Designing Randomized Controlled Trials with External Validity in Mind

Sylvain Chassang Samuel Kapon*,†

Princeton University

December 7, 2022.

Abstract

This paper describes a number of strategies that experimenters may use to improve the external validity of their own findings, and of their research field as a whole. The paper emphasizes a dynamic view of research processes, in which learning about treatment and treatment adoption does not cease after a given study is performed. External validity need not be an unattainable goal in such a context. However, because researchers today need not be the same as researchers and policy-makers tomorrow, dynamic research processes are affected by research externalities, i.e. research practices that have high social value but low private returns. The paper identifies several of these research externalities and argues that funding organizations can have a significant impact at a relatively modest cost by subsidizing external-validity add-on modules specifically targeting research externalities.

KEYWORDS: experiment design, external validity, research externalities, structured speculation, online learning, adaptation, option value.

*Contact: Chassang, chassang@princeton.edu; Kapon kapon@princeton.edu.

†We are grateful to the Alfred P. Sloan Foundation for providing the impetus for this project, as well as funding under grant G-2020-14052. We are especially thankful to Miriam Bruhn, Angus Deaton, Pascaline Dupas, Dean Karlan, David McKenzie and Jonathan Morduch for providing extensive comments, and to Thomas Leavitt for discussing the paper at the EGAP and CSDC workshop on "Knowledge Accumulation and External Validity: Implications for Design and Analysis" (2022). We owe an extensive intellectual debt to Abhijit Banerjee, Gerard Padró i Miquel and Erik Snowberg for many conversations on this topic.

1 Introduction

This paper seeks to identify concrete steps that individual experimenters can take to improve the external validity of their findings, and to support the external validity of findings generated by their field of research. The paper emphasizes the view that research is a dynamic process in which treatment adoption is not once and for all, but rather adapts to novel facts. This view expands the set of steps that can be taken to improve external validity, but also highlights that many of these steps have positive externalities with large social value but low private returns for individual researchers. The paper argues that funders seeking to improve external validity can maximize their return on investment by subsidizing positive research externalities via add-on modules designed to augment any baseline research proposal.

External validity is a difficult problem because it is fundamentally about extrapolation to novel environments for which little data is available. In contrast, internal validity is concerned with the correct evaluation of treatment effects in the context in which studies are performed. However, as the recent replicability crisis affecting many of the social sciences highlights, even internal validity is not a given (Simmons et al., 2011). For this reason, taking steps to ensure internal validity are a pre-requisite to any attempt at enhancing external validity. This includes pre-registering studies and key endpoints, ensuring the timely reporting of study findings, and facilitating access to replication and contradictory data.

A focal approach to external validity has been to provide methods and conditions under which findings from one study context can be consistently extrapolated to a different study context. This strand of work operates under the assumption that different contexts differ only in the distribution of participants' covariates, and that outcomes are independent of context conditional on these covariates (Hotz et al., 2005). Under this strong assumption, consistent extrapolation can be achieved across contexts by appropriately reweighting conditional treatment-effects. Dehejia (2005) and Heckman and Vytlacil (2005) emphasize the use of this approach to simulate out counterfactual treatment policies in which treatment is

targeted according to estimates of treatment value elicited from program administrators, or from the recipients themselves. While consistent extrapolation is only feasible under strong assumptions, researchers can take steps to maximize the chances that the assumptions do hold, including measuring rich covariates, studying contexts with diverse participants, and testing for consistent extrapolation across local contexts.

The paper's main thesis is that there is greater scope for improving external validity when research is viewed as a dynamic process where predictions from previous studies can be refined in view of novel evidence. This approach echoes Deaton and Cartwright (2018)'s view that "RCTs can play a role in building scientific knowledge and useful predictions but they can only do so as part of a cumulative program, combining with other methods, including conceptual and theoretical development, to discover not 'what works', but 'why things work'." Interestingly, results from the online learning literature (Blackwell, 1956, Hannan, 1957, Foster and Vohra, 1999) imply that it is possible to build prediction systems that match the performance of the best expert in hindsight, from any finite group of experts. This provides a robust basis for "as-good-as-possible" extrapolation, even when contexts can be arbitrarily different.

This dynamic view of research and policy adoption also suggests other dimensions of interest for researchers, in particular, building short-term surrogate endpoints (which facilitate rapid exploration), and evaluating treatments as real options that can be abandoned or not by the policy-makers that implement them. The difficulty is that because researchers today are likely different from the researchers and policy-makers that will use the research tomorrow, research practices that have positive social value may have low private returns to experimenters. There are significant externalities between researchers seeking to document treatment effects in their context of interest, and future users of study findings, whether they are other researchers or policy-makers.

Takeaways for funders. We believe that funders can play a catalytic role towards greater external validity by setting up programs encouraging researchers to internalize research externalities, i.e. research practices that have low private returns for the researchers, but high social returns for other researchers, and policy-makers.

Research Practice	**Practical Use**
Pre-register trial design and key endpoints	Lets subsequent researchers and policy makers correctly assess the strength of evidence from a given study
Timely reporting of endpoints	Ensures that available findings do not suffer from reporting bias
Structured speculation about mechanisms and generalizability	Guides adoption in new contexts; provides subjective input to extrapolation methods; clarifies remaining uncertainty
Link findings to previous endpoints and speculation	Facilitates exhaustive search of relevant information across studies
Test validity across local contexts	Helps evaluate generalizability of findings

Table 1: Experimental research practices that require little change in the research design.

Tables 1 and 2 summarize research practices likely to have low private benefits but high social value. Table 1 lists research practices that require little change in the experimental design (e.g. preregistration, standardizing measurement, timely reporting . . .), while Table 2 lists research practices that require some change in the experimental design (e.g. measuring richer covariates, collecting data at multiple stage to test surrogate endpoints . . .). The paper describes the rationale behind these practices in detail.

We believe that resources earmarked towards improving external validity can be used for maximum systemic impact by subsidizing add-on modules to existing research proposals, targeting the research practices highlighted by Tables 1 and 2 for as many studies as possible. This add-on module strategy contrasts with, and may be complementary to a more concentrated approach supporting a few large scales studies specifically designed to

Research Practice	Practical Use
Measure rich covariates	Input for extrapolation models using data from a given study
Document context	Input for extrapolation models using data from multiple studies
Include standardized measures of context and endpoints	Facilitates comparisons and aggregation of findings across different studies
Experiment in contexts with heterogeneous populations	Ensures that study context can be used to learn about other environments through reweighting.
Vary context	Provides a sense of global generalizability of findings
Assess predictability from early outcomes	Lets policymakers assess speed with which one can assess fitness to own context
Provide and test surrogate endpoints	Speeds up ability to make policy-relevant choices in novel contexts
Implement random roll-out design	Lets policymakers assess option value of experimenting with policy

Table 2: Experimental research practices that require changes in the research design.

investigate external validity.

2 Framework and Challenges

A decision-maker is choosing between two policy options $\tau \in \{0, 1\}$. For simplicity, one can think of $\tau = 1$ as a novel policy, while $\tau = 0$ represents a status quo option. Following the Neyman-Rubin potential outcomes framework (Rubin, 2005), let $Y_i(\tau) \in \mathbb{R}$ denote the outcome of individual i following treatment decision τ.

The decision-maker cares about the impact of the policy on the outcomes $Y_i \in \mathbb{R}$ of individuals $i \in I$, drawn from a relevant treatment population I. For simplicity, we assume that the decision-maker seeks to maximize the average outcome in the relevant treatment population:

$$\max_{\tau \in \{0,1\}} \mathbb{E}\left[\frac{1}{|I|}\sum_{i \in I} Y_i(\tau)\Big|\mathcal{F}\right] \tag{1}$$

where $\mathcal{F}$ denotes the information available to the decision-maker at the time of decision making.[1] Note that the real-number Y_i may aggregate multi-dimensional outcomes (e.g. an educational program may affect both earnings and health outcomes) into a welfare index net of program costs.

In this setting, the decision-maker's optimal policy is entirely determined by the expected treatment effect given information

$$\mathbb{E}\left[\Delta Y|\mathcal{F}\right] \quad \text{where} \quad \Delta Y = \frac{1}{|I|}\sum_{i\in I} Y_i(1) - Y_i(0).$$

The optimal policy given information $\mathcal{F}$ sets treatment $\tau \equiv \begin{cases} 1 & \text{if } \mathbb{E}[\Delta Y|\mathcal{F}] \geq 0 \\ 0 & \text{otherwise} \end{cases}$.

Covariates, contexts and data. Individuals $i \in I$ differ by their individual characteristics, captured by covariates $X_i \in \mathcal{X}$ (e.g. gender, age, education, income...), but also by the context $C \in \mathcal{C}$ in which they evolve (e.g. country, year, macroeconomic conditions, institutional environment, treatment implementation).

We assume that potential outcomes $Y_i(\tau = 0)$ and $Y_i(\tau = 1)$ are drawn independently, each from a cumulative distribution function $F(y|X_i, \tau, C)$ that depend on the covariates X_i of person i, treatment status $\tau \in \{0, 1\}$, as well as on the context C in which the treatment is administered. In addition, a context C is associated with a distribution of characteristics X, prob$(X|C)$. For instance, if the context corresponds to macroeconomic circumstances while X is an individual's employment status, then the share of employed individuals will naturally vary with the macroeconomic cycle.

