NBER WORKING PAPER SERIES

ASSESSING EXTERNAL VALIDITY

Hao Bo
Sebastian Galiani

Working Paper 26422
http://www.nber.org/papers/w26422

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2019

We thank Matias Cattaneo, Guido Kuersteiner, and Owen Ozier for their very valuable comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Assessing External Validity Hao Bo
and Sebastian Galiani NBER Working
Paper No. 26422
November 2019, Revised June 2020
JEL No. C18,C52,C93

ABSTRACT

In designing any causal study, steps must be taken to address both internal and external threats to its validity. Researchers tend to focus primarily on dealing with threats to internal validity. However, once they have conducted an internally valid analysis, that analysis yields an established set of findings for the specific case in question. As for the future usefulness of that result, however, what matters is its degree of external validity. In this paper we provide a formal, general exploration of the question of external validity and propose a simple and generally applicable method for evaluating the external validity of randomized controlled trials. Although our method applies only to RCTs, the issue of external validity is general and not restricted to RCTs, as shown in our formal analysis.

Hao Bo
Department of Economics
University of Maryland
Tydings Hall
College Park, MD 20740
hbo4@umd.edu

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

1. Introduction

In designing any causal study, steps must be taken to address both internal and external threats to its validity (see Campbell, 1957, and Cook and Campbell, 1979). Researchers tend to focus primarily on threats to internal validity, i.e., determining whether it is valid to infer that, within the context of a particular study, the differences in the dependent variables are caused by the differences in the relevant explanatory variables. External validity, on the other hand, concerns the extent to which a causal relationship holds over variations in persons, settings, and time. It is important to underscore the fact at the outset that external validity does not extend to modifications in the treatment, although in practice, researchers often try to generalize their results by conflating the two levels of generalization into a question of external validity.

Randomized controlled trials solve the problem of selection bias in the identification of causal effects. Thus, theoretically, cause-effect constructs identified by means of randomized controlled trials are internally valid, that is, they permit the identification of causal effects for the population from which the random sample used in the estimation was drawn. The outcomes of such experiments are interesting in their own right, but researchers sometimes explicitly assume external validity (EV), i.e., that the internally valid estimates obtained for one population can be extrapolated to other populations. In fact, it is not uncommon that after researchers have established a cause-and-effect relationship in a specific population, they proceed to discuss its implications based on the assumption that this relationship is generally valid. In this paper, we formalize the concept of external validity and show that in general, it is unlikely that any given study will be externally valid in any general sense. This is one reason why Manski (2013) says that the current practice of policy analysis "hides uncertainty".

Once researchers have conducted an internally valid analysis, that analysis yields an established set of findings for the specific case in question. As for the future

usefulness of that result, however, what matters is its degree of EV. The most commonly held view in this regard is that the EV problem hinges on assumptions about the relationship between the population for which internally valid estimates have been obtained and another, different population. Apart from researchers who are focusing on EV in a specific context, many researches either ignore the EV problem altogether or approach it subjectively. In this paper, we provide a formal and general reflection on the EV problem and propose a simple and generally applicable method for evaluating the external validity of randomly controlled trials (RCTs).

In this paper we define external validity as the stability of the conditional distribution p(outcomes | treatment) across different populations. We then formalize the degree to which we can make judgments about a new population (density) generated as a subpopulation from an overarching population that also generates the "original" population for which there is an internally valid estimate. Without loss of generality, assume that we have data that allows estimation of the joint distribution p(outcomes, treatment). We then have p(outcomes, treatment) = p(outcomes | treatment) × p(treatment). We say that there is external validity if, for other data with a potentially different joint distribution of outcomes and treatment, the conditional distribution p(outcomes | treatment) stays the same.

Our definition of external validity is the same as that of Janzing, Peters, and Schölkopf (2017). Admittedly, this seems quite stringent. It might be thought that, even with a moderate change of p(outcome | treatment) across different populations, external validity could be maintained. But what exact degree of change in p(outcome | treatment) leads to EV or external invalidity cannot be precisely defined. We need an operationalizable definition of EV, so, in line with a small body of literature (Janzing, Peters, and Schölkopf, 2017), we err on the side ofcaution, although we admit that there are other ways of defining EV that provide interesting and important insights, e.g., Meager (2019).

Based on our theoretical framework, we then propose two alternative measures of external validity. To the best of our knowledge, we are the first to propose formal mathematical definitions of external validity and, on that basis and in the context of an RCT, to propose purely data-driven measures related to theoretical constructs.

The measures of EV we propose in this paper can take advantage of multiple trials to evaluate the degree to which certain empirical conclusions are valid across different populations. Needless to say, ultimately, the external validity of all causal estimates is established by replication in other datasets (Angrist, 2004).[2]

We would like to determine whether a given study or a given set of studies can be generalized to other populations in general.[3] In order to do that, we propose a method that applies to RCTs, but it should be noted that the issue of external validity is general and not restricted to RCTs, as shown in our formal and general reflection below.

The rest of this paper is structured as follows. In Section 2, we provide a formal and general reflection on the EV problem. Based on the model described in that section, in Section 3 we propose a simple and generally applicable method for assessing the external validity of RCTs. Finally, we present final remarks.

2. External Validity

A single experiment (or a set of experiments regarded as a single experiment)[4] allows us to arrive at a point estimate for the population of cause-effect parameters. Assessing the EV of one causal parameter entails estimating treatment effects as a function of different populations. Thus, evaluating the EV of an internally valid estimate of a cause-effect parameter entails assessing a distribution of cause-effect

[2] In the areas of labor and development economics, a number of studies use similar multi-country strategies to generalize cause-and-effect constructs. For example, Cruces and Galiani (2007) examine the effects of fertility on labor outcomes in three countries; Dehejia, Pop-Eleches, and Samii (2019) examine the causal effects of sibling sex composition on fertility and labor supply across many countries and years and characterize how its effects vary in terms of available covariates; Banerjee et al. (2015) study microcredit in six countries; Galiani et al. (2017) study the effects of sheltering the poor in three countries; Gertler et al. (2015) study health promotion in four countries; and Dupas et al. (2016) examine the effects of opening savings accounts in three different countries.

[3] For example, Deaton (2010) writes: "We need to know when we can use local results, from instrumental variables, from RCTs, or from nonexperimental analyses, in contexts other than those in which they were obtained."

[4] When we have a set of experiments and we reach a conclusion from them, we have to find a way to aggregate their outcomes from the experiments so they can be regarded as a single experiment (correspondingly, behind that single experiment there would be a single population formed by a mixture of the populations underlying the different original experiments). However, in this paper we do not discuss how to aggregate outcomes form different experiments, which we consider an issue that is specific to each research project. We instead simply assume that each sample point in each experiment have the same weight in the aggregation process.

parameters based on a single draw from it.

In this section, we formalize the concept of EV. We develop our framework in terms of the stability of density functions across populations because nearly all sample analyses are intended to characterize an underlying population. To explore EV, we assume internal validity has been achieved, that is, it is always possible to obtain consistent estimates of the joint density of outcomes and treatment status or the conditional density of outcomes conditional on treatment status.

