Theory, External Validity, and Experimental Inference: Some Conjectures

Fernando Martel Garcia and Leonard Wantchekon

The ANNALS of the American Academy of Political and Social Science 2010 628: 132
DOI: 10.1177/0002716209351519

Theory, External Validity, and Experimental Inference: Some Conjectures

By
FERNANDO MARTEL GARCIA
and
LEONARD WANTCHEKON

It is often argued that experiments are strong on causal identification (internal validity) but weak on generalizability (external validity). One widely accepted way to limit threats to external validity is to incorporate as much variation in the background conditions and in the covariates as possible through replication. Another strategy is to make the theoretical foundations of the experiment more explicit. The latter requires that we develop trajectories of experiments that are consistent with a theoretical argument. In other words, new experiments should not simply consist of changing the context of old ones, but do so in ways that explicitly test various aspects of a theory in a coherent way.

Keywords: causal inference; randomized experiments; external validity

On December 17, 2007, 1,431 poor families in New York City received a total of $740,000 as part of a new, $53 million conditional cash transfer (CCT) anti-poverty program called Opportunity New York City.[1] The design of

Fernando Martel Garcia (fmg229@nyu.edu) is a PhD student at New York University's Wilf Family Department of Politics. His research focuses on accountability mechanisms in public service delivery and how research design and statistical techniques can help us not only uncover such causal mechanisms, but also test potential improvements, make out-of-sample forecasts, and ultimately, improve our theories. He is currently involved in various field experiments in Mexico. Prior to starting his PhD, Mr. Martel Garcia worked as a staff economist for the World Bank.

Leonard Wantchekon (leonard.wantchekon@nyu.edu) is a professor of politics and economics at New York University. He taught at Yale University (1995–2000) and was a visiting fellow at the Center of International Studies at Princeton University (2000–01). He is the author of several articles on post–civil war democratization, resource curse, electoral clientelism, and experimental methods, which have appeared in the American Political Science Review, World Politics, Comparative Political Studies, Quarterly Journal of Economics, *and* Journal of Conflict Resolution. *He is the founding director of the Institute for Empirical Research in Political Economy, which is based in Benin (West Africa) and at New York University.*

DOI: 10.1177/0002716209351519

Opportunity NYC was informed to a large extent by *Progresa*, a highly regarded CCT program initiated in Mexico in 1997 as a randomized field experiment.[2] Accordingly, it is only natural to ask, will the New York initiative be as successful as its Mexican predecessor?

This is essentially a question of external validity, namely, the validity of inferences about whether a causal relationship holds over variation in treatments, outcome measures, units, and settings (Shadish, Cook, and Campbell 2002, 38). As such, it goes to the heart of policy-motivated research, with its emphasis on determining what works under what conditions. That is, what, if anything, can we say about the causal effect of similar policies under different contexts?

The latter question would not matter were it not that it arises quite often in applied work. Indeed, in the present case, the differences between Opportunity NYC and *Progresa* are just as striking as their common pedigree. To begin with, Opportunity NYC is the first time a CCT program is being tried in a large, wealthy city, whereas *Progresa* started out as a program for the rural poor. Second, the experimental units (the relevant New York households) are significantly richer in absolute and relative terms than their Mexican counterparts. Indeed, poverty in the United States might well be quite different from poverty in rural Mexico, in terms of causes, consequences, and solutions. Third, the intervention in New York is also different, in that "[it] is the first program to include a significant workforce participation component in addition to the traditional health and education components."[3]

Given significant variation in treatment, outcome measures, units, and settings, some have questioned whether Opportunity NYC will work at all:

> [New York City Mayor Michael] Bloomberg has misread the purpose of third-world conditional cash-transfer programs, and thus has misread their applicability to New York. . . . In New York, unlike in the third world, poor parents don't have to pay to send their children to school. Nor do they face the tough choice of educating the kids or having enough money to put food on the table every night. (Gelinas 2006)

According to this critic, then, *Progresa* worked by relaxing a binding budget constraint on poor Mexican households, and because such a constraint is, under this interpretation, not binding among poor New Yorkers, Opportunity NYC will probably fail.

The uncertainty surrounding the generalizability of cause and effect relationships from field experiments such as *Progresa* questions not just Mayor Bloomberg's wisdom, but the very enterprise of randomized experimentation for policymaking. Experiments, it would seem, have to offer only a deeply unsatisfying Faustian bargain between internal and external validity: Yes, we are highly certain CCTs worked in Mexico, but remain deeply ambiguous about their potential success in New York.

This article is motivated by the desire to help improve the terms of this Faustian bargain. In what follows we distinguish between the robustness and analytical approaches to external validity. The analytical approach proposes a series of theoretically motivated replications. This approach sees the problem of generalizability

as intrinsically theoretical, in that theories about causal mechanisms, constructs, and selection are what allow us to generalize beyond sampling particulars in individual cases.

In the context of the New York experiment, this approach could begin by theorizing about mediator and moderator variables that may dampen the extrapolation from *Progresa* and then design a sequence of experiments to test these hypotheses.[4] Indeed, randomized CCTs have been implemented in numerous countries, offering ample opportunity to test alternative causal mechanisms for better prediction out of sample (Rawlings 2005; Das, Do, and Ozler 2005). Moreover, we may design small tests of specific implications of the theory without having to replicate *Progresa* each time.

