated by their definition of external validity from one population onto other, 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.

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 treatments effect can be derived (Nguyen et al., 2017; Andrews and Oster, 2019; Gechter, 2021).

While some information can be gathered from a single dataset, ultimately, EV is established by replicating the same experiments in different populations (Angrist, 2004; List, 2020). We therefore also provide practical guidance in the multi-setting case.

Some papers approach the issue more 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.

Formal measures of external validity using two settings were developed by Hotz et al. (2005). The main challenge is that the two settings may be different and the observable characteristics may not be sufficient for adjustment (there could be important unobservable differences, including, for instance, macro-level factors). 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 the propensity of being in 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 pool data from the two settings, fit a propensity score model to account whether an observation is in the Sample conditional on a set of covariates, trim the data to ensure overlap, and then construct unit-level inverse-propensity weights. They then assess whether observable characteristics are sufficient to make the data comparable by checking if there are statistically significant differences in the weighted outcomes of the control units between the two settings. If there is not, they assess whether the treatment effect in the inference sample generalizes to the target population by testing for statistically significant differences in the treatment unit outcomes between settings.

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 sub-classification. 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.

Aside from reweighting, other work has built models of the conditional average treatment effect (CATE) from subsets of the data and then calculated what error would result in extrapolating that to new settings (Kern et al., 2016; Dehejia et al., 2019; Pritchett and Sandefur, 2014). 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.

This paper proceeds as follows. In Section 2 we investigate single-setting measures. We provide an implementation of the single-setting measures of Bo and Galiani (2021). We apply their measures to the case of a three-country randomized control trial on the effect of upgrading slum dwellings (Galiani et al., 2017). In Section 3 we then consider analyses that split the data into the Sample and the Population groups. We first assess external validity using the propensity-score reweighting methods of Hotz et al. (2005). Finally we assess external validity by

Table 1: Satisfaction

	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	1.031**	0.898**	0.317**	0.273**	0.295**	0.264**
	(0.0866)	(0.128)	(0.0618)	(0.103)	(0.0519)	(0.0686)
	[0.861,1.201]	[0.648,1.148]	[0.196,0.439]	[0.0715,0.474]	[0.193,0.397]	[0.130,0.398]
Observations	656	478	718	630	826	642
Country	ES	ES	UY	UY	MX	MX
Estimation	OLS	Matching	OLS	Matching	OLS	Matching

Outcome is Satisfaction Index and statistics are coefficient, (standard error), and [confidence interval].
Table shows the variation in treatment effect from the original OLS models of Galiani et al. (2017) to the matching estimators needed for EV reweighting. OLS models include baseline controls from Galiani et al. (2017).
Matching estimator is one-to-one (with replacement) bias-corrected and using the same baseline variables.
* (p <0.10), ** (p <0.05), *** (p <0.01)

modeling the CATE. The existing external validity CATE literature focuses on estimating the CATE using a subset of the settings and evaluating the mean-squared error that would result from predicting the CATE on the held-out settings versus estimating the CATE directly on the held-out setting. We, by contrast, focus on building a CATE model where we can construct a hypothesis test for whether it generalizes across settings. To this end, we provide a method using machine learning (ML) to select the components of the CATE model (the dimensions along which we will estimate heterogeneity) that will later be estimated. We then show that this method allows easy extrapolation for forecasting the treatment effect in other, not studied, populations. Section 4 concludes.

2 Single-setting measures

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. We then describe the single-setting measures of Bo and Galiani (2021) and apply them to the housing example.

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.