Data $D[C_1, \cdots, C_N]$ collects treatment outcomes, covariates and treatment status for participants in contexts C_1, ..., C_N: $D[C_1, \cdots, C_N] = \cup_{k=1}^{N}\{(Y_i(\tau_i), X_i, \tau_i, C_k), i \in I_k\}$.

[1] Note that Y_i may evaluate outcomes via some utility index, allowing the decision-maker to express distributional preferences. When the objectives of decision-makers are not known, the impact of treatment on the distribution of outcomes is needed for decision-making.

Note that we abuse notation, and denote by $i \in D[C_1, \cdots, C_N]$ the event that outcomes from person i are included in data $D[C_1, \cdots, C_N]$. In principle, this may be a random event if there is selection into the data (e.g. studies that find inconclusive treatment effects are not reported).

A prior μ over treatment effects across contexts is a probability distribution over context and covariate-dependent potential outcome distributions, $(F(\cdot|X, \tau, C))_{X \in \mathcal{X}, C \in \mathcal{C}}$. Prior μ is needed for a Bayesian decision-maker to extrapolate treatment effect measures across contexts.

Internal validity. Take as given a context C, and related data $D[C]$ collected under context C. The expected treatment effect under context C takes the form $\mathbb{E}[\Delta Y|C] = \sum_{X \in \mathcal{X}} \text{prob}(X|C) \times \mathbb{E}[\Delta Y|X, C]$.

Assume that data D is representative of context C, $\text{prob}(X|D) = \text{prob}(X|C)$, and that treatment is uniformly assigned, $\tau_i \sim U\{0, 1\}$. Data D lets us form an estimator of $\mathbb{E}[\Delta Y|C]$,

$$\widehat{\Delta} Y_C \equiv \frac{2}{|D[C]|} \left(\sum_{i \in D[C], \tau_i = 1} Y_i(\tau_i) - \sum_{i \in D[C], \tau_i = 0} Y_i(\tau_i) \right).$$

Definition 1 (internal validity)**.** *We say that estimate $\widehat{\Delta} Y_C$ is internally valid if $\widehat{\Delta} Y_C$ approaches $\mathbb{E}[\Delta Y|C]$ with probability 1 as sample size $|D[C]|$ gets arbitrarily large.*

Although properly administered randomized controlled trials ensure that estimates are internally valid, the crisis of replicability in the social sciences (Simmons et al., 2011) shows that it is not in fact guaranteed. This is the result of several biases: publication bias in favor of exciting results, difficulties in publishing replications and contradicting studies, difficulties in accessing contradictory information. Well known steps, detailed in Section 3 can help ensure that available data is not biased, so that treatment effect estimates are internally valid.

Figure 1: The challenge of external validity: extrapolation to new contexts

External validity. The decision-maker faces an external validity challenge when they have access to data $D[C_1, \cdots, C_N]$ from contexts $\{C_1, \cdots, C_N\}$, and must make a treatment choice applying to a different context $C_{N+1} \notin \{C_1, \cdots, C_N\}$.

An expert model m is a mapping from data $D[C_1, \cdots, C_N]$ and a novel context C_{N+1} to a treatment effect estimate $m(D[C_1, \cdots, C_N], C_{N+1}) \in \mathbb{R}$. Informally, we say that m is externally valid if

$$m(D[C_1, \cdots, C_N], C_{N+1}) \simeq \mathbb{E}[\Delta Y | C_{N+1}]$$

with high enough probability.

As Banerjee et al. (2017b) observe, without restrictions on prior μ, data from previous studies do not inform extrapolation to new contexts. However, as we argue in Sections 4 and 5, this doesn't mean that we cannot take steps to improve the external validity of findings.

3 Ensuring Internal Validity

Data $D[C]$ can fail to provide internally valid estimates whenever the event that observation i with treatment τ_i is included in $D[C]$, $i \in D[C]$, is correlated with outcome realization Y_i.

This allows for selective reporting of data corresponding to a variety of biases: outright manipulation by the researcher, subsample selection, bias in publishing, or bias in the search for relevant data.

To address this issue, helpful research practices have been proposed, supported by significant organizational efforts:

- The pre-registration of experiments and endpoints; this allows to search for data independently of realized outcomes and prevents the ex post selection of endpoints and subgroups of interest. Several organizations support the pre-registration of experiments, including the National Institute of Health's ClinicalTrials.gov platform, the Center for Open Sciences' Open Science Framework, the American Economic Association's Randomized Controlled Trials Registry, and the Wharton Credibility Lab's AsPredicted platform.

- The timely publication of experimental findings based on a pre-determined schedule; Without such a requirement, pre-registration has no disciplining effect on research entities capable of running many trials and selectively report their results expost. The Food and Drug Administration Amendments Act of 2007 requires parties conducting medical research registered on ClinicalTrials.gov to report study findings within a year of the study end. The website AllTrials.net, and the associated Unreported Trial of the Week column in the British Medical Journal, seek to help enforce reporting by attracting public attention on particularly egregious reporting failures.

  To our knowledge, no such effort seems to have taken place in the social sciences.

- Making data and study design (including survey instruments) systematically available,

allowing to check that findings are not sensitive to the details of the statistical analysis; The Center for Open Science, supported by a number of leading journals, encourages the use of badges publicly indicating compliance with desirable research practices, including the sharing of data, and study materials.

We assume throughout the remaining sections that data made available from past studies is not selected based on outcomes. However, this is not meant to underestimate the magnitude of the challenge presented by the replicability crisis. We are unlikely to address external validity if internal validity remains in doubt.

4 Static Extrapolation

Much of the existing work on external validity (Dehejia, 2005, Hotz et al., 2005) takes a fairly static view: past studies have collected data, and a policy-maker needs to make a once-and-for-all decision in a novel context. Hotz et al. (2005) show that under strong assumptions, it is possible to consistently extrapolate in new contexts. At the same time, static extrapolation need not be successful when available sample size is small and the relevant covariates are high dimensional.[2] Still, we identify concrete steps that researchers can take to give consistent static extrapolation the highest chance of success.

4.1 Consistent extrapolation

Hotz et al. (2005) show that it is possible to consistently estimate treatment effects in new contexts under the strong assumption that conditional on observable covariates, treatment effects are independent of context:

$$\mathbb{E}\left[\Delta Y | X, C_0\right] = \mathbb{E}\left[\Delta Y | X, C_1\right]. \tag{2}$$

[2]Nonetheless, Vivalt (2020) argues empirically that controlling for study characteristics can significantly improve the predictability of treatment effects.

When (2) holds, then as long as the set of covariates in C_0 is rich enough to cover the set of covariates observed in C_1, estimates from C_0 can be consistently extrapolated to C_1 by reweighting observations to reflect differences in the distribution of covariates in context C_0 versus C_1 (Figure 2). The reweighted treatment effect estimator takes the form:

$$\mathbb{E}[\Delta Y|C_1] = \sum_{X \in \mathcal{X}} \text{prob}(X|C_1) \times \mathbb{E}[\Delta Y|X, C_0]. \tag{3}$$

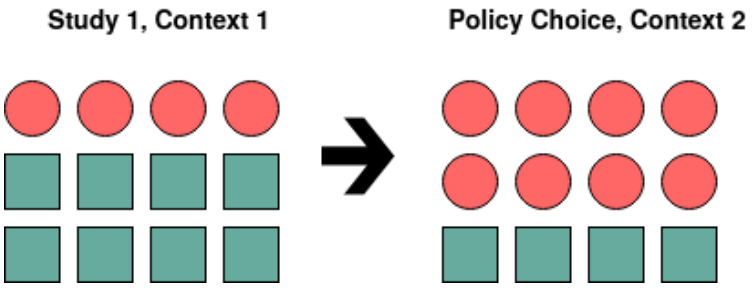

Figure 2: Extrapolating to new contexts by reweighting observations

In practice, naïve reweighting need not be practical if covariates X take many potential values or are high dimensional. In that context, work along the lines of Wager and Athey (2018) and Athey et al. (2019b), can play an important role. By adapting machine learning methods (i.e. random forests, Breiman (2001)) that have proven successful in dealing with high dimensional data and non-linear relationships, they allow analysts to better estimate conditional treatment effects, and therefore – provided (2) holds – extrapolate treatment effects to new contexts.

4.2 Useful steps

Consistent extrapolation by reweighting conditional treatment effects depends on two requirements: (i) condition (2) should hold; (ii) the support of covariates X in context C_0 should include the support of covariates X in context C_1. We now discuss steps that can be

taken to give consistent extrapolation the highest chance of succeeding.

4.2.1 Collect rich covariates

Condition

$$\mathbb{E}\left[\Delta Y | X, C_0\right] = \mathbb{E}\left[\Delta Y | X, C_1\right]$$

only holds if the covariates we are conditioning on are rich enough. If we omit covariates X' that mediate treatment effects, then context will be correlated to treatment effects via the distribution of omitted covariate X'.