In contrast, the robustness approach relies on replication across various settings, treatments, outcome measures, and units. Rather than fret whether *Progresa* will work in New York, the approach is to just go ahead and test it. Such brute force replication is, indeed, an obvious path to dissolve the uncertainty. But besides the potential practical drawbacks, it has severe limitations. Suppose we run the test and the results are negative. What shall we conclude? That CCTs work in Mexico and not in New York? That they work only in poorer societies? Or in all societies but New York? That they did not work in New York in 2008-12 but may work at a later date? Should we therefore repeat the same experiment every few years? The list is potentially endless. For us, theory-driven sequences of experiments are key for optimal learning.

External validity can be greatly improved by connecting individual experiments with a theory, even if those individual experiments are not themselves theoretically grounded, especially at the early stages of a research program. Indeed, in what follows, we reemphasize the distinction between the external validity of theories associated with a research program and the particular experiments on which the program is based, thereby underlining the three-way relation between experiments, research programs, and external validity. After all, external validity is an attribute of inferences and not of any particular method (Shadish, Cook, and Campbell 2002). Indeed, contra commonly held beliefs, experimentation is key to testing theories about external validity.

Besides motivating the desire for optimal learning, the distinction between the analytical and robustness approaches has important practical consequences, as the payoff to more careful replication could be large. For example, the budget for Opportunity NYC has been set at $53 million, financed by the mayor himself and private foundations (it receives no public funds). This is a huge gamble. Some relatively inexpensive pilot testing and diagnosis may well reduce uncertainty, increasing the expected value of this and other future replications. Thus, there may be large private and external gains to be had.

Finally, two caveats: First, whereas the sharp distinction between robustness and analytical approaches is useful as a rhetorical device, it is unlikely to be borne out in practice. In reality, there is likely to be some amount of overlap, in that replications often embody some implicit prior or theory. Second, because our goal, at this stage, is to highlight some conjectures, the approach is discursive and

not explicitly deductive. In ongoing work, we formally apply the statistical learning and decision-making literatures to the notion of external validity.

This article proceeds as follows. First, we discuss what external validity is, why it is problematic and what the criteria are for evaluating external validity. We then explain what the analytical approach is. Next, we discuss the potential pros and cons. Finally, we explore how external validity relates to causality, prediction and understanding, before offering a conclusion.

What Does "External Validity" Really Mean?

Imagine testing the effect of non-partisan get-out-the-vote mailings on turnout amongst N individuals, about whom all we know is their individual identifiers $i = 1, 2, \ldots, N$. Variable Y records the turnout outcome, such that $y_i = 1$ if unit i turned out to vote and $y_i = 0$ otherwise. Suppose the treatment effect is positive and practically and statistically significant. Now, suppose we are asked to place a bet on whether the treatment will work on some new units, units $N + 1$, $N + 2$, and $N + 3$. All we know is that these units were not in the original sample. How shall we place our bet? And how will we know if we have won? That is, when is a replication result close enough to the original inference that external validity may be deemed upheld? We answer these two questions in turn.

External validity as extrapolation

Under a squared loss function, say, a common predictor for an outcome of interest y is the conditional mean $E[P(y|Treatment)]$, where $P(y|x)$ describes the density of y conditional on x.[5] Although this may seem natural, we don't know, with the information we are given in the hypothetical voting experiment mentioned above, whether the outcomes for these new units are identically and independently distributed in relation to the experimental sample. For all we know, the new units are a cat, a dog, and a goldfish, none of which vote.

What we face is a problem of predictive ambiguity as, strictly speaking, the experiment reveals nothing about the density $P'(y|Treatment)$ for the new units. Predictive ambiguity arises whenever choice or behavior depends on an objective function with an unknown probability distribution (Manski 2008). At this point, we therefore need to make assumptions. We may, for example, assume that the new units are not materially different from the old ones, in which case our choice of predictor would be justified (i.e., is in accordance with our assumptions and loss function).

The fundamental problem is that external validity involves extrapolation of treatment effects to new units or, more generally, making predictions off the support of the estimated density function (which includes the experimental setting, outcome measure, and treatment). As such, external validity claims are inherently ambiguous. Such problems with extrapolation arise whenever we ignore whether the units we want to make a prediction about belong together with the units used

to estimate the density, either because we lack observable information about these units, or because some potentially relevant characteristics may be intrinsically unobservable. This, then, is the problem in making the projection from *Progresa* to New York: we just cannot be sure these two policy experiments are sufficiently similar. But similar in what respects? In the next section, we argue that they need to be similar in theoretically relevant aspects or, alternatively, that we have a theory that allows us to bridge their differences.

As was argued above, the problem of extrapolation may be overcome in two ways (Manski 2008). First, according to the robustness approach, we may circumvent it altogether by performing a test that includes a sample from these new values (e.g., testing whether the new units vote after receiving the mailing). Of course, such trial and error can be very expensive and often impractical. A more nuanced version of this approach is to perform a pilot test of a much larger intervention, say, by randomly sampling from the target population of treatments, units, settings, and outcomes. Unfortunately, such a population is ill defined for many experimental aspects, such as treatments, outcome measures, and settings (Shadish, Cook, and Campbell 2002). For this reason, we don't think that replication using random samples from a well-defined population overcomes the problem of external validity altogether, as often this is not possible or it is too costly.