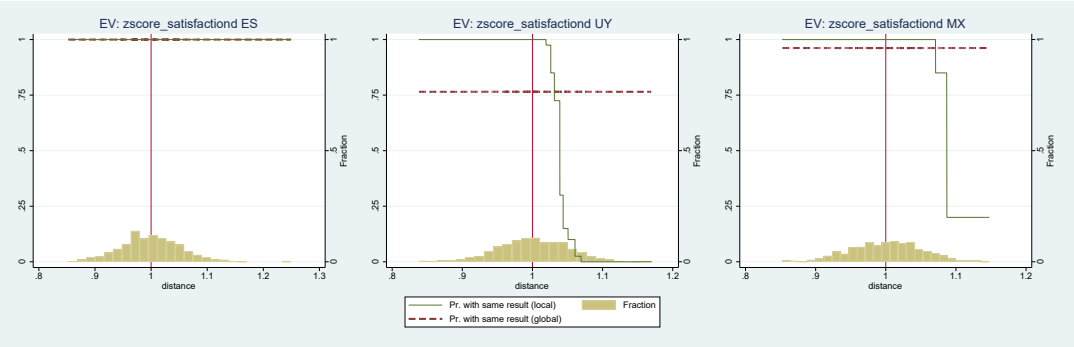
We focus on their main outcome, the "Satisfaction" Index. This is an aggregate index that summarizes several satisfaction sub-measures: Satisfaction with Floor, Wall Quality, Roof Quality, House Protection against Water when it rains, and Qualify of Life. Each of those measures is turned into a Z-score, 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.

As the estimation methods of Bo and Galiani (2021) are based on 1-to-1 matching, we first replicate, for each country, the original estimates of Galiani et al. (2017). We next construct the analogous 1-to-1 matching estimation of Abadie and Imbens (2011) that ensures exact matches on cluster and then matches and bias-corrects for the remaining controls. Results are shown in Table 1. The estimated effect is positive and statistically significant across all three countries for both estimation approaches. Overall, the matching estimates tend to be slightly smaller and less precisely estimated.

We next assess the external validity using the techniques from Bo and Galiani (2021). [1]

[1]Stata package available at `https://github.com/bquistorff/ExternalValidity` and

Figure 1: Satisfaction - EV curve



Plots are of the proportion of reweightings that have the same statistical significance category as well as the density of the reweightings. They are aligned according to the "distance", which is defined by Bo and Galiani (2021) as one minus the correlation of the treated-control outcome difference vector and the reweighting vector.

Table 2: Satisfaction (Matching EV estimates)

	(1)	(2)	(3)
Treatment	0.898**	0.273**	0.264**
	(0.128)	(0.103)	(0.0686)
	[0.648,1.148]	[0.0715,0.474]	[0.130,0.398]
Observations	478	630	642
Country	ES	UY	MX
Pr. Same Stat. Sig.	1	0.765	0.962

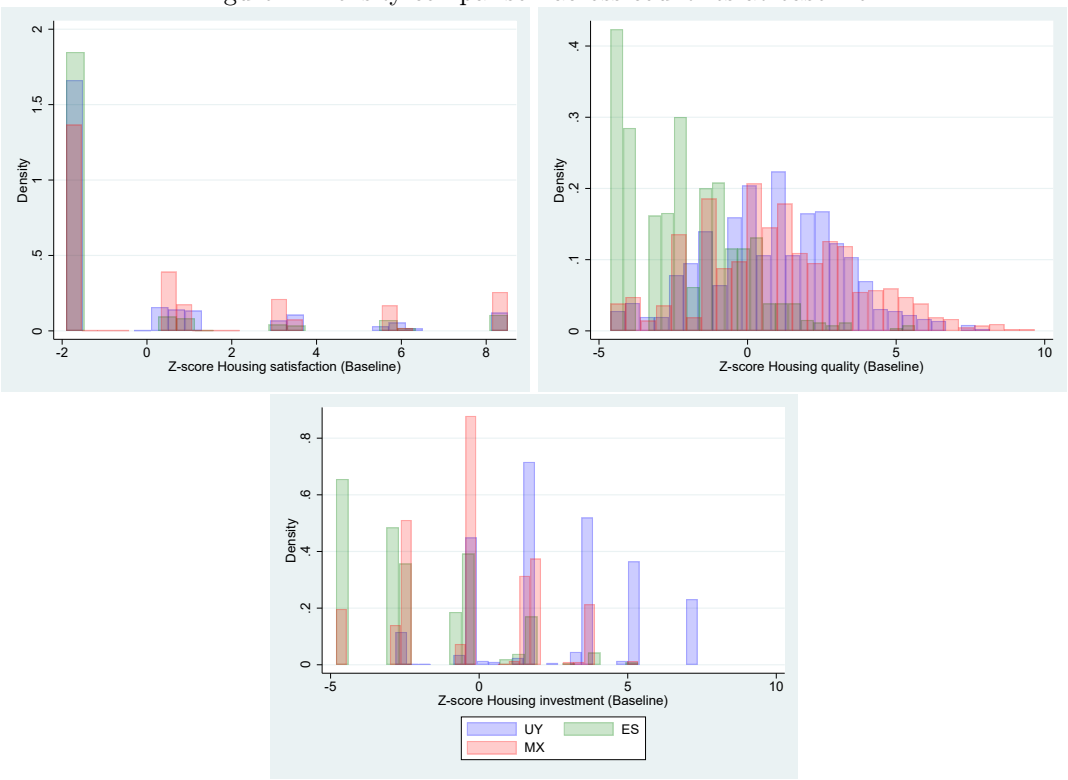
Outcome is Satisfaction Index and statistics are coefficient, (standard error), and [confidence interval]
* (p <0.10), ** (p <0.05), *** (p <0.01)