Consider for instance, evaluating the effect of a drug whose treatment effect depends on the gender, age, and blood type of the recipient, and that the population shares of different blood types vary across contexts C_0 and C_1. Then, we have that

$$\begin{aligned} & \mathbb{E}[\Delta Y | \textsf{age, gender}, C_0] \neq \mathbb{E}[\Delta Y | \textsf{age, gender}, C_1] \\ \text{but} \quad & \mathbb{E}[\Delta Y | \textsf{age, gender, blood type}, C_0] = \mathbb{E}[\Delta Y | \textsf{age, gender, blood type}, C_1] \\ & = \mathbb{E}[\Delta Y | \textsf{age, gender, blood type}] \end{aligned}$$

This suggests that to give consistent extrapolation the best chance of success, it may be beneficial to measure rich covariates likely to mediate treatment effects. Importantly, relevant covariates may include data that is private information to the treatment recipients. This includes:

- *Network Structure.* Banerjee et al. (2014) show that in practice, members in a social network can identify centrally located members of the network, a quantity that may otherwise be difficult to measure but is necessary for generalizing experimental results for which network structure matters.

- *Participant preferences.* Karlan and Zinman (2009) emphasize that well designed experiments can elicit unobservable private information from participants. In the context

of a lending program, they show it is possible to separately identify the role of both adverse selection and moral hazard in determining treatment outcomes.

Chassang et al. (2012) study experiment design as a mechanism design problem and show that maximally informative experiments elicit treatment effects as a function of preferences, and can be attained by letting participants make choices between different lotteries over consumption bundles including treatment status. This allows for the recovery of Marginal Treatment Effects (Heckman and Vytlacil, 2005), which can be used to simulate treatment effects under various treatment pricing mechanisms.[3]

Ashraf et al. (2010), Cohen and Dupas (2010) and Jack (2013) study treatments in which participants receive subsidized goods, respectively a water purification solution, insecticide treated bed nets, and a tree planting contract. They recover treatment effects conditional on values, and evaluate the effectiveness of prices (set via take-it-or-leave-it offers, as well as auctions) in effectively targeting subsidies.

More recently, Narita (2021) highlights that by eliciting values, it is possible to run more ethical trials that improve participants welfare at a limited cost in information.

4.2.2 Include diverse contexts

A second requirement for reweighted estimator (3) to be implemented is that covariates X that occur with positive probability in context C_1 should all have positive probability in context C_0. Otherwise $\mathbb{E}[\Delta Y|X]$ cannot be estimated using data from context C_0 for each covariate X in the support of context C_1. Although reweighting means that studies don't have to be representative to have some external validity, they need to cover the range of

[3] A related literature seeks to evaluate the extent to which one can extrapolate LATE estimates Angrist et al. (1996), which correspond to treatment effects on the subset of participants whose treatment adoption decision is affected by the experiment ("compliers"), to a representative population treatment effect. Kowalski (2018) demonstrates a way to use outcomes for "always takers" and "never takers" along with a monotonicity assumption to obtain bounds on population wide treatment effects. Bertanha and Imbens (2019) proposes tests of the assumption that treatment effects are independent of participants' compliance decision.

possible relevant covariates. For this reason, studies executed in *contexts that are diverse*, i.e. that include recipients exhibiting a larger support of covariates, will help generate more externally valid predictions.

One way to achieve this is to broaden the set of individuals eligible to receive treatment within the initial context C_0. This may have a cost in terms of achieving internally valid estimates. In general, sampling noise and integer constraints mean that treatment and control groups need not be strictly balanced. This does not matter much when sampling from a treated population that exhibits very homogeneous characteristics. If instead a study samples a broader range of covariates, then balance issues may be more concerning. A number of techniques can then be used to ensure balanced assignment without sacrificing robust inference (Morgan and Rubin, 2012, Banerjee et al., 2020).

Another possibility is to expand the study sample to include more than one context. This is especially useful if no single context exhibits sufficient diversity in covariates X to cover the entire range of possible covariates $\mathcal{X}$. Figure 3 provides an illustration. The range of possible covariates $\mathcal{X}$ is pictured by the square to the left. Initial context C_0 only includes covariates in the lower left triangle, and alone cannot be used to extrapolate treatment effects to contexts C_1, C_2 or C_3. However an experiment that suitably diversifies the range of covariates being sampled – here a study including both contexts C_0 and C_2 – would permit extrapolation to C_1, and C_3.

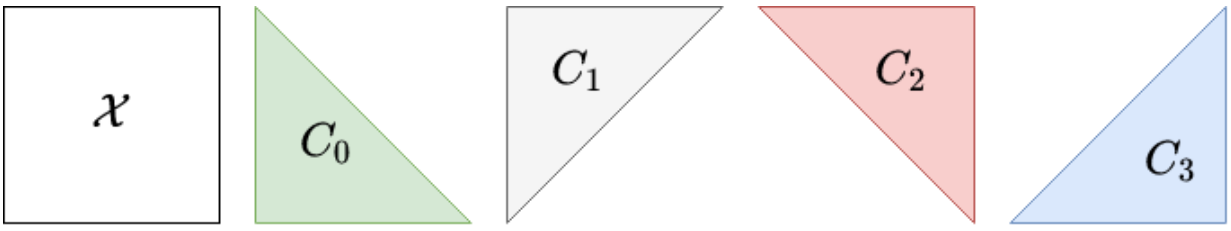

Figure 3: Varying contexts helps span new environments

Two objectives of variation in context would be to improve

- *Extrapolation across time, and macro-economic circumstances.* Rosenzweig and Udry (2020) demonstrate the tenuous generalizability of cross-sectional estimates at one point in time to a future date in the context of agricultural investments, making a case for the collection of aggregate variables as well as multi-period experimental design for evaluating the temporal external validity of estimates.

  Dehejia et al. (2019) use data from a variety of different contexts and time periods to replicate the empirical approach in Angrist and Evans (1998). They find that the macro-level environment (e.g. GDP, women's labor force participation rate) is critically important to explain variation in estimates in different contexts. This implies that external validity of the point estimates requires taking the macro-level environment into account.

- *Extrapolation across cultures and norms.* An active and growing literature documents the importance of culture for economic outcomes. This has implications for external validity, because unmeasured cultural variables may be important when extrapolating results across environments. Padró i Miquel et al. (2015) show how culture can impact the effectiveness of democratic institutions, demonstrating that the presence of temples, interpreted as a proxy for civic capital, enhances the power of democracy to deliver public goods.

  Corno et al. (2020) study how agricultural output affects the risk of child marriage. They show that cultural norms affect the response of child marriage to agricultural shocks, and stress the importance of culture in evaluating the external validity of measured relationships between variables that are important for the design of policy responses to child marriage. The paper finds that droughts have different effects on the timing of marriage in Sub-Saharan Africa and in India. In Sub-Saharan Africa, the marriage payment is a bride price, while in India it is a dowry, and this interacts with droughts to affect the timing of marriage in opposite ways: in Sub-Saharan Africa,

droughts increase the marriage rate, while in India they reduce it.

Jayachandran and Pande (2017) shows that eldest son preference in India is a driving mechanism behind stunting in later-born children, and is strongest in regions that practice patrilineality. The decision of how to allocate resources among children is therefore partially determined by culture, and so exporting policies aimed at increasing child health in one location to another must account for parental preference over resource allocation to their children.

Another important variation in context is scale (Al-Ubaydli et al., 2017, Muralidharan and Niehaus, 2017). As we argue in Section 5 there is an option value in only scaling up the treatments that seem to generate sufficient returns. For this reason, we think the question of external validity at scale can be tackled by dynamically expanding a program in stages, building on previous findings to decide whether and how to expand at each stages. Banerjee et al. (2017a) provides a concrete description of such a process in the context of scaling-up a proof-of-concept study evaluating a novel educational strategy in Indian elementary schools.

4.2.3 Testing for external validity across local contexts

A difficulty with condition (2) is that it cannot be tested if data from a single context is used. This may cause concern for a decision-maker considering whether, and how to extrapolate to their own context.

Reassuring evidence can be provided by testing the external validity of findings within a given study. Indeed, many studies naturally include some local variation in contexts, i.e. variation in location, groups, or time. Imagine that context C_0 can be decomposed into two distinct contexts:

$$C_0 = C_{0,A} \cup C_{0,B}.$$

If consistent extrapolation holds, then we expect that

$$\mathbb{E}[\Delta Y|X, C_{0,A}] = \mathbb{E}[\Delta Y|X, C_{0,B}].$$

This equality can then be tested using only data from C_0.

5 Dynamic Extrapolation

In general, there is no reason to expect that condition (2), allowing for consistent extrapolation, should hold. This could be because relevant covariates are unknown, or cannot be reliably measured. Treatment effects could also just be quite sensitive to context. In recent work, Gechter et al. (2018) point to the difficulty of extrapolating the treatment effects of conditional cash transfers even using more sophisticated models of conditional treatment effects. This echoes LaLonde (1986)'s argument that the econometric analysis of observational data fails to replicate experimental results: misspecification and data-limitations just cannot be ignored.

Although consistent extrapolation from a single study is an implausible goal, this does not mean that there is no useful way to approach extrapolation. Instead it means that (i) extrapolation from a single study is necessarily subjective, and (ii) research and the evidence it produces should be viewed as a dynamic learning process refining our understanding of what might work, and in what context, rather than as one-shot evidence collection opening and closing the book on one particular question. This dynamic view of evidence suggests ways to approach external validity even when condition (2) does not hold, as well as steps researchers can take to improve learning across studies.