Second, the analytical approach relies on testable theories about cause and effect relationships to impose global shape restrictions, such as invariance, linearity, or monotonicity of the estimated density. For example, to deal with extrapolation one could rely on theoretically motivated assumptions regarding invariance (e.g., $P(y|x) = P'(y|x')$, $x \neq x'$). Here, we assume that the new units, despite being off the support of $P(y|x)$, are sufficiently similar in theoretically relevant aspects to the N experimental subjects that we can predict their outcomes using the estimated density. Gelinas's (2006) criticism of the New York replication, above, essentially questions this invariance assumption, by arguing that New York City is not at all similar to Mexico.[6]

Alternatively, we may assume linearity or monotonicity, which allow us to adjust for relevant differences across units, settings, and so on. The relevant assumption in this case is that the model is well specified for cases on and off the support and that relevant moderators have been incorporated as factors into the original experiment. Accordingly, to make an ex-post prediction, we collect data on the relevant covariates specified by our theory for the units we want to make a prediction about and feed these inputs into the model to get a vector of predictions as the output. If the model predicts well out of sample, we may gain further confidence in the assumption that it is well specified for some universe of cases, and so we may have more confidence in its predictions as more and more successful replications accumulate.

External validity as subjective

Now that we have, it is hoped, justified some assumptions and settled on a predictor to inform our bets, it remains for us to agree on a criterion that determines

which bets win. That is, what criterion determines external validity, or "whether a causal relationship holds over variation in treatments, outcome measures, units, and settings"?

By what criteria do we judge the success or failure of external validity? One criterion is to declare the experiment successful if the conditional estimates are identical to the ones in Mexico. However, this would yield few if any successes. Even if the underlying data-generating processes were the same and we could condition away differences between experiments, sampling variability would still make exact matches between estimated parameters highly unlikely.

To fix ideas, suppose large enough samples allowed us to ignore sampling variability altogether. Would a moderate difference between the estimated parameters question the external validity of the inferences from *Progresa*? And what if, despite this moderate difference, the estimate still preserves the direction and practical relevance of the causal effect across both settings? Does the original prediction remain externally valid? And, if not, what exactly do we mean by the external validity of a causal inference?

Shadish, Cook, and Campbell (2002, 34) interpret validity as "the approximate truth of an inference" yet, to the extent that we never really get to observe truth, this is highly problematic. Besides, even if we did, we still need to deal with the issue of "moderate" differences in true parameters as, for most practical purposes, exact matches are not what we are after. Instead, we propose a more pragmatic interpretation of validity, one built on the pillars of prediction and statistical decision theory.

Under the pragmatic criterion we propose, our concern with external validity stems from the need to decide whether or not to implement a program like *Progresa*, say, in New York. That decision is often in the hands of a politician or high-level bureaucrat. If so, the preferences, opportunities, and constraints of the decision-maker should enter into the ex-post assessment of the validity of the inference (Granger and Machina 2006). Consequently, to a first approximation, it is decision-makers that determine the criteria for success.[7]

Although different decision-makers are likely to have different loss functions, and hence evaluate external validity differently, it is not a case of anything goes. At a minimum, an externally valid extrapolation of a causal relation should be one that, once realized, results in a causal effect that is in the same direction as the prediction. In addition, the size of the causal effect must remain practically significant at some predetermined level. That is, the decision-maker must evaluate the outcome in a fashion consistent with his ex-ante loss function. An implication of this understanding, though one not pursued in this article, is the need for better use of statistical decision theory in political science.[8]

To recap, external validity is problematic because it involves extrapolation, that is, making predictions off the support of the estimated density. Whether the inferences regarding the applicability of the parameters of that density off the support are externally valid or not will therefore depend on the accuracy of its predictions out of sample—that is, across combinations of treatments, outcome measures, units, and settings not in the original sample. The accuracy of these forecasts, in turn,

depends on a vector of ex-post prediction errors and a loss function evaluating these that is specific to the decision-maker (Manski 2008). Accordingly, we are open to the idea that the validity of an inference lies, within the bounds defined above, in the eye of the beholder. External validity is a matter of degree.

The Analytical Approach to External Validity

In the previous section, while defining what we mean by external validity, we suggested the idea that theory could be used to justify global shape restrictions to make extrapolations. In this section, we expand on precisely the type of theorizing needed to justify such restrictions. We do so in four parts. First, we provide some examples of why general knowledge is, indeed, theoretical knowledge, explaining how theories about constructs, causal mechanisms, and compliance and selection are key for external validity. Second, we provide some examples. Third, we consider some pros and cons that might be made against the analytical approach. Fourth, we discuss what is and is not new in this scheme. To avoid a long digression, for the instant purposes we will define theories as falsifiable statements about causal relationships between classes of events. These theories may be complex and micro-founded, or more simple and aggregate; either way, they ought to explain causal regularities rather than individual particulars.