Focusing on population densities might seem unnecessary, since most researchers need to model only the first and second moments of a population (density) to obtain their parameters of interest. The first moment is needed for a point estimate, while the second moment is used for evaluating sampling variability. Nevertheless, both moments are a function of a population density, i.e., a density of outcomes conditional on treatment status. We focus on population densities here because we want to emphasize that the nature and difficulty of assessing EV lies in the differences among populations. In addition, at a conceptual level, this simplifies the analysis, since it is necessary to make comparisons for only one entity (the density) instead of two (the first and second moments).

2.1. From one population to another

Conducting external inference from internally valid estimates entails switching the population under study. Assume that there is an overarching population that consists of all vectors (y, z, w) from the probability density D(y, z, w; θ). Also assume that we obtain a sample from a subpopulation defined by the density D(y, z; $w = w_0, \theta$). When we say something, based on estimates from this sample, about another subpopulation defined by the density D(y, z; $w = w_1, \theta$), for some $w_1 \neq w_0$, we are conducting external inference.

This is the general setup: w defines, in a general way, the difference between populations; internally valid inferences will usually yield different estimates of the cause-effect constructs of interest. θ governs how the differences in w affect those constructs across populations. Alternatively, we could define different populations by assuming that their w's are distributed in different ranges instead of assuming that

they take different point values, but this change would not add any insight to the analysis. Actually, the assumptions are conceptually equivalent.

w_0 is a given realization of w and is constant for the subpopulation for which the sample is used to conduct the empirical analysis. For example, if we draw a sample within a country, all different sample points have the same country identity (w_0). With the sample drawn from D(y, z; $w = w_0, \theta$), whether we estimate the joint density of (y, z) or a conditional density of y on z, we always conduct inference under the condition w = w_0. Then, conducting external inference implies assessing whether such estimates are valid for a different population (in our example, from another country), characterized by $w = w_1$.

Researchers generally do not have information about how changing w would change the density D(y, z, w; θ). Usually, they have to make assumptions in this regard in order to conduct external inference. We now explore this question formally. It is informative to express D(y, z; $w = w_0, \theta$) as follows:

$$D(y, z; w = w_0, \theta) = D(y|z; \theta_1(w_0, \theta) \times D(z; \theta_2(w_0, \theta))$$

Assume that the estimand is the conditional density D(y | z; $\theta_1(w_0, \theta)$). In the case of an RCT, the marginal density $D(z; \theta_2(w_0, \theta))$ can be ignored, since z is randomly assigned. Thus, assume we estimate $D(y|z;\ \theta_1(w_0, \theta))$ and then want to know how well we would assess another population characterized as $w = w_1$ if we rely only on our internally valid estimate. Assuming the conditional density function is differentiable almost everywhere with respect to w and applying the mean value theorem, we obtain:

$$\int [D(y|z;\ \theta_1(w_0, \theta)) - D(y|z;\ \theta_1(w_1, \theta))]^2 \, dy$$

$$= (w_1 - w_0)^2 \int \left(\frac{\partial D(y|z;\ \theta_1(w_b, \theta))}{\partial \theta_1} \frac{\partial \theta_1}{\partial w_b} \right)^2 dy$$

where w_b takes a value between w_1 and w_0. If w_1 is not known but can be assumed to be close to w_0, we can approximate w_b by w_0 (where the higher order

residual in a Taylor expansion is negligible). Then, for a close w_1, the more sensitive the conditional density with respect to w is, the more we will miss the target while conducting external inference by relying only on an internally valid estimate of the estimand of interest in our sample. Importantly, this is independent of the sample size.

Relying on this setup, we define:

(1) Punctual local external validity when $w_1 = w_0$;

(2) Local external validity as $\int(\frac{\partial D\left(y\middle|z;\ \theta_1(w_1,\theta)\right)}{\partial\theta_1}\frac{\partial\theta_1}{\partial w_1})^2 dy = 0$ for all w_1 within a small interval of w_0;

(3) External validity as $\int(\frac{\partial D\left(y\middle|z;\ \theta_1(w,\theta)\right)}{\partial\theta_1}\frac{\partial\theta_1}{\partial w})^2 dy = 0$ for all w_1;

(4) Indirect external validity where there exist $f(w;\theta_1) = \int(\frac{\partial D\left(y\middle|z;\ \theta_1(w,\theta)\right)}{\partial\theta_1}\frac{\partial\theta_1}{\partial w})^2 dy$, either known or estimable that can be used to adjust $D\left(y\middle|z;\ \theta_1(w_0,\theta)\right)$ to calculate $D\left(y\middle|z;\ \theta_1(w_1,\theta)\right)$.

In the literature, researchers conducting external inference have attempted to either test (1), (2), or (3) or to exploit (4). As our discussion makes clear, the question of external validity rests on the relationship between the population for which we have an internally valid estimate and the population about which we are to make judgments. The function $f(w;\theta_1) = \int(\frac{\partial D\left(y\middle|z;\ \theta_1(w,\theta)\right)}{\partial\theta_1}\frac{\partial\theta_1}{\partial w})^2 dy$ formalizes this relationship.

2.2. From one population to any population

We now extend our analysis. The estimand is a conditional model based on data generated from the density $D(y,z;\ w = w_0,\theta)$. We denote the conditional model by $D(y \mid z;\ w = w_0,\theta_1)$, where θ_1 is the parameter governing the conditional model and is a function of w_0 and θ. Applying Bayes' law and assuming z is weakly exogenous (see Engle, Hendry, and Richard, 1983) for θ (this is always the case when z is randomly assigned), we have:

$$D(y \mid z;\ w = w_0, \theta_1) = \frac{D(y,z;\ w=w_0,\theta_2)}{D(y,z;\ \theta)} \times D(y|z;\theta_3),$$

where $D(y|z;\theta_3) = \int D(y|\ z,w;\theta)\, dw$, $D(y,z;\ \theta) = \int D(y,z,w;\theta)\, dw$, and θ_3 is a function of θ. Note that here we use $D(y,z;\ w = w_0,\ \theta_2)$, where θ_2 is a function of θ, rather than $D(y,z;\ w = w_0,\ \theta)$, because we want to emphasize that, when moving from D(y, z, w; θ) to $D(y,z;\ w = w_0,\ \theta_2)$, the parameter vector θ may change. θ_2 can be seen as a function of θ and w. However, there is nothing wrong with using the general notation $D(y,z;\ w = w_0,\ \theta)$ instead. External inference in this setting entails assessing $D(y|z;\theta_3)$ based on the estimation of $D(y \mid z;\ w = w_0,\theta_1)$ (i.e., from a subpopulation to the overarching population). Note also that θ_2 and θ_3 are not variation-free (Engle, Hendry, and Richard, 1983), so θ_3, estimated by optimizing a loss function based on $D(y \mid z;\ w = w_0,\theta_1)$, generally does not coincide with the result obtained by estimating on the basis of an optimization of a loss function based on $D(y|z;\theta_3)$, which researchers cannot optimize in any event, since the available data is generated only from the subpopulation with $w = w_0$. Taking these considerations into account, we define:

(1) Overarching external validity as $\frac{D(y,z;\ w=w_0,\theta_2)}{D(y,z;\ \theta)} = 1$.