5.1 Framework

Section 4 couches the question of external validity in a static once-and-for-all manner: a corpus of experimental data has been accumulated; an expert must make a decision that will be maintained for the foreseeable future.

In practice, learning does not stop at one point in time. Instead, each implementation in a new context provides a new opportunity for learning. Additional data can be confronted to anticipated outcomes based on previously available evidence.

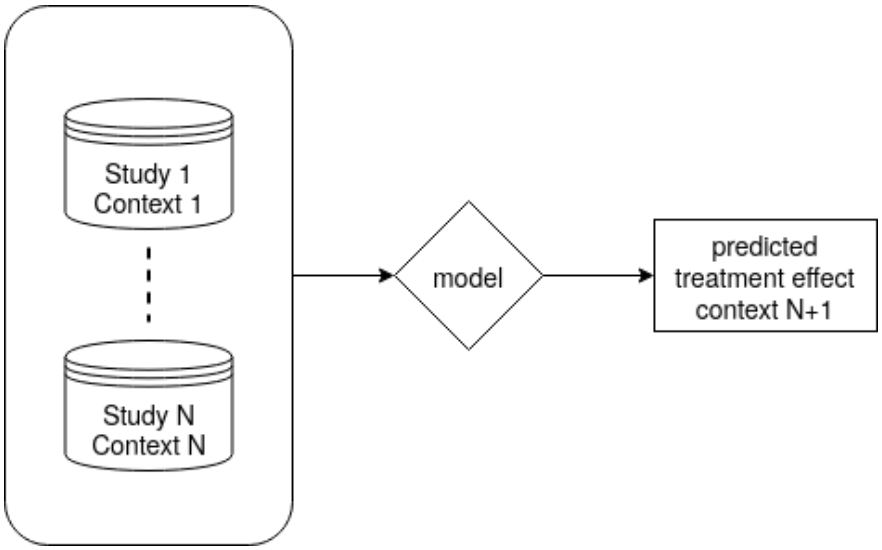

Figure 4: A predictive expert model, applied at one of many stages of evidence collection.

We refer to any procedure m, used to makes predictions about treatment effects in a new context C_{N+1} on the basis of data $D[C_0, \cdots, C_N]$ collected over previous studies, as an expert model. Model predictions $m(D[C_0, \cdots, C_N], C_{N+1})$ can potentially guide further policy making, as well as continued research. Because a model m can be tested many times along the process of research, there exist adaptive learning procedures that will ensure non-trivial learning without having to make a prior assumption about the data-generating process. For instance, it will be possible to learn which of two models decision-makers should use when choosing policies on the basis of extrapolated treatment effects.

What's a model. We emphasize that models here can be any process by which information from past studies is associated to predictions about treatment effects in new environments. This includes both purely statistical procedure based on objective data, and structural models based on a more explicit theory of change. Alternatively, the model may take the form of a human expert with considerable field experience, capable of formulating predictions about treatment effects in different settings. DellaVigna and Pope (2018) show that experts (and especially graduate students) are quite good at predicting the effectiveness of various incentive schemes. Finally decision-making committees combining the views of multiple experts, as well as statistical analysis, may be treated as a "model", since it effectively maps data to predictions about outcomes. This is roughly what meta-analyses performed by policy institutes such as Cochrane Reviews accomplish.

In this dynamic setting the goal of externally valid inference is to identify the best extrapolation model among available competing options.

5.2 Aggregating predictions

A key reason why dynamically extrapolating across many context is an intrinsically more hopeful exercise than extrapolating from a single study is that there are good decision-making algorithms – known in the statistics literature as online learning algorithms (Blackwell, 1956, Hannan, 1957, Foster and Vohra, 1999) – which provide prior-free performance guarantees when the number of implementation opportunities grows large, even if contexts are related in arbitrary ways.

Concretely, imagine that we are trying to learn which of two models, model A and model B, is best at producing policy relevant treatment estimates. Formally, let us denote by $\widehat{\Delta}Y_{A,k}$ and $\widehat{\Delta}Y_{B,k}$ the treatment effect estimates respectively produced by model A and model B in context C_k using data $D[C_0, \cdots, C_{k-1}]$. Let us denote by ΔY_k^* the true treatment effect in context k. We assume that both estimated and true treatment effects are bounded in

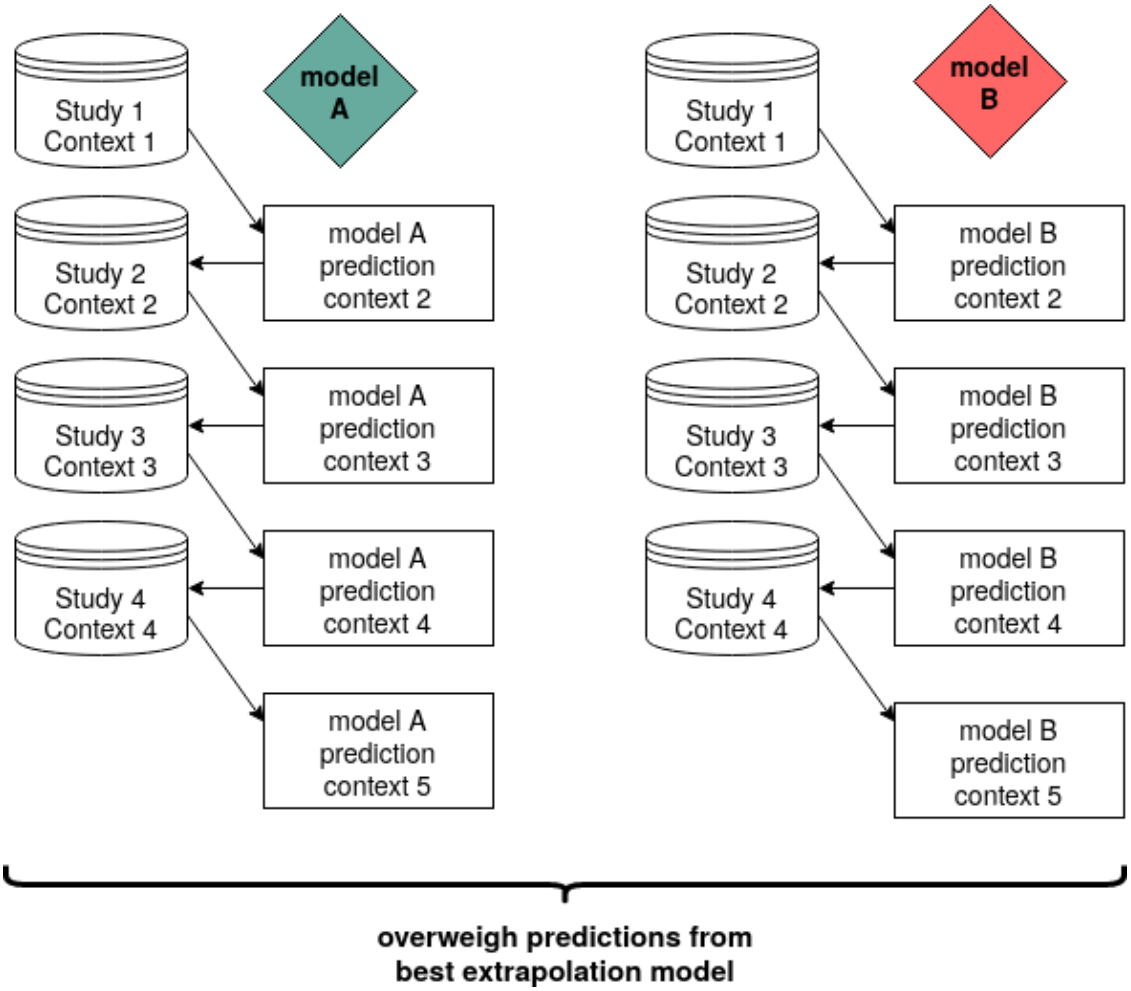

Figure 5: Evaluating and choosing expert models

absolute value by $M > 0$.

An aggregation strategy σ associates past data to weights $\sigma_{A,k} \in [0,1]$ and $\sigma_{B,k} \in [0,1]$ (such that $\sigma_{A,k} + \sigma_{B,k} = 1$) used to form estimate

$$\widehat{\Delta}Y_{\sigma,k} \equiv \sigma_{A,k}\widehat{\Delta}Y_{A,k} + \sigma_{B,k}\widehat{\Delta}Y_{B,k}.$$

For illustrative purposes let us assume that policy decisions made using an estimate $\widehat{\Delta}Y_k$ generate a payoff $u_k(\widehat{\Delta}Y_k) \equiv -(\Delta Y_k^* - \widehat{\Delta}Y_k)^2$ that is maximized by making exact predictions.

Let us denote by

$$\mathcal{R}_T^A = \frac{1}{T+1}\sum_{k=0}^{T} u_k(\widehat{\Delta}Y_{A,k}) - u_k(\widehat{\Delta}Y_{\sigma,k})$$
$$\mathcal{R}_T^B = \frac{1}{T+1}\sum_{k=0}^{T} u_k(\widehat{\Delta}Y_{B,k}) - u_k(\widehat{\Delta}Y_{\sigma,k})$$

the performance loss – also known as regret – of σ relative to model A and relative to model B. The literature on online learning establishes that the following is true (Blackwell, 1956, Hannan, 1957, Foster and Vohra, 1999).