After constructing the 1-to-1 matches, 1000 reweighting vectors for the treated-control pairs are drawn uniformly from the distribution of all possible reweightings and the matching estimate is re-calculated for each. If all the pair-level outcome differences are quite similar then the reweighting estimates will be very similar to the original estimate. To the extent that there is variation, the effect will change with how the reweighting vector modulates heterogeneity. We then calculate the proportion of reweightings that have the same category of statistical significance. The overall proportion is listed in Table 2. Practically all reweightings for the two countries with the higher effect t-statistic (El Salvador and Mexico) have the same statistical significance class whereas for Uruguay it is still high (0.765), but smaller.

One can also view how the EV measure changes "locally" as the reweighting vector becomes more or less "favorable". This is done by calculating a "distance" variable, defined in Bo and Galiani (2021) as one minus the correlation between the reweighting vector and pair-level treated minus control outcome difference vector. The domain is $[0, 2]$, where a low value indicates reweightings that increase the estimated treatment effect, a value of 1 indicates reweightings similar to the unweighted estimate, and a high value indicates reweightings that decrease the treatment effect. We graph how the EV measure changes locally with distance in Figure 1. Since the estimated treatment effects are positive, the EV measure will be high for low distances and may decrease at higher distances. The local measure decreases faster for Uruguay as its initial estimate had the lowest t-statistic.

`http://www.sebastiangaliani.com/`.

Figure 2: Density comparison across countries at baseline



3 Multi-setting measures

Given the positive results from the previous single-setting analyses, we would hope that the results would generalize across countries and so next 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 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 2. 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 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 indicator for being in the Sample:

[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.

Table 3: 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 and statistics are coefficient and (standard error).

Omitting non-sample-interacted coefficients (e.g., base treatment).

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

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

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, 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 3. 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.

The difference in the ATE across the countries could be due to 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. Analogous to estimating causal effects with selection on observables, we will assess external validity using both a reweighting and regression (CATE) modeling approaches.

3.1 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:

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

Covariates in the prediction model are the main covariates and missing indicators (cluster dummies would perfectly predict being in the Sample) and we estimate the model using a logistic regression. Results are shown in Table 4. 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

Table 4: 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]
* (p <0.10), ** (p <0.05), *** (p <0.01)

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 4 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 5. In this model, the coefficient on "Is Sample" reports whether the reweighting made the control outcomes similar. 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.