(2) Indirect overarching external validly when the function $g(\theta) = \frac{D(y,z;\ w=w_0,\theta_2)}{D(y,z;\ \theta)}$ is either known or estimable and can be used to adjust a loss function based on $D(y \mid z;\ w = w_0,\theta_1)$ that makes it possible to estimate $D(y|z;\theta_3)$. Clearly, this requires a known (or assumed) relationship between the subpopulation under study, $D(y,z;\ w = w_0,\ \theta_2)$, and the overarching population, $D(y,z;\ \theta)$, which is given by $D(y,z;\ \theta) = \int D(y,z,w;\theta)\, dw$.

From the above discussion, it is clear that it is more likely that an internally valid estimate will have punctual or local external validity than external validity or overarching external validity. Thus, not surprisingly, the existing literature has focused on specific populations for extrapolation, making specific assumptions about the relationship between the population for which there are internally valid estimates and the target population. Next, we review that literature.

Remark: In the above framework, we provide insights on EV based on the stability and equivalence of population densities, either of densities from two sub-populations or of densities from a sub-population and an overarching population. It could be said that his is unnecessarily stringent. But in order to provide useful insights on the EV problem, we decided to err on the side of being stringent to make things operationalizable. Every model is wrong, but some are useful (Box, 1976).

2.3. Literature Review

One reason why internally valid estimates of causal constructs might lack external validity is changes in the population over time. We could posit, for example, that w in the above framework varies over time. In such a setup, Rosenzweig and Udry (2018) provide an innovative way of conducting external inference. Using repeated cross-sections, they estimate the causal effect of interest over time, where in each period the vector w is fixed at some specific value. They focus on one dimension of w for which they have a measurement, i.e., rainfall. They then estimate the response of the casual construct of interest to rainfall. Using the empirical distribution of the underlying shock (rainfall), they can infer both how the causal parameter of interest varies with this shock and its average effect. Thus, they also estimate the effect for the overarching population. This method requires that other time-varying unobservable variables in w are not correlated with the observable one, which, in their application, may be the case, since rainfall is determined outside the economic system, although it still might trigger adjustments in some unobservable variables.

Andrews and Oster (2019) propose a method for estimating the average treatment effect (ATE) for a target population based on another population (often a trial population) for which a researcher is assumed to have an internally valid estimate. They assume that the conditional ATE-given covariates and unobservables are the same in the trial and target populations (and that the covariates and unobservables are uncorrelated). First, they adjust the ATE by differences in covariates between the trial and the target populations. Second, they model how unobservables and covariates simultaneously affect individual treatment effects and the likelihood that individuals in the target population were also in the trial population. Relying on this model, they derive a formula to adjust the ATE for differences in unobservables.

Other papers have dealt with the issue of non-compliance in instrumental variables estimation. One line of discussion about external validity relates to the local average treatment effect (LATE) (Angrist, Imbens, and Rubin, 1996). The standard setup assumes the presence of a binary endogenous treatment variable instrumented with a binary ignorable variable (for example, random assignment to treatment). Assuming monotonicity, the population is divided into three groups: (1) compliers, whose treatment status is affected by the instrument; (2) always-takers, who always receive treatment regardless of the value of the instrumental variable; and (3) never-takers, who never receive treatment regardless of the value of the instrumental variable. The second and third groups are usually combined and labeled as non-compliers. The variation (information) in the instrument only makes a difference for compliers. Since, in this setting, internally valid estimates are derived from the variation in the instrument, the estimates do not provide a basis for internally valid inferences about the whole population, but only about a hypothetical population of compliers. The question with regard to external validity that usually arises in this setting has to do with when and how estimates for compliers can be used to infer the parameter of interest for the whole population. Naturally, the first step in answering this question is to understand the relationship between compliers and non-compliers. Examples in the literature include Angrist (2004) and Angrist and Fernandez-Val (2010). Angrist (2004) examines a few possible relationships between compliers and non-compliers that yield externally valid inferences and then estimates the ATE using information for the LATE under each relationship. Angrist and Fernandez-Val (2010) assume that the instrumental variables, conditional on covariates, are as good as if they were randomly assigned and that observable covariates fully determine covariate-specific treatment effects. The relationship between compliers and non-compliers is then reduced to different compositions of observable covariate values. Since the LATE and ATE are both weighted sums of (observable) covariate-specific treatment effects, the EV problem of differentiating the ATE from the LATE becomes one of modifying the weights used in the LATE to align them with the weights used in the ATE.

Another line of analysis concerning external validity involves regression discontinuity methods. A regression discontinuity estimator is, by definition, a local estimator: it only identifies causal constructs for the subpopulation of subjects whose

forcing variable values are near a discontinuity threshold (and are also compliers in the case of a fuzzy design). The external validity question usually asked in such a setting is how estimates for the subjects near the threshold (and that are also compliers) apply to the sample population. Naturally, the first step in assessing external validity in this setting is to understand the difference between the subjects whose treatment effect can be identified and other subjects. Dong and Lewbel (2015) exploit the sensitivity of estimates of the forcing variable to shed light on the relationship between subjects with different forcing variable values. Angrist and Rokkanen (2015) advocate testing whether the forcing variable and the "treatment" outcome are uncorrelated conditional on variables, which, if it were the case, would be informative about the relationship between subjects with different forcing variable values and, in turn, could be of use in addressing external validity questions. Bertanha and Imbens (2018) focus on the fuzzy regression discontinuity design and provide a test to determine whether compliers are systematically different from non-compliers conditional on the forcing variable as well as exogenous covariates. They argue that external validity requires the null hypothesis of no differences in order to be valid.

These studies share a common feature: the pivotal element in approaching the issue of external validity is assumptions about the relationship between the population for which there are internally valid estimates of causal parameters and the population for which the researcher would like to make an external inference. The relationships exploited in the above-cited studies are the one between compliers and non-compliers, the one between subjects with different forcing variable values, and the one between different periods. The effectiveness of any practical evaluation of external validity is determined by these relationships, as we explain in Sections 2.1 and 2.2.

These papers focus on specific populations as basis for extrapolations and do not explore EV in any general form. This is natural enough, since, for any given population, the relationship between it and the population that was originally studied can be easily assumed or modeled, while it is very hard to undertake an evaluation of EV in general. Thus, in terms of our analysis, the literature has focused mostly on methods relating to the concept of local external validity.

In the next section, we propose a method for assessing EV both for specific populations and generally. Our method is based on the insights derived from Sections 2.1 and 2.2. The focus is on determining the likelihood that an internally valid estimate could be generalized not only to the overarching population (overarching external validity) but also to specific populations, starting from those close (local external validity) to the one studied and then moving away from it to more different populations (external validity).

3. Assessing external validity

We propose a method for evaluating the degree to which a conclusion based on a given population applies to populations represented by samples that have been formed by randomly reweighting the original sample. Our method is therefore both data-driven and generally applicable. It is important to note that for researchers with several experiments and underlying populations at hand, they still need to ponder on to what degree their conclusions from the several experiments can carry on generally. So, they can take advantage of our proposed measures by regarding their experiments as a single experiment and the underlying populations as a single population. In the following discussion and examples, we always start from a single sample or experiment (i.e. single population).