Proposition 1. *There exist data-driven rules σ such that for all sequences of true treatment effects ΔY_k^*, and model estimates $\widehat{\Delta}Y_k^A$ and $\widehat{\Delta}Y_k^B$, estimates produced by aggregation rule σ are approximately as good as those of the best model* in hindsight*. Formally,*

$$\mathcal{R}_N^A \leq M\frac{1}{\sqrt{N+1}} \quad \textit{and} \quad \mathcal{R}_N^B \leq M\frac{1}{\sqrt{N+1}}.$$

One such data-driven rule is to set

$$\sigma_{A,k} \equiv \frac{[\mathcal{R}_{k-1}^A]^+}{[\mathcal{R}_{k-1}^A]^+ + [\mathcal{R}_{k-1}^B]^+} \quad \text{and} \quad \sigma_{B,k} \equiv \frac{[\mathcal{R}_{k-1}^B]^+}{[\mathcal{R}_{k-1}^A]^+ + [\mathcal{R}_{k-1}^B]^+},$$

where $[\mathcal{R}]^+ \equiv \max\{0, \mathcal{R}\}$ denotes the positive part of $\mathcal{R}$.

We emphasize that Proposition 1 holds regardless of the expert models used to make predictions, whether they use subjective evaluations from human experts or not. The important takeaway is that even if we don't know how different contexts relate to one another, an adaptive use of available evidence allows researchers to construct extrapolations that place most of their weight on the best available models.

5.3 Useful steps to support learning over many contexts

Proposition 1 provides hope that some degree of external validity can be achieved when making decisions over many contexts. Still, there are steps that can be taken to increase the effectiveness of aggregation procedures. As we highlight, these steps are often positive externalities from current researchers to future users of the evidence they create.

5.3.1 Document contexts

One implication from the dynamic view of research is that it is socially valuable for researchers to carefully document the context in which they operate, even if it is fairly homogeneous, and therefore would not lend itself to informative subgroup analysis within a given study.

Incentives in this case are poorly aligned. The study-level benefits for documenting the context are low. If the context (e.g. macroeconomic conditions, norms, implementation partners ...) is common across participants, then its role in determining outcomes cannot be determined within a single study. The benefits of carefully documenting the context in which a study occurs accrue to future researchers and policymakers, who can combine that information with data from other studies conducted in different contexts to improve the external validity of their models.

5.3.2 Standardize outcome and covariate measures

Practitioners of meta-analyses routinely bemoan the fact that different but related studies differ in the way they measure covariates and outcomes. In a recent meta-analysis of the impact of micro-credit, variation in measurement leads Meager (2016) to ignore effects on income and assets. Measuring covariates and outcomes differently across studies hampers our ability to aggregate findings effectively.

Consider for instance a decision-maker interested in extrapolating the treatment effect of

a loan product to a new context of interest. Two studies have been published on the issue, evaluating the extent to which the effect of the loan product is mediated by underlying levels of trust in the population. If the two studies measure trust in the same way and find different treatment effects even conditional on trust, then the decision-maker may legitimately infer that there is significant uncertainty in treatment effects even conditional on underlying trust. If instead the two studies measure trust in different ways, then differences in treatment effects conditional on *measured* trust don't necessarily indicate that treatment effects conditional on trust are actually different. Differences in the relationship between the effect of loan products and trust may be driven by differences in measurement. In other words, variation in measurement may invalidate consistent extrapolation along the lines discussed in Section 4, even when it is possible.

It is useful to note that there may be real hurdles to standardization. First, there may be a lack of consensus among researchers. In that case, organizational resources would have to be spent to formulate measurement methods that everybody can live by. Second, it is often convenient to reinvent the wheel – some measurement methods can be reasonably intuitive, and individual researchers may have idiosyncratic preferences regarding the best way to measure. Finally, adaptation to the local context could make some designs more attractive than others. For instance, measuring income in rural communities versus urban ones is likely to raise different challenges. How should home production be valued? How about payments in kind and favor exchanges? Currently researchers have no incentive to internalize how their measurement choices will facilitate the aggregation of their evidence with that provided by other studies. As a result, there is a real role for funding organizations to encourage the standardization of measurements. This does not mean that measurement adapted to local conditions must be discouraged, but rather that, at least as far as key endpoints and key participant and context characteristics are concerned, it is desirable to include at least some standardized measures in the set of measures used.

Because standardization is fundamentally an attempt to resolve an externality, fund-

ing organizations have a role to play here. Specifically, they can (i) support open-source databases of reference survey modules that researchers can include relatively cheaply in their studies; (ii) ensure that funding recipients internalize the value of standardizing their survey instruments; (iii) require that funding recipients provide open source access to their survey instruments.

We note that EGAP's Metaketa Initiative offers one possible model to start building such standardized measures by coordinating research teams around pre-identified important themes. In addition, EGAP enforces several important research practices, including pre-registration, formalizing hypotheses, and pre-analysis plans.

5.3.3 Engage in structured speculation

In the course of conducting a field study, experimenters often form a fairly sophisticated subjective understanding of the mechanisms that mediate treatment effects in their context. In a business setting, this sophisticated experience-based understanding of the environment would be perceived as a valuable asset. In a research context however, this sophisticated understanding cannot be substantiated to the usual standards of evidence, and rarely finds its way in publications. As a result, the value of such experiential knowledge is lost.

Consider for instance, an experimenter evaluating the impact of introducing weather insurance in a rural setting. During field-work, the researcher may observe that local farmers have unusually high trust in their government for historical reasons. Since farmers share this common experience, the experimenter is unable to provide data-driven evidence that the reputational capital of public institutions is an important driver for the adoption of weather insurance products. As a result, this subjective belief does not get reported in published research.

Banerjee et al. (2017b) argue that once we take a dynamic view of research, subjective assessments need no longer be cheap talk, provided they are expressed in clear and falsifiable ways (i.e. the study design needed to test them and the relevant test statistic should

be clear). Indeed, later studies provide a natural opportunity to validate or not subjective assessment. This incentivizes researchers to be both truthful and realistic about their subjective assessments. Banerjee et al. (2017b) are agnostic about the procedure used to speculate beyond the study environment: they may be obtained through a formal structural analysis, through reflection and intuition guided by experience, or a mix of both.

However, the manner in which subjective assessments are formulated is important. Banerjee et al. (2017b) propose that researchers include a "Structured Speculation" section in their papers or research reports, satisfying the following two requirements: the speculative nature of the exercise should be transparent; predictions about outcomes from treatment in other environments (or outcomes of other treatments as in counterfactual analysis) should be formulated in a clear, unambiguous, and falsifiable manner – it should essentially suggest an unambiguous blueprint on how to test the claim.

Several dimensions of external validity seem worth speculating about. First, under what environments should we expect treatment to be effective? Second, under what environments should we expect treatment to be ineffective? Third, what are important aspects of treatment about which we have significant residual uncertainty? Fourth, what are plausible improvements to the treatment policy?

Formally embracing speculation offers a number of potential benefits:

1. It reveals practical know-how useful to policy-makers that would not typically find its place in usual research publication.

2. Along the lines suggested by Figure 6, subjective assessments can be used as an additional input for expert models, contributing domain-specific, forward-looking insight. Furthermore, this additional prediction data may be used to evaluate which experts, and which extrapolation techniques seem to be most successful.

   This complements a view formulated by DellaVigna et al. (2019): they suggest eliciting experts' beliefs prior to experimental evaluation to identify successful forecasters.

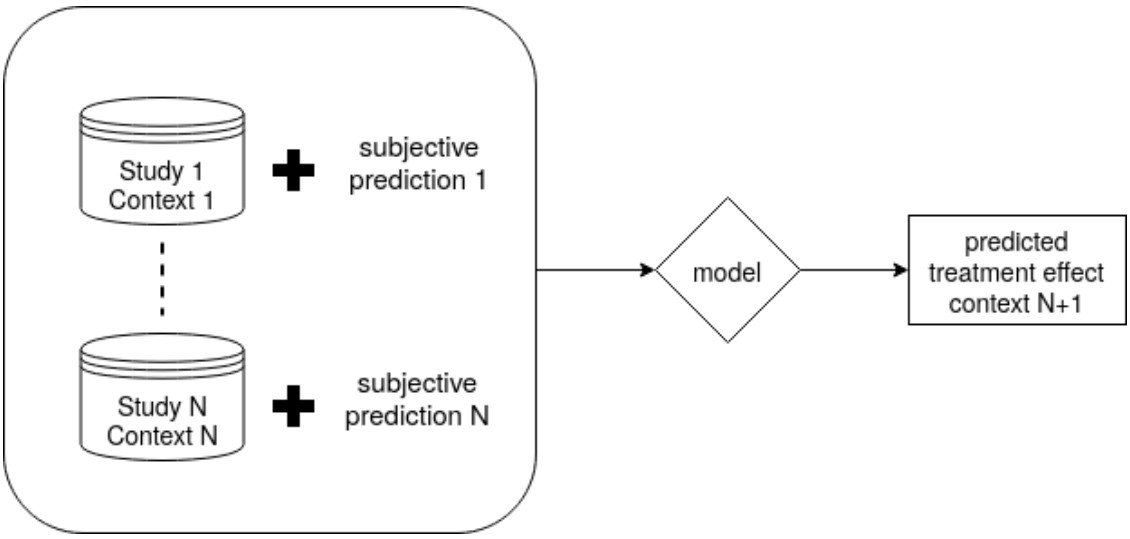

Figure 6: Using structured speculation to improve expert models

Whether the prediction exercise occurs at the end of the last study, or before the next one is second-order. We share their view that subjective assessments are useful inputs, and that it is beneficial to elicit such views from a larger group.