Table 5: Reweighted pooled treatment-effect estimation

	(1)	(2)	(3)
Treatment	1.080**	0.320**	0.361**
	(0.165)	(0.106)	(0.0622)
Is Sample	-0.437*	-0.141	0.934**
	(0.240)	(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.
* (p <0.10), ** (p <0.05), *** (p <0.01)

3.2 CATE Modelling

In this section we outline an alternative approach based on the 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:

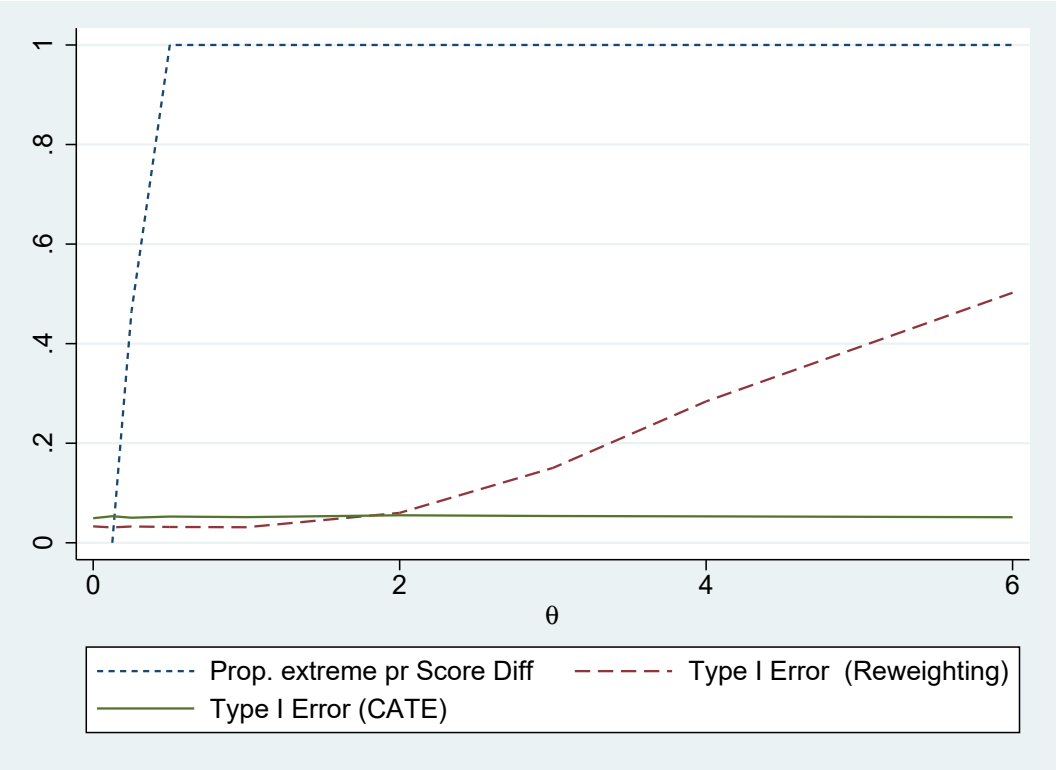
$$Y_i = (D_i \times \tilde{X}_i \beta_{\tilde{X}} + X_i \gamma_X) + (D_i \times \tilde{X}_i \times \mathbb{1}_{i \in S} \delta_{\tilde{X}} + X_i \times \mathbb{1}_{i \in S} \delta_X) + \varepsilon_i \tag{4}$$

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.

3.2.1 Simulation comparison

We motivate the use of the CATE model partially due to potential concerns about high false-positive rates when using the reweighting method. 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

Figure 3: Simulations of False Positives by estimation method



Diagnostics from simulations according to Equation 5. "Prop. extreme pr Score Diff" notes the proportion of simulations that had standardized propensity score differences of at least 0.25. 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.

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 distirbutions become more dissimilar. Results are shown in Figure 3. 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. The reweighting technique also does not improve when the true propensity scores are used in place of estimated propensity scores (Appendix Figure 4).