In order to maintain consistency with the theoretical framework outlined in Section 2, we assume that each reweighting of the original sample corresponds to a value of w, that is, represents a new population. After defining and constructing new samples by reweighting, we propose a way of measuring the extent to which the conclusion reached about the original sample holds true for the reweighted samples (or the new populations). This is a global measure of EV (based on the concept of overarching external validity). We also propose a local measure of EV (based on the concepts of local external validity and external validity) by grouping new populations based on a specific criterion and then measuring the degree to which the conclusion for the original sample still holds for each group.

Reweighting is usually conceptualized as a way to explore within population sampling variability. But this is largely an epistemological stance based on a mental construction that treats a sample as an empirical "estimator" of a population, rather than an identity in nature. As we get into more details, we will explain why certain

reweighting works for our purpose of generating new populations, as well as how we take account of within populations sampling variability.

Next, we provide the details and examples of our proposed measures of EV. It is very important to notice now that we are proposing measures of EV, not statistical tests of it.

3.1 Defining new populations and constructing representative samples

We assume that a researcher starts from a randomized controlled trial (RCT) dealing with a given population and then wants to assess the degree of EV of the causal constructs that have been estimated. We provide a general method for doing this. Our method, inspired by our formal and general reflection above, takes advantage of RCTs as shown below. However, it should be noted that the problem of external validity is general and not restricted to RCTs.

Assume that the sample size of the control group is m and the sample size of the treatment group is l, so the sample size analyzed is n=l+m. Assume also that there is baseline pre-treatment information (Y, X)', which is usually the case. We pair each observation in the treatment group with its nearest observation in the control group in terms of the Mahalanobis distance[5] using the baseline information (the Mahalanobis distance between vectors of baseline information (Y, X)'). We thus choose the nearest neighbor in the control group to each observation in the treatment group as a counterfactual and pair the two observations. Repetitive use of observations from the control group are allowed, and unused observations from the control group are discarded after all observations from the treatment group are paired. For each pair, we assign an index $i \in \{1,2,\ldots l\}$.

We define a reweighting vector P for the indices: $P = (p_1, p_2, \ldots, p_l)' = \frac{(G_1, G_2, \ldots, G_l)'}{\sum_1^l G_i}$, with $p_i \geq 0$ and $\sum_1^l p_i = 1$. $(G_1, G_2, \ldots, G_l)'$ is a random vector, with each element drawn independently from the Gama(1,1) distribution. Reweighting the original

[5] The Mahalanobis distance can be defined as the Euclidean distance with each variable rescaled to have unit variance. Though this distance is the most commonly used measure in the literature, there are many alternative matching criteria (Rosenbaum, 2010) that researchers can use for their specific purposes. In addition, researcher can match one sample point for a group (treatment or control group) with many sample points from the other group.

sample based on the Gama(1,1) generates reweighting vectors uniformly distributed over all possible reweighting vectors (Efron and Hastie, 2016), so our exploration of how the original conclusions apply to new populations treats every possible population evenly. For each reweighting $P = (p_1, p_2, \ldots, p_l)' = \frac{(G_1, G_2, \ldots, G_l)'}{\sum_1^l G_i}$ of the original sample, we create a reweighted sample in which p_i is the weight for the pair indexed by i=1, 2, … , l. We do multiple reweighting (as many as 1,000 times), and we regard each reweighting outcome as a sample representing a new population. Weighting the actual sample by $\left(\frac{1}{l}, \frac{1}{l}, \ldots, \frac{1}{l}\right)'$ is a consistent estimator of the original population. Therefore, each weighting vector corresponds to a (consistent estimate of) a new population.

Though it should be obvious, for those who strongly believe that reweighting can only be used to evaluate within population sampling variability, we have the following argument for using reweighting to generate new populations:

(1) Reweighting the original sample based on the Gama(1,1) generates reweighting vectors uniformly distributed over all possible reweighting vectors (Efron and Hastie, 2016), so our exploration of how the original conclusions apply to new populations treats every possible population evenly.
(2) For example, if you have a sample with 50 men and 50 women, and then you generate a reweighted sample with 20 man and 80 women, it's a matter of perspective choice to see the new sample as representing a new population. Using reweighted sample to evaluate within population sampling variability is largely an epistemological stance based on a mental construction that treats a sample as an empirical "estimator" of a population, rather than an identity in nature.
(3) Also a reweighted sample, though not drawn form a new population, can be regarded to represent a new population in the light of the following observation by Fisher: *"[…] the only populations that can be referred to in a test of significance have no objective reality, being exclusively the product of the statistician's imagination […]." (Fisher, 1956)*

3.2 A global measure of EV

We now propose a global measure of EV for the average treatment effect (ATE), but our measure can be applied to any estimand obtained by contrasting treatment and control groups in an RCT. For each reweighted sample, we calculate the ATE and its standard error. Our proposed global measure of EV for the original internally valid analysis is based on the percentage of new populations for which the original conclusion still holds true. We say that the original conclusion holds when one of the following holds: if, in the original sample, the estimate was not statistically significant at a certain level, then the same is true for the reweighted sample or, if the estimate for the original sample was statistically significant at a certain level, then, in the reweighted sample it is also significant, at least at that level, and with the same coefficient sign as in the original population. Framing things this way, we also account for within-population sampling variability while trying to measure EV.

The above is how we operationalize whether the original conclusion holds. Some may argue that a given case will be scored as externally valid where the original estimate takes a value of 1 with a standard error of 0.45 and the reweighted estimate takes a value of 10 with a standard error of 2, but the original estimate will not be scored as externally valid if the reweighted estimate takes a value of 0.99 with a standard error of 0.55, even though this is extremely close to the original estimate.

In terms of number magnitude, the people who make this argument have a point, but what happens in practice? In cases where there is a value of 0.99 with a standard error of 0.55, the researcher will usually report no effect, that is, the original conclusion of a positive significant effect does not hold. In cases where there is a value of 10 with a standard error of 2, the researcher will usually report a significant positive effect, that is, the original conclusion of a positive significant effect does hold, although the effect magnitude is a question that calls for further exploration. While the above argument makes a point, we have preferred our method of operationalization because it can lead to something that we think is useful. What is more, we are by no means suggesting that this should generally be the only method of operationalization.

We start from an RCT and estimate the new populations by means of a matching estimator with replacement. Abadie and Imbens (2006) set out conditions for the

consistency of this matching estimator.[6] However, in finite samples, there are two sources of bias due to imperfect matching on observables and non-matching on unobservable. For the second source of bias, researchers can perform a sensitivity analysis as introduced by Rosenbaum (2010) and Imbens and Rubin (2015).