3. Finally, structured speculation provides guidance for later researchers, helping elaborate models and theories of change, as well as pointing out dimensions of residual uncertainty where further investigation is most valuable. In the event that findings do not replicate well, clearly instantiated predictions from an explicit study design also weaken the argument that subsequent studies were somehow poorly conducted.

Dupas (2014) provides an early example of structured speculation by discussing the types of health products and settings in which her findings on the impact of free access on long-term adoption are likely to extend, based on an explicit theory of change (Weiss et al., 1995, Weiss, 1997). More recently some papers have formally engaged in structured speculation. This includes Rosenzweig and Udry (2020), Björkman Nyqvist et al. (2018) and Fink et al. (2020), who explicitly reference and engage in structured speculation.[4] Fink et al. (2020), for instance, study the effects of relaxing seasonal liquidity constraints, by experimentally

[4]Burchardi et al. (2019) describes an approach to extrapolating their results across contexts.

introducing loans due only after the harvest season in a large number of rural villages in Zambia, on the allocation of labor and agricultural output. After describing their experiment and results, the authors use a calibrated model to make predictions about quantities that were not experimentally measured.

5.3.4 Link studies

In principle, the fact that past findings will be revisited in future research enhances the validity of findings over time. However, this is only true if the decision maker is able to get a complete view of the set of studies related to a particular question. In practice this is often difficult. One reason is that later studies may be less well published than the original ones, or not published at all. In the extreme, there may be bias against publishing contradictory findings, assigning discrepancies to measurement issues. Alternatively, if the decision-maker uses citation count as a way to prioritize their investigation of the literature, then it is natural that early seminal papers will figure prominently. In contrast, relevant later work may end up being ignored by evidence reviews.

Recent community-driven attempts to link studies can facilitate the search and discovery of comprehensive information. The website PubPeer shows how one might go about building such a resource. Individual papers are given a webpage, on which the research community can post information as well as link other studies. The Center for Open Science offers a set of software tools, the Open Science Framework (OSF), whose mission is to

> effectively share the story of your research project and eliminate data silos and information gaps. The OSF allows all of those tools to work together, removing barriers to collaboration and knowledge.

Such a resource may be organized to facilitate information aggregation by humans or machines. Steps may include posting properly formatted data, or at least results tables. Funders could help build such a resource incrementally by: (1) asking researcher to specify

in advance which previous studies their data acquisition hopes to refine; (2) automatically linking their research findings to these previous studies.

5.4 Useful steps to support adaptive policy-making

In practice, policy-makers do not blindly adopt at scale a policy that has only been implemented in other environments. Instead, many policy-makers will choose to implement the policy on a temporary basis, i.e. run a pilot, and base their adoption decisions on results from the pilot.[5] This suggests two insights: the first is that instead of extrapolating from studies performed in significantly different contexts, the relevant problem may in fact be to extrapolate from previous evidence from different contexts *and* a pilot study in the relevant context.[6] This seems like a much more hopeful exercise. The second insight is that since policy-makers will only implement treatment at scale when it seems effective, treatment should not be valued for its average effect, but instead for its option value. Treatment effect estimates that do not take into account this option value underestimate the dynamic value that policy-makers actually get from experimenting with treatments.

5.4.1 Extrapolating from context-relevant pilots and surrogate endpoints

The challenge of external validity is that new contexts may just be very different from the contexts of past studies. In many settings however, the policy-maker is not compelled to implement treatment at scale on the first try. Instead, a policy-maker will frequently run pilot experiments in their context, and decide whether to scale up on the basis of both previous studies, and findings from their own pilot. The question therefore becomes whether it is feasible to use short-term and small scale, but *context relevant*, outcome data to predict

[5] Consistent with this view, Vivalt et al. (2019) argues that evidence from their own country receives higher weight in policy-makers' decision process.

[6] We note that if the decision-maker is averse to randomization, the pilot need not take the form of a randomized controlled trial. In that case, treatment outcomes should be compared to a suitable prior about outcomes under the control policy.

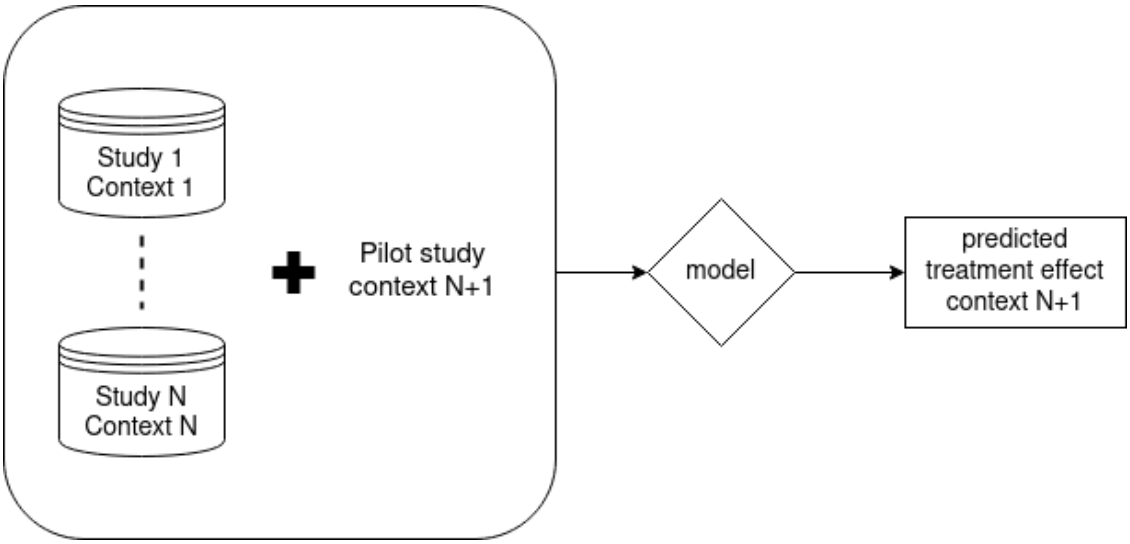

Figure 7: Using short-run pilot data to improve expert models

the long-run treatment effects of a policy implemented at scale. While this is a non trivial challenge, the fact that pilot data reflects specificities of the new context increases the chances for successful extrapolation.

A key difficulty is that there may be significant delay between the implementation of a treatment and the realization of the welfare relevant outcome. This poses a significant practical challenge to adaptation. In that setting, it is socially valuable for researchers to help develop surrogate short-term measures predictive of long term outcomes. Athey et al. (2019a) shows how to do so in the context of predicting long-term effects of the GAIN job assistance program. They provide conditions under which surrogate measures are valid endpoints (outcomes must be independent of treatment conditional on surrogate endpoints), and highlight the value of combining multiple surrogate measurements in an aggregate surrogate index. Importantly, the surrogate index is built using observational long-term panel data, and does not necessarily require observing long-term experimental outcomes. As a result surrogate indices can be pre-registered before an experiment is run.

Additionally, observing medium-term experimental outcomes can be used to test the assumption that surrogate endpoints are sufficient statistics for long-term outcomes of interest.

Systematically developing surrogate endpoints, and acquiring the medium term data needed to validate them, would considerably enhance policy-makers' capacity to extrapolate from context relevant pilots, as well as other researchers' ability to evaluate the long-term effects of the treatments they investigate.

5.4.2 Measuring the option value of treatment

The fact that policy-making is adaptive also changes the way we value treatment effects. Consider a policy-maker that implements a policy on a temporary basis, evaluates it after some given amount of time, and then decides whether to continue the treatment or revert to a default option.[7]