Theories about constructs, causal mechanisms, and compliance and selection

Theorizing for generalizability comes at three levels: theories about constructs, theories about causal mechanisms, and theories about compliance and selection. First, theories about constructs are essential in order to be able to talk meaningfully about the results of experiments. For example, Wantchekon (2008) studies the impact of informed campaigns on voting behavior. Yet, before we can generalize his results from the particular implementation, we need to be clear what the theoretically relevant attributes of a campaign are that make it an informed campaign, the idea being that campaigns sharing these essential features would have the same causal effect ceteris paribus.

Second, we can specify a theoretical causal mechanism, one where both mediator and moderator variables are conjectured, measured, and tested, and what we believe are irrelevancies are left out ex ante (unless the budget allows for more testing). With the parameters of this fully specified model at hand, we can then measure the level and prevalence of such moderators and mediators in the new target population of interest, and use them as inputs into the model to predict the average treatment effect in that population. An obvious application of this is in studies of the genetic basis of disease and their interaction with potential remedies.

Third, most social science experiments are, explicitly or implicitly, encouragement designs (Horiuchi, Imai, and Taniguchi 2007). As such, treatment assignments do not guarantee compliance. To the extent that compliance rates

vary across treatments, outcomes, units, and settings, so will the average causal effect. To provide better predictions one could then proceed as above, modeling the population by characterizing it as a combination of compliers, defiers, and never- and always-takers (Angrist, Imbens, and Rubin 1996), according to their individual attributes; then using this information to predict the proportion of compliers and others in the new target population; and, finally, computing the implied average treatment effect (or other effect of choice) in the target population using the predicted proportions (see Frangakis, Rubin, and Zhou [2002] and Horiuchi, Imai, and Taniguchi [2007] for specific applications).

In practice, all three issues—construct validity; systematic differences in levels of mediators and moderators; and systematic differences in shares of compliers, defiers, and never- and always-takers—will impact our estimates of average treatment effects, and so all three need to be considered simultaneously. Accordingly, what we ultimately need are good theoretical models of the latent construct, the causal mechanism, and selection into treatment—what are commonly referred to as structural equation models.[9] Such models embody global shape restrictions that, if correct, are what allow us to make good predictions out of sample. However, our purpose is not to argue for a return to the large structural equation models of the type once sponsored by the Cowles Commission for Research in Economics, say, but simply to note that conceptualizing such models may help in the design of research programs, even if individual experiments remain much less complicated.

Theory, research programs, and individual experiments

In practice, most experiments should not attempt to estimate such complex structural models, nor do they need the backing of a fully specified theory. Indeed, in the early stages of a research program concerns about external validity are often secondary to construct or internal validity. Rather, our argument is that, to the extent that external validity is desired in mature research programs, there are more efficient ways of achieving it than testing large structural models or blindly replicating, as in the robustness approach. Rather, experiments ought to quite deliberately test the appropriateness of assumed global shape restrictions.

For example, Mook (1983) cites the example of Ekman and Friesen (1971), who asked whether recognition of emotional facial expressions depended on culture. Rather than do innumerable replications across all possible cultures, they theorized that if facial expression were interpreted similarly across cultures, then this must be true across the most distant cultures. Hence, they "stress tested" the theory by comparing Americans to the most distant culture they could think of, the Fore of Papua New Guinea. The finding that they both recognize happiness in each other suggests the universality of emotional expression.

Similarly, going back to the Mexico–New York example above, if we think that a history of substance abuse is a significant moderator of the effect of CCTs, say, then we may design an experiment that stress tests this aspect. Despite its narrow focus, the fact that this experiment is embedded in a larger research project may

inform us greatly about the external validity of predictions based on *Progresa*, insofar as it may allow us to condition our expectations on the prevalence of substance abuse in the New York target population.

Another example is the Benin electoral experiments (see Wantchekon 2003, 2008). One of the findings of the 2001 experiment is that voters are more likely to react positively to a "public goods" message when it comes from a co-ethnic candidate. A possible explanation of this result is that voters trust a candidate from their ethnic group more than they trust a candidate from another group. This means that the mediating variable between ethnic ties and vote is trust, or the credibility of the candidate. By testing in the context of the following experiment in 2006 the relationship between credibility of candidates and voting behavior, Wantchekon (2008) improved the external validity of the results of the 2001 experiment.

Pros and cons of the analytical approach

The analytical approach to external validity we are proposing may be preferable for at least three reasons. First, the question of external validity is largely a theoretical one. Research on external validity asks not just whether a CCT program will work in New York, say, but why or why not. That is, it demands an explication in terms of a causal mechanism or relevant differences between units in and out of sample. Indeed, for policy analysis, as well as for the purpose of scientific advancement through comparative research, the question of where, under what conditions, and among which subpopulations a treatment is likely to work is often as interesting as whether or not the treatment worked in the first place (Heckman 2005). As such, the analytical approach is of interest in and of itself.

Second, answering these questions may provide us with tighter bounds on future predictions out of sample, and increase our confidence in the maintained shape restrictions. Moreover, a program of research that focuses on external validity will subject theories to the strongest possible tests, out-of-sample tests, thereby offering the potential for large updates in our priors.

Third, we conjecture that theoretically driven research programs will permit us to design experimental replications for optimal learning—or, to paraphrase Milton Friedman, to come up with sequences of experiments that explain as much as possible in the shortest possible sequence.