Correcting the first source of bias can be done using the methods discussed in Imbens and Rubin (2015, sect. 18.8). We use one of those methods to adjust the estimates for new populations and therefore, from now on, when we refer to the estimated ATE for new populations, we are referring to bias-adjusted estimates, as follows:

1. With the data from the original control group, we have regressed the outcomes on observable variables and recorded coefficients for baseline variables as *B*.
2. We have adjusted each pair's treatment-minus-control outcome by *– (observable variables of the treated in the pair)*B + (observable variable of the control in the pair)*B.*

Up until now, we had assumed that all the subjects were compliers in the original sample, or we had focused on the intention-to-treat effect (ITT). However, there is no problem with including non-compliers in the analysis and focusing on a parameter such as the LATE since the reweighted samples are expected to be balanced for non-compliers, especially when the sample size is large.

Alternatively, one possibility is to estimate the average effect on compliers. Under standard assumptions, we know who the never-takers in the treatment group are. In their case, we can improve on the matching with the control group because we can add the residuals of the treatment effect analysis (in addition to the baseline variables already used) to the matching variables. Thus, we can also match on unobservables.

[6] The data and treatment assignment from an RCT satisfy the conditions required for consistency given in Abadie and Imbens (2006). The matching estimator here is a weighted sum, but given our way of generating reweighting vectors, when the sample size goes to infinity, the probability that a finite number of pairs receive all the weight goes to zero, so by applying a law of large numbers, such as Chebychev's weak law of large numbers, consistency is proved.

This matching strategy does not work for compliers, since their residuals are affected by their heterogeneous treatment effects. Always-takers can be matched in the same fashion. After matching, researchers can restrict their analysis to compliers both in the original sample and in the EV exercises proposed in this paper.

3.3 Local measures of EV

With a large number (e.g., 1,000) of reweighted samples, each representing a new population, these populations can be grouped based on a given criterion. We now have a vector of treatment-control matched pairs from the original sample. Each pair yields a treatment-minus-control outcome, adjusted as proposed in the previous section based on Imbens and Rubin (2015). First, we calculate the correlation between the vector of these adjusted treatment-minus-control outcomes and each reweighting vector. The higher this correlation is, the more weight a reweighting vector gives to pairs with high treatment-minus-control outcomes. Second, we calculate 1 minus this correlation for each reweighting vector (i.e., for each new population); this calculation gives the distance of a new population from the population with the largest effect magnitude. Note that the original population has a distance of 1 because the above correlation for the original population is zero, so the extent of the difference between other populations and the original population can be summarized by how their distance measure differs from 1.

Intuitively, populations with a distance measure close to 1 give similar weights to pairs with high or low adjusted treatment-minus-control outcomes, so they are "near" the original population, which gives equal weight to every pair. Populations with a distance measure greater (smaller) than 1 have a weighting vector that is negatively (positively) correlated with the vector of the adjusted treatment-minus-control outcomes, so they give more weight to pairs with small (large) effect magnitudes.

With the above definition of distance, we then also propose using the EV curve to measure the degree to which the conclusion regarding the original population holds for new populations as their distance from the population with the largest effect magnitude (moving away from distance zero) or from the original population increases (moving away from distance 1). The EV curve is defined below, with

distance denoted by d.

$$\mathrm{EV(d)} = \frac{\text{Number of new populations for which the original result holds at distance} \in [\mathrm{d}, \mathrm{d}+\epsilon]}{\text{Number of new populations at distance} \in [\mathrm{d}, \mathrm{d}+\epsilon]}$$

In the definition of the EV curve, the original result holds when, as above, one of the following holds: if, in the original sample the estimate was not statistically significant at a certain level, then the same is true for the reweighted sample or, if the estimate was originally significant at a certain level, then it is also significant, at least at that level, and with the same coefficient sign as the new population.

Before we provide examples illustrating the two methods proposed above, we need to add a caveat that applies to both of them. If the sample (or multiple samples) at hand or the population (or populations) represented by the sample (or samples) being studied does not contain characteristics that would generate a new population and are relevant for the statistical inference, then our method is moot. One cannot make bricks without straw. Any statistical method can be useful only up to the point that its information inputs allow. Our method is designed to provide the best possible assessment of EV purely based on the data that researchers have, without theoretical or structural assumptions being involved.

3.4. Two Simulated Examples

We now provide two simulated examples to illustrate our method for assessing EV. First, we start from an internally valid analysis assumed to have a significantly positive ATE. We then calculate the global measure of EV and present the EV curve to be used to assess EV locally. Second, we do the same exercise starting from an internally valid analysis that is assumed to have an ATE that is not significantly different from zero.

3.4.1 Assessing EV for an internally valid significant positive result

In the first simulated example, the sample size is 100, with 50 observations in the treatment group and 50 in the control group. There are two observable variables, x1 and x2, and one unobservable variable, u. For each of these three variables, 100 values are drawn independently from the standard normal distribution. The

(potential) heterogeneous treatment effect for each observation, whether it is in the treatment group or control group, is given by $\tau_i = x_{1i} + x_{2i} + u_i + v_i + 10 * 1\{c_i > 0.8\}$, where v_i is drawn from a unit-variance normal distribution with a mean=-1, and c_i is drawn from the 0-1 uniform distribution.

We define the treatment status vector, a vector with a length of 100, as D, of which the first 50 elements are equal to 1 and the other 50 elements are equal to 0. D_i=1 indicates that observation i is in the treatment group; D_i=0 indicates that observation i is in the control group. The outcome variable is thus defined as: $y_i = 1 + \tau_i * D_i + x_{1i} + x_{2i} + u_i$. The internally valid estimate of ATE in our simulation is 1.52 with a standard error equal to 0.677 and a p-value for a test of the null hypothesis (ATE equals 0) equal to 0.027.

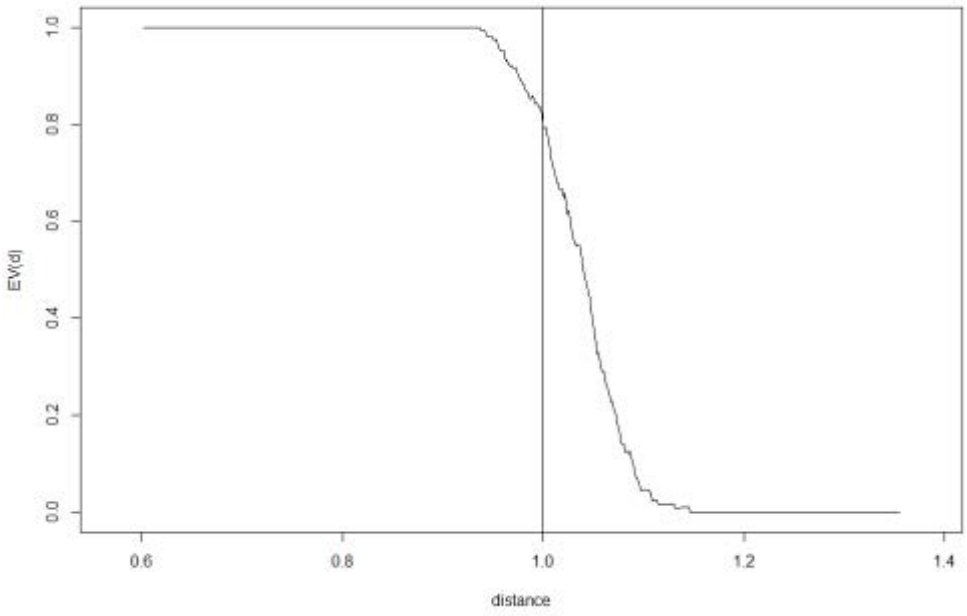
We generate 1,000 reweighting vectors over the pair indices as explained in Section 3.1 and calculate the global measure of EV introduced in Section 3.2. In this case, this measure is the proportion of these 1,000 new populations for which the lower bound of the confidence interval is greater than zero. Since the original analysis is significant at the 95% level, we choose the lower bound associated with two standard errors. The value of the global measure of EV in our simulated example is **0.584**. This means that the conclusion reached in the original internally valid analysis holds for 58.4% of the uniformly generated new populations.