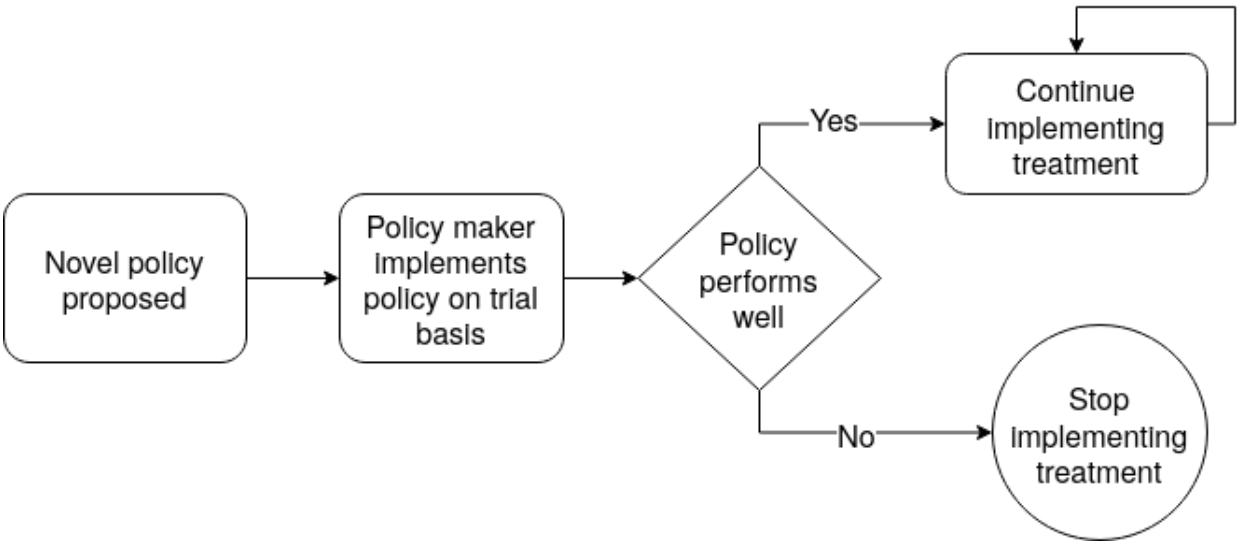

Figure 8: Treatments evaluation should reflect the process of adoption

Dynamic adoption means that we should value treatments as real options. This can significantly change our assessment of whether a treatment is beneficial or not. Imagine an educational treatment that uses a different teaching approach to motivate underperforming children. Half of the children are responsive to this different approach in which case their

[7]Option value considerations only apply when treatment is repeated, although it need not be repeated on the same individuals.

performance improves. Half of the children are not responsive to this new approach, and their performance decreases (the new program crowds out more productive educational content). On average the effect on a randomly chosen child is zero.

Now consider program adoption. Enrollment in the program lasts for a year. At the end of the year, students choose to either continue or enroll in other programs. Naturally, those who experience better outcomes than expected will continue the treatment. Those who experience worse outcomes than expected exit. Hence, even though the program has zero effect in the first year, it has a positive treatment effect over the long run once adaptation is taken into account. Note that the dynamic adoption strategy in which students continue the program only if their experienced returns are high enough only generates a positive option value if (i) early treatment effects are positively correlated to later treatment effects; and (ii) there is no large negative long run effect of following treatment for a year, and then returning to control.[8]

For this reason, it is valuable for experiments to not just evaluate the mean value of a treatment, but rather evaluate the option value of treatments, building in selective adoption on the basis of early treatment outcomes. Chassang et al. (2021) delineate how to do so in both experimental settings and using observational data. Option value considerations may be important whether adoption decisions are made by the experimental subjects themselves, or by policy-makers if the policy is a systemic one that cannot necessarily be targeted at the individual level because of practical concerns, or important externalities. One example is the adoption of systematic health campaigns (including vaccination, deworming ...).

One difficulty with evaluating the option value of treatments is to take into account potential negative impacts of first adopting, and then abandoning treatment. In the medical field, many drugs (e.g. antidepressants) have severe withdrawal effects. In an economic setting, Dupas (2014) and Fischer et al. (2019) evaluate the extent to which offering subsi-

[8]For instance, one may be concerned that students that experience very bad outcomes under treatment are hindered in the future by holes in their past education, or suffer from a loss of confidence.

dized health products can diminish adoption compared to control groups once the subsidy is removed, presumably by resetting consumers' expectations about value and what is a fair price.

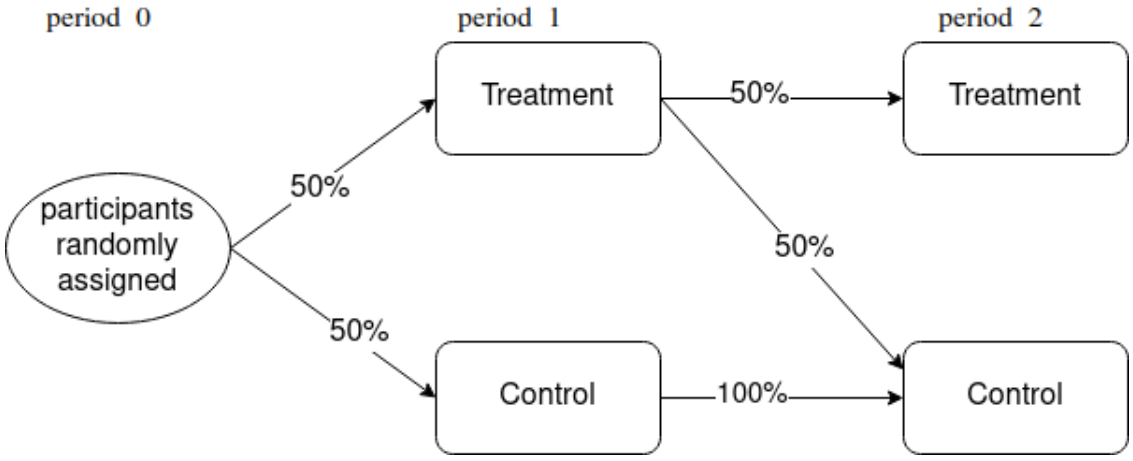

Figure 9: To evaluate the option value of treatment, it is useful to reassign some participants away from treatment

To measure potential withdrawal effects it is necessary to modify standard randomized controlled trial designs in which participants are assigned to treatment and control once. Instead, Chassang et al. (2021) consider dynamic designs in which a share of participants initially assigned to treatment are reassigned to control (9) after one period has passed ($t = 1$) and *interim outcomes have been measured*. Denote by Y_t outcomes in periods $t \in \{1, 2\}$. If participants are randomly reallocated to the control group, then such a trial lets us measure the treatment effect of dynamic rules of the form

$$\tau_1 = 1 \quad \text{and} \quad \tau_2 = \begin{cases} 1 & \text{if } Y_1 > y \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

for all thresholds y. Realistically, it may not be feasible or ethical to randomly assign all participants once information about their individual returns is available. In this case, it may only be possible to randomize period 2 treatment for those treated in period 1 if their

outcome Y_1 in period 1 falls in a specific range: $Y_1 \in [\underline{y}, \overline{y}]$. In that case, the design of Figure 9 can be used to evaluate the option value of adoption strategies (4) for any threshold $y \in [\underline{y}, \overline{y}]$.

By designing experiments to evaluate the option value of treatments, researchers generate useful information indicating to policy-makers whether experimenting with a policy is worth it for them. Note that building on Athey et al. (2019a) we could use dynamic adoption rules based on aggregates of short-term outcomes that are better predictors long-term outcomes. In fact, if experimenting with, and then abandoning treatment has a negative treatment effect, then it will be useful to try and shorten the initial evaluation period as much as possible.

6 Comments

We are grateful to several field experts for providing extensive comments on an early draft. Instead of claiming their thoughts as our own, we found it preferable to collect them in an independent section.

Testing mechanisms and theories of change. In Section 5 we define models as any mean used to extrapolate from past contexts to new ones. This is an intentionally broad definition that encompasses human experts formulating opinions, purely statistical models, expert judgment supported by theories of change, and structural models built on a detailed understanding of the incentives of different economic actors. We choose to remain broad to highlight that as far as dynamic extrapolation is concerned, these are all valid contenders, and that good procedures exist to disproportionately listen to the best available models, regardless of their type.

Multiple field experts have expressed significant hope that models expressing an understanding of mechanisms will lead to successful extrapolation models, and that we should

systematically encourage researchers to formulate their findings in terms of what we learn about mechanisms rather than what we learn about specific treatments. Paraphrasing one expert, mechanisms may be more externally valid than policies, so that an understanding of mechanisms could translate better across contexts. For instance

> "Parents respond to incentives that bring forward in time the benefits of vaccinating kids"

may generalize, while

> "Parents will increase vaccination rates by 5 pp when given lentils at clinic"

may not. This is an important message for funders who sometimes prefer funding learning about policies over learning about mechanisms.

Another field expert suggested several practical ways that can help researchers use the evaluation of a given policy to build a deeper understanding of mechanisms. A first difficulty is that policies correspond to bundles of hypotheses that need to be unbundled to get at mechanisms. One helpful practical way to identify underlying hypotheses is to ask

> Why could this policy fail?
>
> Why could this fail to generalize?

Once possible theories of change are identified, one opportunity to test such theories is whether they help carry over findings from the pilot to the main experiment.[9] Finally, qualitative surveys can provide valuable insights into the mechanisms at work. Dupas and Robinson (2013) use such surveys when evaluating the reasons why simple lock boxes can help increase savings.

[9] In this sense, pilots and main implementations necessarily offer some variation in context. The field expert also suggests that potential theories of change may be pre-registered.

Selecting contexts and policies. Several field experts highlighted that both contexts and treatments exist in fairly high dimensional spaces and that many variants of both context and treatment may be available to experimenters. For instance, the experimenters may have an opinion on which contexts are likely to lead to higher versus lower treatment effects. Alternatively, the experimenter may vary the inputs of treatment, for instance by selecting implementation partners, choosing the amount of training provided to subjects or enumerators.[10]

One expert highlighted that a dynamic view of the research process affects the order in which researchers may decide to evaluate various contexts and treatment variations. Favorable and less favorable environments can provide upper and lower bounds to the benefits of policies. Initial evaluations of novel policies whose mechanics are still poorly understood may focus on favorable environments to provide an opportunity for learning, and establish possibility results. For instance, initially, it may be preferable to implement complex policies in partnership with well trained administrators, the rationale being that total disasters are uninformative. Subsequent evaluations may then focus on less favorable contexts. For instance, the quality of program implementation may be adjusted to be more representative of implementation at scale.