Common approaches for dealing with this include trimming the sample, though this presents several difficulties. First, it is difficult to know what thresholds to use, and the thresholds will change the estimated effect. Indeed, in this example, the point when θ increased enough such that

most of the simulation samples were too different according to the rule-of-thumb for standardizes differences in propensity scores did not match the point when we see increased error rates in the reweighting method. Second, trimming by propensity scores changes the interpretation (estimand) of effect in a complex way as it depends on a potentially large non-linear model. This makes interpretability of the results difficult. Finally, it is challenging to know how to extrapolate from an existing analysis to estimate an average treatment effect in a new setting.[3] While we can use the existing propensity model to predict (and trim) propensity values for the new setting, this is not really the relevant model to use. In order to extrapolate to a new setting we need to estimate a new propensity model between existing and new data in order to reweight the data. This will result in different propensity scores for the existing data and so the same trimming rule will result in different data used (highlighting the problem with the estimate being conditional on a complex model).

We note that the trimming procedure that bounds covariate values is much more straightforward to use. This, coupled with the potentially increased false-positive rate of the reweighting approach motivate our use of the CATE approach.

3.2.2 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. 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 the CATE parameters. 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 for the CATE parameters.

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.[4] The Lasso estimator is used to selected 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. In our setup, this is

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

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 sufficiently high (unlike the Ridge Regression with an L_2 penalty, which never sets coefficients to exactly zero).[5]

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. If our primary

[3] A related challenge is how to trim when the initial set of settings is more than two. Estimating logit equations for each pair of Sample-Population would result in conflicting predicted probabilities for observations. A solution here, though, is to use a single multinomial logit for all the initial settings.

[4] We note that while there are more flexible ways to estimate a non-linear CATE (e.g., Wager and Athey 2018), they do not allow to test for differences in the global model across settings.

[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.

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. Additionally, a set of variables may jointly be rather predictive even if each individually is statically significant.

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. To keep the set from being 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 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,

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

When using a selection technique, such as the 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.[6] We will 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

[6] If one is willing to make stronger assumptions on the data generating process, one could use the Post-Lasso OLS using the whole data as in Belloni and Chernozhukov (2013)

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.

3.2.3 Full approach

The final component of our approach is to reframe the test of generalizability to provide a straightforward result when there are $S > 2$ settings, where S now stands for the number of settings. The approach used with reweighting provided S separate pair-wise tests. As S increases, however, is not clear that with a single failed test we should conclude that an effect does not generalize.[7] We therefore modify the approach to provide a single combined test of generalizability, by expanding Equation 4 to have each setting interacted with the CATE.

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

We then conduct a joint test of the combined vector of coefficients $\hat{\delta}_{*,\tilde{X}} = (\hat{\delta}_{2,\tilde{X}}, ..., \hat{\delta}_{S,\tilde{X}})$.

The full algorithm is shown in Algorithm 1. We estimate this 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 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 than zero.

We show $\hat{\delta}_{\tilde{X}}$ from the setting-interacted CATE model in Table 6. We do not reject the joint test that the estimated 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, for complete reference we replicate the previous treatment effect approaches (the simple ATE comparison and the reweighting approaching) using the same subsample in Appendix Tables 7

[7] There is also a subtle multiple testing issue as the data from each country are used multiple times.

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.

3.2.4 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 Tables 10 and 11.

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.

3.2.5 Effect of Trimming

We note that the preceding treatment effect estimations was 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 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.

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		
R^2	0.339		
p-val no CATE difference	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.

4 Conclusion

In this paper we evaluate various strategies for assessing external validity (EV) of treatment effects estimates and apply them 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. When we apply the single-setting EV measures of Bo and Galiani (2021) to each country individually, the results suggest that the treatment effects is fairly generalizable, though in two of the countries it may not generalize for large shifts in the covariates. As there is a large difference in several covariates across the countries, this becomes quite important when we evaluate EV procedures that compare across countries.

We evaluate two ways of controlling for changes in the covariate distribution that assessing if there is a common conditional average treatment effect (CATE) 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 view these results on this dataset as compatible with the results of Bo and Galiani (2021). 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.

References

Alberto Abadie and Guido W. Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11, January 2011. doi:10.1198/jbes.2009.07333.