This being said, the analytical approach is not foolproof. One could argue that the external validity criticism applies just as readily to a more fully specified causal mechanism, construct, and selection process than to a simple one. After all, moderator or mediator variables in one setting may be different, or have different impacts, in new settings, populations, treatments, or outcome measures, and so extrapolation is not possible. For example, household income may be a moderator of the effects of *Progresa* in Mexico but not in New York, perhaps due to some unobserved interacting variable. All this is certainly possible, but there are at least three powerful rejoinders.

First, we can test this criticism. If the moderators are found to behave similarly in Mexico and New York, then we have more confidence in projecting treatment effects to Los Angeles and Kuala Lumpur, say, than if we had replicated without a theory. The latter does not allow us to make accurate predictions because we are not using all the available information efficiently, by conditioning on the relevant moderators and mediators. As a result, robustness replication yields more uncertain predictions and represents a very inefficient way of cumulating knowledge. Our view is that theorists explain and empiricists condition; doing neither greatly limits external validity.

Second, this is a *reductio ad absurdum* criticism. It simply negates the whole basis of comparative research. If causal and selection mechanisms are different across all variations in settings, populations, treatments, or outcome, then we have no basis for generalizations. Without theorizing, each construct, treatment, outcome measure, unit, and setting is unique—the possibility of scientific learning denied. This may well be the case, but it is a testable proposition.

Third, the burden of proof is on generalizability skeptics to propose why they think the inference is not generalizable. Indeed, this is what Rosenbaum (2002, 9) calls tangible criticism: "a specific and plausible alternative interpretation of the available data; indeed a tangible criticism is itself a scientific theory, itself capable of empirical investigation." In contrast, dismissive criticism "rests on the authority of the critic and is so broad and vague that its claims cannot be studied empirically."[10]

To sum up, think of the replication process as an optimal learning problem. Our contention is that the analytical approach will require fewer iterations to achieve a given uncertainty tolerance (about external validity) than the robustness one or, alternatively, reduce uncertainty further for any given number of replications. As Guala and Mittone (2005, 499) state, "external validity is an important issue for experimenters *and* theorists alike" (emphasis in original).

The difference between the robustness and analytical approaches somewhat resembles the difference between active and passive learning (Castro et al. 2008) or the problems faced in dynamic control programming. As that literature implies, there are trade-offs to be made between the analytical and robustness approaches. For example, testing some theories will require factorial designs, which in turn may involve larger samples. Yet this may still be cheaper than testing hypotheses separately and, to the extent that they generate sensible findings, may substantially improve our predictions out of sample.

Old wine in new bottles?

Theory already permeates everything we do, from the questions we ask, to the data we collect, to how we define our concepts.[11] For example, take one of the most successful and comprehensive experimental research programs in political science, the series of field experiments on voter mobilization spawned by the work of Gerber and Green (2000) and reviewed in Green and Gerber (2008). The use of theory is clear in the choice of treatments: researchers have studied

whether mass mailing, door-to-door canvassing, and so on impact turnout, but not the impact of mailings printed with Times New Roman versus Arial font, as we have no theoretical basis for presuming these might be causally relevant.

The fact is, most replications are already theory driven, explicitly or not. For example, the finding by Gerber and Green (2000) that phone calls are ineffective in getting out the vote relative to personal face-to-face contact motivated Nickerson (2006) to test whether this was due to the face-to-face component or the extra personal attention associated with door-to-door canvassing. He finds that the quality of the phone calls matter and that brief, nonpartisan phone calls can raise voter turnout if they are sufficiently personal. We can use this finding to help us predict how other means of communication may fare conditional on the degree of personal attention provided. Also, at times, these replications shed light on unexpected results, motivating new theory and further experimentation (Gerber 2004).

If theory already permeates many of our experimental research programs, what is our point? First, we want to dispel criticisms that individual experiments have little or no external validity: so long as they contribute to general theories of voter behavior, say, experiments may expand the range of prediction significantly. Moreover, we can test external validity claims using experiments strong on internal validity. Second, for their part, experimenters ought to make more explicit the theoretical context of their experiments, even if it is only an enumeration of potential causal pathways lacking in formalisms. Given that experimental replication is highly decentralized, it is up to each replication to do its part in expanding the external validity of the whole.

When Are Concerns about External Validity Justified? Prediction versus Understanding

We began this article by noting the policy relevance of external validity questions, in particular because of the perceived need to make predictions about causal effects out of sample. But is predictive success the hallmark of a good theory? What about explanation and understanding? And when should experimentalists focus on external validity? Besides, what about the idea that the best predictors are often atheoretical associations à la Sims (1980)?

Prediction and understanding are closely related. Yet an example due to Rubin (1996, 475) perhaps best highlights their differences: suppose an unfair coin yields heads with probability .6. A model that predicts heads with probability 1 will get it right 60 percent of the time, one predicting heads with probability of .6 will get it right 52 percent of the time. On the basis of predictive success, we would be tempted to choose the wrong model, even if it cannot explain the observed sequence of tosses and, in particular, the 40 percent of tails. So prediction cannot be all that there is to it. However, our point is not that external validity ought to be the only goal of science; rather, our claim is that the best way to test external validity is to test predictions out of sample. Different goals require different tests.