Now we compute the EV curve, EV(d), introduced in Section 3.3. To apply the definition of EV(d)[7], we note that when the original result holds for a new population, this means that the confidence interval lower bound for the new population is greater than zero. Since the original analysis is significant at the 95% level, we choose the lower bound for new populations as point estimates minus two standard errors. If we look at the curve starting from a distance equal to 1, we see the positions of the new populations relative to the original population, whose distance measure equals 1. (Remember that the distance is a measurement of the distance of a new population from the population with the largest possible effect

[7] With respect to the choice of ϵ in the definition of EV(d), we choose a value of 0.05 in both this and the next example in order to make the curve smooth. This works like a moving average that smooths out a graph. Note that if ϵ is too small, the curve will be very rugged locally; if it is too large, the curve will not be locally informative, since it will simply be an overall average.

magnitude.) In Figure 1 below, we see that the original conclusion of positive significance is very likely to hold at a small distance (around 1 and smaller than 1) and it is very unlikely to hold at a large distance. Intuitively, new populations with small distances have more weights on pairs with large effect magnitudes and those with large distances have more weights on pairs with small or even negative effect magnitudes. In this example, EV is assessed locally by seeing how quickly EV(d) drops as the distance moves to the right and away from distance 1 (the distance for the original population), around which EV(d) is about 80%.

Figure 1



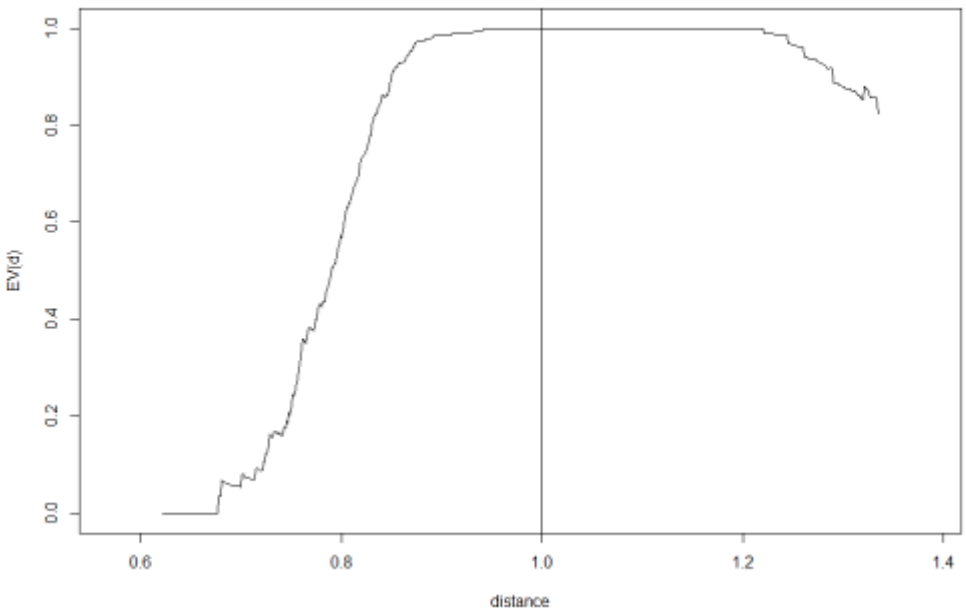
3.4.2 Assessing EV for an internally valid result without a significant difference from zero

In the second simulated example, the setup is the same as in the first example except that the (potential) heterogeneous treatment effect is instead given by $\tau_i = x_{1i} + x_{2i} + u_i + w_i + 10 * 1\{c_i > 0.8\}$, where w_i, instead of v_i as in the previous example, is drawn from a unit-variance normal distribution with a mean=-2. x_{1i}, x_{2i}, u_i, c_i, D_i, and y_i are generated in the same way as in the previous example. The internally valid estimate of ATE in our simulation is then 0.52 with a standard error equal to 0.677 and a p-value for a test of the null hypothesis (ATE equals 0) that equals 0.446.

We generate 1,000 reweighting vectors over the pair indices as explained in Section 3.1 and calculate the global measure of EV introduced in Section 3.2. In this case, this measure is the proportion of the new populations with confidence intervals including zero in these 1,000 new populations. Since the original analysis yields an estimate not significantly different from 0 at the 95% significance level, we choose the lower bound of the confidence interval associated with two standard errors. The value of the global measure of EV in our simulated example is **0.908**. This means that the conclusion reached in the original internally valid analysis holds for 90.8% of the uniformly generated new populations from reweighting the original sample.

Now we compute the EV curve, EV(d), introduced in Section 3.3. Applying the definition of EV(d), we note that when the original result holds for a new population, this means that the new population's confidence interval includes zero. Since the original analysis yields an estimate not significantly different from 0 at the 95% significance level, we choose the range of confidence intervals for new populations as point estimates $\pm$ two standard errors. In Figure 2 below, as expected, we see that the original conclusion of no significance (neither significantly positive nor significantly negative) is very unlikely to hold at very large or very small distances. We also see that, as new populations move closer to the original population, whose distance measure is equal to 1, EV(d) increases. In Figure 2, we see that EV(d) is 100% in the small neighborhood of the original population and that EV(d) eventually drops as the distance measurement moves away from 1.

Figure 2



3.4.3 Monte Carlo exercises

We have proposed a measure for use in a given analysis, not an estimator for a fixed parameter in the population or a test. Continuing with the above examples, in this section we perform a simple Monte Carlo exercise to see how the EV curves and the global EV measures vary across 50 repetitions.[8]

Now we repeat the simulated example given in 3.4.1 50 times. Specifically, we start from the same sample used in 3.4.1, repeat the 1,000-reweighting 50 times, and calculate the global measures and the EV curves 50 times. Note that we start from the same sample rather than generating 50 samples from the population to start with because analyses based on different samples from the same population have different levels of EV, and our measures are being used to evaluate EV for given analyses. Figure 3 and Figure 4 below show how the global EV measure and the EV curve vary. The takeaway is that they do not vary much.

[8] Why 50 times? We obtain quite similar results when we repeat the example 100 times or 200 times, but at those levels the results, although consistent with the results obtained with 50 repetitions, simply appear as black blotches.