Isaiah Andrews and Emily Oster. A simple approximation for evaluating external validity bias. *Economics Letters*, 178:58–62, May 2019. doi:10.1016/j.econlet.2019.02.020.

Joshua D. Angrist. Treatment effect heterogeneity in theory and practice. *The Economic Journal*, 114(494):C52–C83, March 2004. doi:10.1111/j.0013-0133.2003.00195.x.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, July 2016. doi:10.1073/pnas.1510489113.

Susan Athey and Guido W. Imbens. The econometrics of randomized experiments. In *Handbook of Field Experiments*, pages 73–140. Elsevier, 2017. doi:10.1016/bs.hefe.2016.10.003.

Orazio Attanasio, Adriana Kugler, and Costas Meghir. Subsidizing vocational training for disadvantaged youth in Colombia: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 3(3):188–220, July 2011. doi:10.1257/app.3.3.188.

Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2), May 2013. doi:10.3150/11-bej410.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2): 608–650, April 2014. doi:10.1093/restud/rdt044.

Nicholas Bloom, James Liang, John Roberts, and Zhichun Jenny Ying. Does working from home work? Evidence from a Chinese experiment. *The Quarterly Journal of Economics*, 130(1): 165–218, November 2014. doi:10.1093/qje/qju032.

Hao Bo and Sebastian Galiani. Assessing external validity. *Research in Economics*, 75(3):274–285, September 2021. doi:10.1016/j.rie.2021.06.005.

Donald T. Campbell. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4):297–312, 1957. doi:10.1037/h0040950.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, January 2018. doi:10.1111/ectj.12097.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, January 2009. doi:10.1093/biomet/asn055.

Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii. From local to global: External validity in a fertility natural experiment. *Journal of Business & Economic Statistics*, 39(1):217–243, August 2019. doi:10.1080/07350015.2019.1639407.

Sebastián Galiani, Paul J. Gertler, Raimundo Undurraga, Ryan Cooper, Sebastián Martínez, and Adam Ross. Shelter from the storm: Upgrading housing infrastructure in Latin American slums. *Journal of Urban Economics*, 98:187–213, March 2017. doi:10.1016/j.jue.2016.11.001.

Michael Gechter. Generalizing the results from social experiments: Theory and evidence from Mexico and India. Mimeo, October 2021. URL `http://www.personal.psu.edu/mdg5396/Gechter_Generalizing_Social_Experiments.pdf`.

V. Joseph Hotz, Guido W. Imbens, and Julie H. Mortimer. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, 125 (1-2):241–270, March 2005. doi:10.1016/j.jeconom.2004.04.009.

Holger L. Kern, Elizabeth A. Stuart, Jennifer Hill, and Donald P. Green. Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, 9(1):103–127, January 2016. doi:10.1080/19345747.2015.1060282.

Hannes Leeb and Benedikt M. Pötscher. Can one estimate the unconditional distribution of post–model–selection estimators? *Econometric Theory*, 24(02), April 2008a. doi:10.1017/s0266466608080158.

Hannes Leeb and Benedikt M. Pötscher. Recent developments in model selection and related areas. *Econometric Theory*, 24(02), April 2008b. doi:10.1017/s0266466608080134.

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

Rachael Meager. Understanding the average impact of microcredit expansions: A Bayesian Hierarchical Analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91, January 2019. doi:10.1257/app.20170299.

Karthik Muralidharan, Abhijeet Singh, and Alejandro J. Ganimian. Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review*, 109(4):1426–1460, April 2019. doi:10.1257/aer.20171112.

Trang Quynh Nguyen, Cyrus Ebnesajjad, Stephen R. Cole, and Elizabeth A. Stuart. Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, 11(1), March 2017. doi:10.1214/16-aoas1001.

Lant Pritchett and Justin Sandefur. Context matters for size: Why external validity claims and development practice do not mix. *Journal of Globalization and Development*, 4(2), January 2014. doi:10.1515/jgd-2014-0004.

Elizabeth A. Stuart, Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf. 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, April 2011. doi:10.1111/j.1467-985x.2010.00673.x.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, January 1996. doi:10.1111/j.2517-6161.1996.tb02080.x.