For example, in a widely cited article, Mook (1983) criticized the preference then prevalent in psychology for experiments with strong external validity. His argument relied on a distinction between prediction and understanding, or, as he put it, between two modes of research: the analogue and the analytical models. In the analogue model, the objective is to model the real world for prediction purposes. Thus, the variables that account for the most real-world variance are the most important, since they are the ones that speak most directly to our ability to make predictions about causal relations.[12] In the analytic model of research, by contrast, the objective is to understand the workings of a system, to test the internal validity of theories. These theories may apply to real life, but there is no attempt at generalization at this point.

In a similar vein, Przeworski (2007) is interested, not only in the effects of causes, but also in the causes of effects: the list of factors $X_1, X_2, \ldots X_n$ that explain the observed (in nature) outcome Y, say lung cancer prevalence.[13] Przeworski's point is that we often desire to understand not only what the potential causes of Y are, but also what the actual causes of Y are in a particular population and how these come about. Demonstrating that X can cause changes in Y is a necessary but insufficient condition for the inference that X explains Y, or that X can be used to manipulate changes in Y in any particular target population in any period. Translated to Mook's language, the analytical mode of research (or the effects of causes) asks, "Could X have caused Y?" whereas the analogue mode of research (or the causes of effects) asks, "Does X typically cause Y?"

Note that our answer to these questions may significantly influence how we answer their corollary, "Can X be manipulated so as to change Y in a desired direction?" This is important for two reasons: first, because by knowing the actual causes of a disease, say, we might be better able to design prevention measures; and, second, because causes that are effective in the lab may not be efficacious in the field.

For example, take the efficacy of bed nets in preventing malaria. Under control conditions bed nets have been shown to be highly effective in preventing malaria: households randomly "treated" with bed nets experience a reduction in malaria incidence relative to households randomly allocated to control conditions. These controlled experiments identify the "effects of causes"—in this case, bed net use reduces malaria incidence. Based on this evidence, numerous programs have been implemented that freely distribute bed nets in areas of high malaria incidence. And yet, to date, the jury is still out as to the effectiveness of these interventions in reducing malaria incidence in the treated areas. Why? One answer is lack of compliance—providing a free bed net does not imply that it will be used appropriately.

This example illustrates the point that just because X can be shown to cause changes in Y, it does not follow that it explains any of the observed variance in Y in the real world nor, indeed, that it can be an effective cause in the real world (a problem of external validity), where we may lack sufficient control to ensure full compliance. In other words, there are other factors (potentially unobserved) that moderate the causal effect in real applications. In Mook's terminology, we gain

understanding of potential causes for reducing malaria, but we may still end up with bad policy predictions.

Understanding what the actual causes of some phenomenon are should inform our predictions and research. This is what motivated psychologist Egon Brunswick to advocate "representative designs" where, in particular, levels of the causal variable in question would be chosen according to their conditional distribution in the natural world (Albright and Malloy 2000).

Accordingly, the idea that distribution of free bed nets will somehow generate the treated counterfactual observed in controlled trials may appear far-fetched. For all we know, the given distribution of bed nets in any country is an equilibrium—everyone who wants one has one—so reducing their price to zero may have a tiny effect in that, at some point, there is no longer a binding budget constraint (Gelinas's [2006] point above). In fact, it is entirely possible that there is simply no way to generate the experimental counterfactual in the field without at the same time changing the levels of other (potentially unobserved) covariates. Perhaps malaria eradication requires a process of modernization that includes improved education as to the pathogenic nature of the disease, better drainage, urbanization, air-conditioning, better medical facilities, and so on—which, by the way, speaks to the important development literature on the sequencing of reforms and the need for theories of change.

Linking external validity to the capacity to make reasonable causal predictions out of sample in no way undermines the importance of theory to the enterprise, nor the value of understanding. Prediction and understanding may have slightly different goals, but, ultimately, a good measure of useful knowledge is a test of the external validity of its predictions. In addition, sound understanding of the actual causes of effects may help us theorize about causal mechanism and optimal interventions.

Conclusion

In this article, we argue that claims to external validity of randomized field experiments are stronger when theoretical connections between experiments are established and tested. In an experiment that establishes a causal relationship between the two variables (the treatment and an outcome of interest) under a set of conditions, we can improve external validity in at least two ways: (1) by replicating the relationship between the two variables under new conditions (the robustness approach) or (2) by establishing that the relationship is mediated or moderated by the set of variables—that is, the analytical approach.

We believe the analytical approach may turn out to be more effective than the robustness approach. This is because the mediator is likely to represent a larger set of experimental conditions. We recommend that follow-up experiments be primarily focused on testing the theoretical argument of original experiments, instead of simply replicating them in a different context. If external validity is the Achilles' heel of randomized experiments, then testing mechanisms underlying already established causal relationships should be the top priority of the experimental research.

Notes

1. "[Conditional cash transfer] programs provide monetary incentives to households living in poverty when they complete activities aimed at increasing human capital development and breaking the cycle of poverty." From http://www.nyc.gov/html/ceo/html/programs/opportunity_nyc.shtml (accessed February 14, 2009). For an overview of Opportunity NYC see de Sá e Silva (2008).

2. Rockefeller Foundation, http://www.rockfound.org/efforts/nycof/opportunity_nyc.shtml (accessed February 14, 2009). The 2002 successor program to *Progresa* is called *Oportunidades*.