Another expert highlighted the trade-off between homogeneous versus less-homogeneous experimental populations. Homogeneous populations are good for power, and less good for external validity. The opposite holds for heterogeneous populations. Early studies may focus on homogeneous populations while later studies evaluate returns in heterogeneous populations.

Documenting contexts and policies. Multiple experts emphasized to us the importance of documenting implementation details, including the background, training, and incentives

[10]The frontier between context and treatment is sometimes blurry. Selecting a particularly effective, or high social capital implementation partner can be viewed as an aspect of treatment, and it affects the experimentation context.

of the people implementing the intervention. This helps extrapolate across context but also allows decision makers to judge the cost and difficulty of replicating a policy. Evans and Popova (2014) emphasize the fact that implementation costs vary across contexts, including program scale.

Another expert suggested drafting a checklist of concrete context measurements (e.g. GDP growth, enumerator education level...) as well multilingual versions of standardized survey questions for common measurements of interest (e.g. income, time use, expenses, firm profits...).

Getting internal validity right. One expert highlighted that because standard errors are often poorly calculated (Young, 2019), extrapolation issues may really be replication issues: the effects were not really there in the first place.

Several experts also emphasized to us the importance of understanding the distribution on outcomes and treatment effects, as well as understanding the heterogeneity in returns, especially across the relatively rich and poor.

Research incentives. Multiple field experts brought up caveats regarding the way researchers are incentivized to follow socially desirable research practices. One was concerned that badges and "gold standards" lead researchers to embrace oversimplified heuristics and develop misguided lexicographic preferences about study design.

The other expert highlighted that researchers' willingness to speculate about external validity and identify potential mechanisms may vary over the course of a literature. It is possible that early on, researchers are less keen to discuss mechanisms since the facts alone are novel. As a literature matures, clarifying mechanisms is a natural way for researchers to provide differentiating and value-added analysis.

References

Al-Ubaydli, O., J. A. List, and D. L. Suskind (2017): "What can we learn from experiments? Understanding the threats to the scalability of experimental results," *American Economic Review*, 107, 282–86.

Angrist, J. D. and W. N. Evans (1998): "Children and their parents' labor supply: evidence from exogenous variation in family size," *The American Economic Review*, 88, 450–477.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, 91, 444–455.

Ashraf, N., J. Berry, and J. M. Shapiro (2010): "Can higher prices stimulate product use? Evidence from a field experiment in Zambia," *American Economic Review*, 100, 2383–2413.

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (2019a): "The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely," Tech. rep., National Bureau of Economic Research.

Athey, S., J. Tibshirani, S. Wager, et al. (2019b): "Generalized random forests," *The Annals of Statistics*, 47, 1148–1178.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017a): "From proof of concept to scalable policies: Challenges and solutions, with an application," *Journal of Economic Perspectives*, 31, 73–102.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2014): "Gossip: Identifying central individuals in a social network," .

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg (2020): "A theory of experimenters: Robustness, randomization, and balance," *American Economic Review*, 110, 1206–30.

Banerjee, A. V., S. Chassang, and E. Snowberg (2017b): "Decision theoretic approaches to experiment design and external validity," in *Handbook of Economic Field Experiments*, Elsevier, vol. 1, 141–174.

Bertanha, M. and G. W. Imbens (2019): "External validity in fuzzy regression discontinuity designs," *Journal of Business & Economic Statistics*, 1–39.

Björkman Nyqvist, M., L. Corno, D. De Walque, and J. Svensson (2018): "Incentivizing safer sexual behavior: evidence from a lottery experiment on HIV prevention," *American Economic Journal: Applied Economics*, 10, 287–314.

Blackwell, D. (1956): "An Analog of the Minimax Theorem for Vector Payoffs," *Pacific Journal of Mathematics*, 6, 1–8.

Breiman, L. (2001): "Random forests," *Machine learning*, 45, 5–32.

Burchardi, K. B., S. Gulesci, B. Lerva, and M. Sulaiman (2019): "Moral hazard: Experimental evidence from tenancy contracts," *The Quarterly Journal of Economics*, 134, 281–347.

Chassang, S., M. Chen, and M. Kolesár (2021): "Measuring the option value of treatments," *Princeton University Working Paper*.

Chassang, S., P. I. Miquel, E. Snowberg, et al. (2012): "Selective trials: A principal-agent approach to randomized controlled experiments," *American Economic Review*, 102, 1279–1309.

Cohen, J. and P. Dupas (2010): "Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment," *The Quarterly Journal of Economics*, 1–45.

Corno, L., N. Hildebrandt, and A. Voena (2020): "Age of marriage, weather shocks, and the direction of marriage payments," *Econometrica*, 88, 879–915.

Deaton, A. and N. Cartwright (2018): "Understanding and misunderstanding randomized controlled trials," *Social Science & Medicine*, 210, 2–21.

Dehejia, R., C. Pop-Eleches, and C. Samii (2019): "From local to global: External validity in a fertility natural experiment," *Journal of Business & Economic Statistics*, 1–27.

Dehejia, R. H. (2005): "Program evaluation as a decision problem," *Journal of Econometrics*, 125, 141–173.

DellaVigna, S. and D. Pope (2018): "Predicting experimental results: who knows what?" *Journal of Political Economy*, 126, 2410–2456.

DellaVigna, S., D. Pope, and E. Vivalt (2019): "Predict science to improve science," *Science*, 366, 428–429.

Dupas, P. (2014): "Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment," *Econometrica*, 82, 197–228.

Dupas, P. and J. Robinson (2013): "Why don't the poor save more? Evidence from health savings experiments," *American Economic Review*, 103, 1138–71.

Evans, D. and A. Popova (2014): "All those external validity issues with impacts? They apply to costs too!" *World Bank Blog: Development Impact*.

Fink, G., B. K. Jack, and F. Masiye (2020): "Seasonal liquidity, rural labor markets, and agricultural production," *American Economic Review*, 110, 3351–92.

Fischer, G., D. Karlan, M. McConnell, and P. Raffler (2019): "Short-term subsidies and seller type: A health products experiment in Uganda," *Journal of Development Economics*, 137, 110–124.

Foster, D. and R. Vohra (1999): "Regret in the Online Decision Problem," *Games and Economic Behavior*, 29, 7–35.

Gechter, M., C. Samii, R. Dehejia, and C. Pop-Eleches (2018): "Evaluating ex ante counterfactual predictions using ex post causal inference," *arXiv preprint arXiv:1806.07016*.

Hannan, J. (1957): "Approximation to Bayes Risk in Repeated Play," *Contributions to the Theory of Games*, 3, 97–139.

Heckman, J. J. and E. Vytlacil (2005): "Structural equations, treatment effects, and econometric policy evaluation 1," *Econometrica*, 73, 669–738.

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

Jack, B. K. (2013): "Private information and the allocation of land use subsidies in Malawi," *American Economic Journal: Applied Economics*, 5, 113–35.

Jayachandran, S. and R. Pande (2017): "Why are Indian children so short? The role of birth order and son preference," *American Economic Review*, 107, 2600–2629.

Karlan, D. and J. Zinman (2009): "Observing unobservables: Identifying information asymmetries with a consumer credit field experiment," *Econometrica*, 77, 1993–2008.

Kowalski, A. E. (2018): "How to Examine External Validity Within an Experiment," Tech. rep., National Bureau of Economic Research.

LaLonde, R. J. (1986): "Evaluating the econometric evaluations of training programs with experimental data," *The American economic review*, 604–620.

Meager, R. (2016): "Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature," *Manuscript: MIT*.

Morgan, K. L. and D. B. Rubin (2012): "Rerandomization to improve covariate balance in experiments," *The Annals of Statistics*, 40, 1263–1282.

Muralidharan, K. and P. Niehaus (2017): "Experimentation at scale," *Journal of Economic Perspectives*, 31, 103–24.

Narita, Y. (2021): "Incorporating ethics and welfare into randomized experiments," *Proceedings of the National Academy of Sciences*, 118.

Padró i Miquel, G., N. Qian, Y. Xu, and Y. Yao (2015): "Making Democracy Work: Culture, Social Capital and Elections in China," .

Rosenzweig, M. R. and C. Udry (2020): "External validity in a stochastic world: Evidence from low-income countries," *The Review of Economic Studies*, 87, 343–381.

Rubin, D. B. (2005): "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, 100, 322–331.

Simmons, J. P., L. D. Nelson, and U. Simonsohn (2011): "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological science*, 22, 1359–1366.

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

Vivalt, E., A. Coville, and K. Sampada (2019): "Weighing the Evidence: Which Studies Count?" Tech. rep., Working Paper, 2021. and,"How do policymakers update.

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

Weiss, C. H. (1997): "Theory-based evaluation: past, present, and future." *New directions for evaluation*, 76, 41–55.

Weiss, C. H. et al. (1995): "Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families," *New approaches to evaluating community initiatives: Concepts, methods, and contexts*, 1, 65–92.

Young, A. (2019): "Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results," *The Quarterly Journal of Economics*, 134, 557–598.