Figure 3

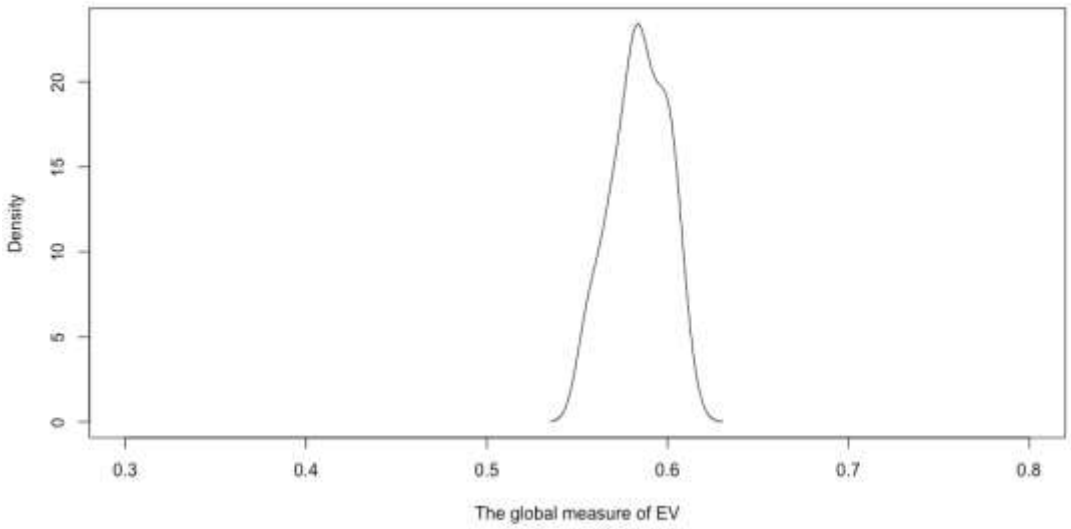
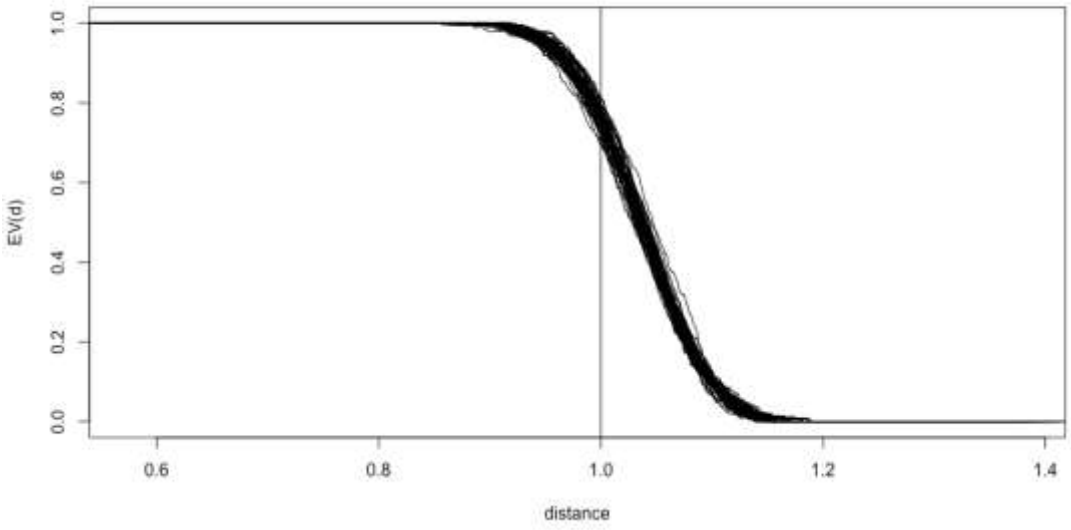


Figure 4



Now we repeat the simulated example in 3.4.2 50 times. Figure 5 and Figure 6 below show how the global EV measure and the EV curve vary. The takeaway is that they do not vary much.

Figure 5

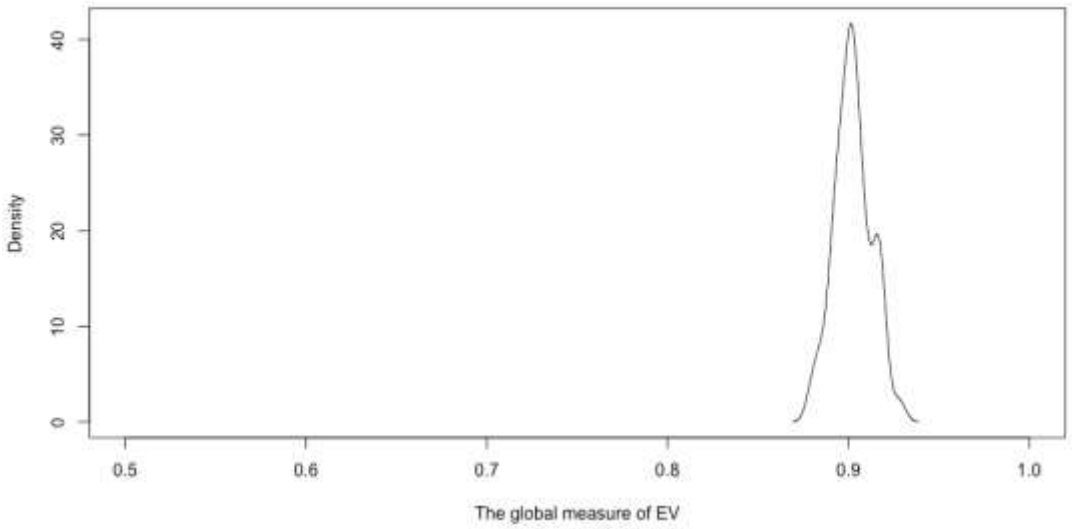
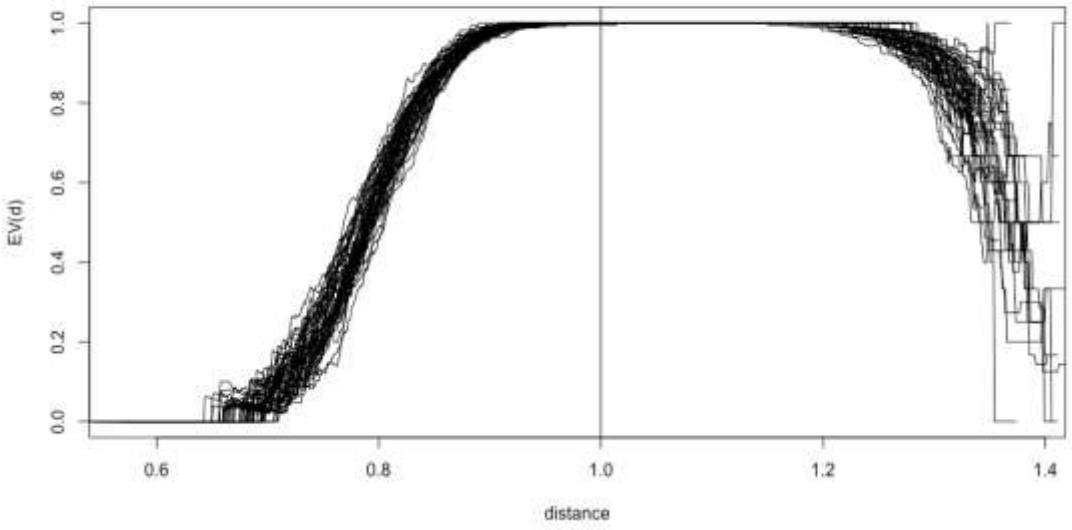