3. Rockefeller Foundation, ibid.

4. On the definition of moderator and mediator, see Baron and Kenny (1986). There are numerous ways to label variables in the context of a causal mechanism, yet "conceptually, moderators identify on whom and under what circumstances treatments have different effects. Mediators identify why and how treatments have effects" (Kraemer et al. 2002, 877). The former may be understood as interaction effects, whereas the latter identify possible mechanisms by which the treatment comes to have its effect. On the complexities of testing mediator effects, see Bullock, Green, and Ha (2008).

5. Let y_i denote test scores for pupil i, T_i denote whether he was randomly assigned to receive school vouchers or not, x_i be a set of covariates strongly correlated with y, and ε_i be a random (iid) error term such that we may write $y_i = \alpha + \beta_0 T_i + x_i\beta + \varepsilon_i$, $i = 1, 2, \ldots N$. We are interested in $\partial Y/\partial T$, or β_0. Accordingly, the question of whether *Progresa* will work in NYC asks, loosely, whether the estimated β_0 in the NYC sample will be statistically significant and of same sign and magnitude as in Mexico. We may test this directly with the NYC program because it is randomized. If, instead, everyone got treated in NYC, we would have only ex ante and ex post measures with no ex post controls. Although not ideal, we could use pre-test data to predict outcomes in the absence of treatment by generating a predicted counterfactual. Finally, note that it is possible to have $\partial Y/\partial T > 0$ and yet $\Delta Y = 0$ if some other event acted in the opposite direction (say a teacher strike). This is why, in the absence of randomized replication, it is crucial to have good predictors of Y, as our inferences will depend on the model for the counterfactual being correctly specified.

6. According to her, the poor in New York have access to free education and their children are idle, features that do not rhyme with the Mexican story of binding budget constraints keeping Mexican children at work and away from school.

7. At this stage, for simplicity, we ignore the issue of who the decision-maker is, whether she is a high-level bureaucrat, a politician, a set of voters, some abstract welfare maximizer, or, indeed, a researcher evaluating a theory. We will simply postulate a decision-maker, typically a politician, and leave it at that, as the relevant person may differ between applications. We are aware this renders scientific knowledge somewhat subjective, but then again, a long literature in the philosophy of science questions the possibility of objective science.

8. See Gerber, Green, and Kaplan (2002) for a rare exception in political science, and Berger (1985) and Manski (2008) for a more general discussion. This has three important corollaries, especially for policy-motivated experiments, that often go unappreciated. First, only if the decision-maker is planning to repeat the experiment elsewhere will she care about its conventional statistical significance over and beyond its practical significance. Second, having a well-specified loss function allows the researcher to move away from simple point estimation, relaxing some assumptions and the somewhat exaggerated obsession with bias (see Rosenbaum [2002] for a discussion of sensitivity analysis and Manski [2008] for interval estimation). As such, the unbiased estimation made possible by randomized experiments may only be needed whenever good enough priors on the bias are lacking, preventing us from adjusting observational estimates to correct for the bias (Gerber, Green, and Kaplan 2002). Third, we need more empirical data on the kind of decision criteria used by policymakers: Bayes, Maximin, or Minimax-regret, say. That is, on the basis of what criteria do policymakers choose policy experiments? Laboratory experiments on high-level officials may help reveal these.

9. This is very much in line with the argument made in Heckman (2008).

10. Experimenters ought to make the experiment as sound as possible to begin with, by stating potential mediators, say, and, whenever the budget allows, testing them. Our point is that in order to help discover whether a design is indeed flawed, the critic needs to specify why he thinks the design is faulty in the first place. Off-the-cuff criticism of the sort "experiments have no external validity" are, on this account, unhelpful and often dismissive.

11. See, inter alia, McDermott (2002); Druckman et al. (2006); Guala and Mittone (2005); Levitt and List (2007); Lucas (2003); Lynch (1999); Moffitt (2004); Shadish, Cook, and Campbell (2002); Schram (2005).

12. We distinguish between making causal predictions of the type "a change in X will cause a change in Y of Δ percent," say, versus simply forecasting Y, say, although to reduce the variance of our causal predictions, good forecasts of Y are often welcome.

13. Confusingly, Mahoney and Goertz (2006) interpret "causes of effects" as explaining individual cases, e.g., what caused Joe Doe to get lung cancer. This, in our view, is not the standard interpretation. We adhere to Przeworski's and Heckman's interpretation.

References

Albright, Linda, and Thomas E. Malloy. 2000. Experimental validity: Brunswik, Campbell, Cronbach, and enduring issues. *Review of General Psychology* 4 (4): 337-53.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91 (434): 444-55.

Baron, Reuben M., and David A. Kenny. 1986. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51 (6): 1173-82.

Berger, James O. 1985. *Statistical decision theory and Bayesian analysis*. 2nd ed. New York: Springer.

Bullock, John G., Donald P. Green, and Shang E. Ha. 2008. Experimental approaches to mediation: A new guide for assessing causal pathways. Unpublished manuscript, Yale University. Accessed online at http://www.ipeg.org.uk/events/field_experiments/Documents/A%20Critique%20of%20Conventional%20Mediation%20Analyses%20--%20Bullock%20Green%20and%20Ha.pdf

Castro, Rui M., Charles Kalish, Robert Nowak, Ruichen Qian, Timothy J. Rogers, and Xiaojin Zhu. 2008. Human active learning. Technical report. Columbia University, New York.