Eva Vivalt. How much can we generalize from impact evaluations? *Journal of the European Economic Association*, 18(6):3045–3089, September 2020. doi:10.1093/jeea/jvaa019.

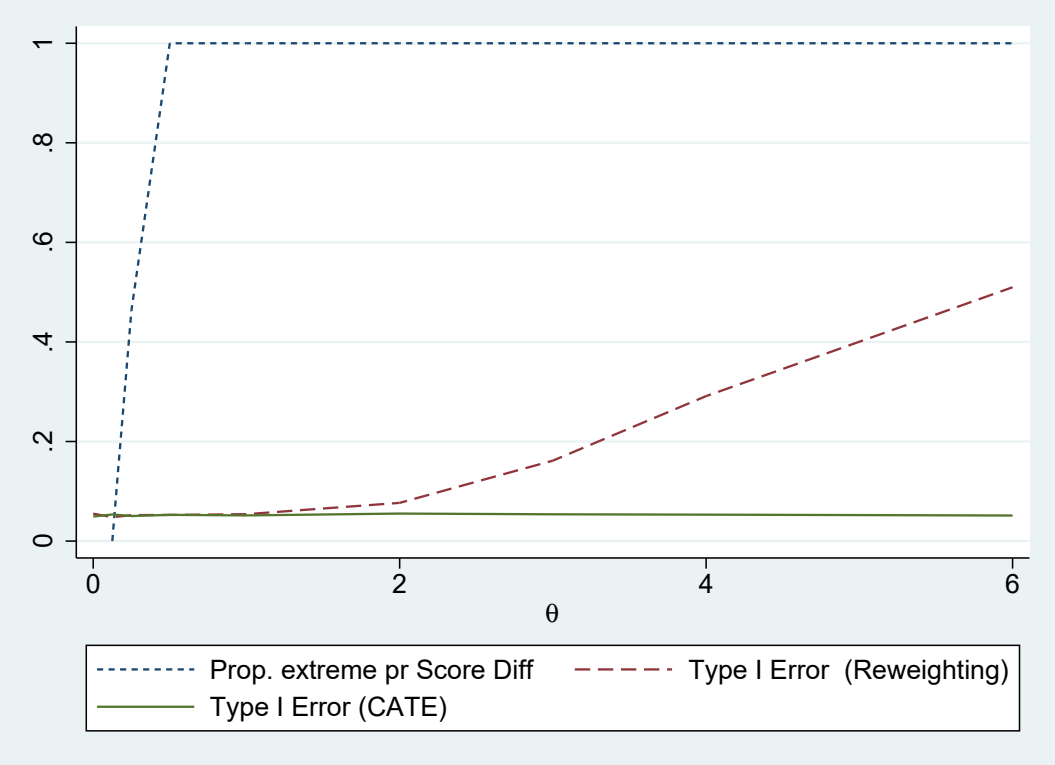
Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, June 2018. doi:10.1080/01621459.2017.1319839.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, February 2006. doi:10.1111/j.1467-9868.2005.00532.x.

Hui Zou. The Adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, December 2006. doi:10.1198/016214506000000735.

A Appendix

Figure 4: Simulations of False Positives by estimation method (using true propensity)



Diagnostics from simulations according to Equation 5. "Prop. extreme pr Score Diff" notes the proportion of simulations that had standardized propensity score differences of at least 0.25. 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 true propensity score rather then the estimated one.

Table 7: ATE Sample-interacted (excl. Sample 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 for Sample. Omitting non-sample-interacted coefficients.

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

Table 8: Reweighted pooled treatment-effect estimation

	(1)	(2)	(3)
Treatment	1.111**	0.326**	0.354**
	(0.160)	(0.105)	(0.0625)
Is Sample	0.154	-0.563**	0.956**
	(0.271)	(0.239)	(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. Exludes Sample training data.

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

Table 9: Sample-interacted CATE (including initially trimmed obs.)

	(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

Variables align with those in Table 10.