Figure 6



3.5 A comment regarding multiple samples

Nowadays researchers can obtain data from multiple samples, but they still need to determine to what degree the conclusions that they draw from several different experiments or samples are generally applicable. In order to deal with this question, they can take advantage of our proposed measure by regarding their several experiments as a single big experiment and the underlying populations as a single population (as a mixture[9]). The above discussions assume a single sample, but they

[9] Of course, there is the problem of determining the weightings when dealing with a mixture,

can be straightforwardly extrapolated to apply to cases involving multiple samples.[10] Actually, in the above simulated examples, we use $\tau_i = x_{1i} + x_{2i} + u_i + v_i + 10 * 1\{c_i > 0.8\}$ and $\tau_i = x_{1i} + x_{2i} + u_i + w_i + 10 * 1\{c_i > 0.8\}$ to generate data, and the last term in each equation shows that the simulated examples can also be regarded as having been derived from two different samples from the perspective of an EV evaluation.

4. Final Remarks

Our method of evaluating EV is based on our theoretical definitions of external validity. It has become clear that, in order to achieve external validity in a practical sense, we need to identify new populations whose relationship with the population (or the population as a mixture of several populations) represented by the original sample (or multiple original samples) is reasonable and workable. In particular, we assume that each specific w which defined new populations in Section 2 corresponds to a new weighting.

Our method of evaluating external validity is purely data-driven, but theory can play an important role in valid extrapolation (Deaton, 2010; Wolpin, 2013). As discussed above, our method goes only as far as the information contained in the sample(s) (or original population(s)) allows it to go. When a researcher wants to say something about a new population with inference-relevant characteristics that are absent from the original population, he or she needs to make further assumptions and to model certain mechanisms. One fruitful line of future work could be to use a combination of theoretical and experimental approaches to measure the generalizability of those mechanisms.

Once researchers have conducted an internally valid analysis, that analysis yields an established set of findings for the specific case in question. As for the future usefulness of that result, however, what matters is its degree of EV. To design for EV, what is wanted is a sample that includes as many different subjects as possible, ones that do not necessarily represent the original population. Specifically, if, for the

but that is a research-specific issue that is outside the scope of this paper.

[10] For example, matching can be done within each sample/experiment and then the reweighting can be done across matched pairs from all samples. Pairs from a specific sample can also be given more weight than other pairs, as the researcher sees fit, depending on the purpose of the research.

population studied in an internally valid analysis, very small weights or no weight at all are assigned to some kinds of subjects, then a random sample at hand may include very few such subjects or even none at all; if this is the case, such subjects will have very little chance of being represented in new populations. This limitation of the original sample limits the assessment of EV. Thus, stratification at sampling may enhance EV analysis. Similarly, the use of non-representative samples may also facilitate EV analysis. This issue requires further investigation, however.

References

Abadie, A., and Imbens, G.W., 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica*, 74(1): 235-267.

Andrews, I., and Oster, E., 2019. "Weighting for External Validity." NBER Working Paper No. 23826.

Angrist, J. D., 2004. "Treatment Effect Heterogeneity in Theory and Practice." *Economic Journal*, 114(494): C52–83.

Angrist, J.D., and Fernandez-Val, I., 2010. "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." NBER Working Paper No. 16566.

Angrist, J.D., Imbens, G.W., and Rubin, D.B., 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, 91(434): 444-455.

Angrist, J.D., and Rokkanen, M., 2015. "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff." *Journal of the American Statistical Association*, 110(512): 1331–1344.

Banerjee, A.V., Karlan, D., and Zinman, J., 2015. "Six Randomized Evaluations of Microcredit: Introduction and Further Steps." *American Economic Journal: Applied Economics*, 7(1): 1-21.

Bertanha, M., and Imbens, G.W., 2018. "External Validity in Fuzzy Regression Discontinuity Designs." NBER Working Paper No. 20773.Cameron, C., and Trivedi, P., 2005. *Microeconometrics: Methods and Application.* Cambridge University Press.

Box, George, 1976. "Science and Statistics." Journal of the American Statistical Associations, 71(365): 791-799.

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

Cook, T., and Campbell, D., 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings.* Houghton Mifflin.

Cruces, G., and Galiani, S., 2007. "Fertility and Female Labor Supply in Latin America: New Causal Evidence." *Labour Economics*, 14: 565-573.

Deaton, A., 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature*, 48(2): 424–455.

Dehejia, R., Pop-Eleches, C. and Samii, C., 2019. "From local to global: External

validity in a fertility natural experiment." *Journal of Business & Economic Statistics*, pp.1-27.

Dong, Y., and Lewbel, A., 2015. "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models." *Review of Economics and Statistics*, 97(5): 1081-1092.

Dupas P., Karlan, D., Robinson, J., and Ubfal, D., 2106. "Banking the Unbanked? Evidence from Three Countries." NBER Working Paper No. 22463.

Efron, B., and Hastie, T., 2016. *Computer Age Statistical Inference*. Cambridge University Press.

Engle, R., Hendry, D.F., and Richard, J.F., 1983. "Exogeneity." *Econometrica*, 51(2): 277-304.

Fisher, R.A., 1956. *Statistical Methods and Scientific Inference*. Oliver and Boyd.

Galiani, S., Gertler, P., Undurraga, R., Cooper, R., Martinez, S., and Ross, A., 2017. "Shelter from the Storm: Upgrading Housing Infrastructure in Latin American Slums." *Journal of Urban Economics*, 98: 187-213.

Gechter, M., Samii, C., Dehejia, R. and Pop-Eleches, C., 2018. Evaluating Ex Ante Counterfactual Predictions Using Ex Post Causal Inference. arXiv preprint arXiv:1806.07016.

Gertler, P., Shah, M., Alzua, M.L., Cameron, L., Martinez, S. and Patil, S., 2015. How Does Health Promotion Work? Evidence from the Dirty Business of Eliminating Open Defecation. NBER Working Paper No. 20997.Hahn, J., Kuersteiner, G. and Mazzocco, M., 2015. "Estimation with Aggregate Shocks." arXiv preprint arXiv:1507.04415.

Imbens, G.W., and Rubin, D.B., 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Janzing, Dominik, Peters, Jonas, and Schölkopf, Bernhard, 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT press.

Manski, C.F., 2013. *Public Policy in an Uncertain World: Analysis and Decisions*. Harvard University Press.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. American Economic Journal: Applied Economics, 11(1), 57-91.

Rosenbaum, P.R., 2010. *Design of Observational Studies*. Springer Press.

Rosenzweig, M., and Udry, C., 2018. "External Validity in a Stochastic World:

Evidence from Low-Income Countries." Working Paper, Economic Growth Center, Yale University.

Wolpin, K., 2013. *The Limits of Inference without Theory*. MIT Press.