Das, Jishnu, Quy-Toan Do, and Berk Ozler. 2005. Reassessing conditional cash transfer programs. *World Bank Research Observer* 20 (1): 57-80.

de Sá e Silva, Michelle Morais. 2008. Opportunity NYC: A performance-based conditional cash transfer programme. A qualitative analysis. Working Paper 49, International Poverty Centre, Brasilia, and Columbia University, New York.

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. The growth and development of experimental research in political science. *American Political Science Review* 100 (4): 627-35.

Ekman, Paul, and Wallace V. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17:124-29.

Frangakis, Constantine E., Donald B. Rubin, and Xiao-Hua Zhou. 2002. Clustered encouragement design with individual noncompliance: Bayesian inference and application to advance directive forms. *Biostatistics* 3:147-64.

Gelinas, Nicole. 2006. New York isn't Mexico. *City Journal*. Accessed from http://www.city-journal.org/html/eon2006-10-20ng.html.

Gerber, Alan S. 2004. Does campaign spending work? Field experiments provide evidence and suggest new theory. *American Behavioral Scientist* 47 (5): 541-74.

Gerber, Alan S., and Donald P. Green. 2000. The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review* 94 (3): 653-63.

Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2002. The illusion of learning from observational research. Institution for Social and Policy Studies Working Paper, Yale University, New Haven, CT.

Granger, Clive W. J., and Mark J. Machina. 2006. Forecasting and decision theory. In *Handbook of economic forecasting*, vol. 1, 81-98. Amsterdam: Elsevier.

Green, Donald P., and Alan S. Gerber. 2008. *Get out the vote: How to increase voter turnout*. 2nd ed. Washington, DC: Brookings Institution.

Guala, Francesco, and Luigi Mittone. 2005. Experiments in economics: External validity and the robustness of phenomena. *Journal of Economic Methodology* 12 (4): 495-515.

Heckman, James J. 2005. The scientific model of causality. *Sociological Methodology* 35 (1): 1-98.

Heckman, James J. 2008. Econometric causality. Social Science Research Network Electronic Library.
Horiuchi, Yusaku, Kosuke Imai, and Naoko Taniguchi. 2007. Designing and analyzing randomized experiments: Application to a Japanese election survey experiment. *American Journal of Political Science* 51 (3): 669-87.
Kraemer, Helena C., G. Terence Wilson, Christopher G. Fairburn, and Stewart W. Agras. 2002. Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry* 59 (10): 877-83.
Levitt, Steven D., and John A. List. 2007. What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives* 21 (2): 153-74.
Lucas, Jeffrey W. 2003. Theory-testing, generalization, and the problem of external validity. *Sociological Theory* 21:236-53.
Lynch, John G., Jr. 1999. Theory and external validity. *Journal of the Academy of Marketing Science* 27 (3): 367-76.
Mahoney, James, and Gary Goertz. 2006. A tale of two cultures: Contrasting quantitative and qualitative research. *Political Analysis* 14 (3): 227-49.
Manski, Charles F. 2008. *Identification for prediction and decision*. Cambridge, MA: Harvard University Press.
McDermott, Rose. 2002. Experimental methodology in political science. *Political Analysis* 10 (4): 325-42.
Moffitt, Robert A. 2004. The role of randomized field trials in social science research: A perspective from evaluations of reforms of social welfare programs. *American Behavioral Scientist* 47 (5): 506-40.
Mook, Douglas G. 1983. In defense of external invalidity. *American Psychologist* 38 (4): 379-87.
Nickerson, David W. 2006. Volunteer phone calls can increase turnout: Evidence from eight field experiments. *American Politics Research* 34 (3): 271-92.
Przeworski, Adam. 2007. Is the science of comparative politics possible? In *The Oxford handbook of comparative politics (Oxford handbooks of political science)*, ed. Carles Boix and Susan C. Stokes, 147-71. Oxford, UK: Oxford University Press.
Rawlings, Laura B. 2005. Evaluating the impact of conditional cash transfer programs. *World Bank Research Observer* 20 (1): 29-55.
Rosenbaum, Paul R. 2002. *Observational studies*. New York: Springer.
Rubin, Donald B. 1996. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 91 (434): 473-489.
Schram, Arthur. 2005. Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology* 12:225-37.
Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. 2nd ed. New York: Houghton Mifflin.
Sims, Christopher A. 1980. Macroeconomics and reality. *Econometrica* 48 (1): 1-48.
Wantchekon, Leonard. 2003. Clientelism and voting behavior: Evidence from a field experiment in Benin. *World Politics* 55:399-422.
Wantchekon, Leonard. 2008. Expert information, public deliberation and electoral support for good governance: Experimental evidence from Benin. Mimeo. Paper presented at the Freeman Spogli Institute for International Studies on December 2nd 2008, in Palo Alto, California. Accessed online at http://fsi.stanford.edu/events/expert_information_public_deliberation_and_electoral_support_for_good_governance_experimental_evidence_from_benin/