Holding back from theory: limits and methodological alternatives of randomized field experiments in development economics

Judith Favereau and Michiru Nagatsu

QUERY SHEET

This page lists questions we have about your paper. The numbers displayed at left are hyperlinked to the location of the query in your paper.

The title and author names are listed on this sheet as they will be published, both on your paper and on the Table of Contents. Please review and ensure the information is correct and advise us if any changes need to be made. In addition, please review your paper as a whole for typographical and essential corrections.

Your PDF proof has been enabled so that you can comment on the proof directly using Adobe Acrobat. For further information on marking corrections using Acrobat, please visit <http://journalauthors.tandf.co.uk/production/acrobat.asp>; <https://authorservices.taylorandfrancis.com/how-to-correct-proofs-with-adobe/>

The CrossRef database (www.crossref.org/) has been used to validate the references. Changes resulting from mismatches are tracked in red font.

AUTHOR QUERIES

QUERY NO.	QUERY DETAILS
Q1	Please provide missing affiliations for all authors.
Q2	The disclosure statement has been inserted. Please correct if this is inaccurate.
Q3	Please provide missing city and publisher for the “Anderson (2008)” references list entry.
Q4	Please provide missing city for the “Banerjee et al. (2011)” references list entry.
Q5	Please provide missing city for the “Bardsley et al. (2010)” references list entry.
Q6	Please provide missing city for the “Collier (2007)” references list entry.
Q7	Please provide missing city for the “Easterly (2006)” references list entry.
Q8	Please provide missing city for the “Moyo (2009)” references list entry.
Q9	Please provide missing page range for the “Ravallion (2009)” references list entry.
Q10	Please provide missing city for the “Sachs (2005)” references list entry.
Q11	Please provide missing city for the “WHO (2009)” references list entry.

Holding back from theory: limits and methodological alternatives of randomized field experiments in development economics

Q1 Judith Favereau* and Michiru Nagatsu*

ABSTRACT

In this paper, we critically and constructively examine the methodology of *evidence-based development economics*, which deploys randomized field experiments (RFEs) as its main tool. We describe the context in which this movement started, and illustrate in detail how RFEs are designed and implemented in practice, drawing on a series of experiments by Pascaline Dupas and her colleagues on the use of bednets, saving and governance in Kenya. We show that this line of experiments have evolved to address the limitation of obtaining policy-relevant insights from RFEs alone, characterized as their lack of external validity in the literature. After examining the two prominent responses by leading figures of evidence-based development economics, namely machine learning and structured speculation, we propose an alternative methodological strategy that incorporates two sub-fields, namely experimental economics and behavioral economics, to complement RFEs in investigating the data-generating process underlying the treatment effects of RFEs. This strategy highlights promising methodological developments in RFEs neither captured by the two proposals nor recognized by methodologists, and also guides how to combine different sub-fields of economics.

ARTICLE HISTORY

Received 19 December 2019
Accepted 14 January 2020

KEYWORDS

Randomized field experiments; randomized controlled trials; development economics; experimental economics; behavioral economics; external validity

JEL CODES

B41; C93; D04; I15

1. Introduction

In recent years, development economics has undergone a 'credibility revolution' (Angrist & Pischke, 2010), namely the extensive use of randomized field experiments (henceforth RFEs). The recent Nobel Prize awarded to Abhijit Banerjee, Esther Duflo and Michael Kremer acknowledges this credibility revolution. The use of RFEs in development economics has been promoted most prominently by the Jameel Abdul Latif Poverty Action Lab (J-PAL), founded by Abhijit Banerjee, Esther Duflo and Sendhil Mullainathan in 2003 at the Massachusetts Institute of Technology. The goal of this laboratory is twofold: (1) to produce evidence on the effectiveness of development programs and then (2) to use this evidence to guide policy makers. These goals are motivated by the fact that many development programs have failed, and that researchers, aid agencies, and policy makers still do not know 'what works' to fight poverty, thus requiring relevant evidence (Banerjee & Duflo, 2011). An RFE assesses a development program's impact by comparing its effects on (at least two) groups that are randomly assigned different treatments. The random allocation of the program treatments in this experimental design removes statistical biases that other experimental or non-experimental methods do not. Hence, randomized experiments guarantee the reliability of the experimental results, offering a strong internal validity, which is why such an experimental method is often considered as the 'gold standard'.¹ In order to produce evidence following the recent movement of evidence-based policy, the J-PAL's approach emphasizes the role of RFEs as a major tool for development policies.

CONTACT Judith Favereau ✉ judith.favereau@gmail.com

*Both authors equally contributed to the work.

© 2020 Informa UK Limited, trading as Taylor & Francis Group

The J-PAL's approach can be characterized as *evidence-based development economics* (EBDE),² which aims to be 'theory agnostic' before implementing the experiment. Thereafter one builds a theoretical framework based on the evidence from the experimental results (Banerjee, 2005), without taking a stance in the existing theoretical debates. Instead, it tries to find effective means to alleviate poverty relying on evidence from RFEs. The rationale, according to Banerjee and Duflo, is that theory can be a bad guide to experiments.³ Indeed, they claim that development economic theories tend to be contradictory, which confuses the public and stakeholders, preventing them from taking political actions against poverty. The emblematic case of such contradictory positions in development economics discussed in Banerjee and Duflo (2011) is the development aid debate. Two main positions are confronted in this debate. The first one, advocated by Sachs (2005), claims that the poor are trapped in poverty and the only way for them to escape is to benefit from a massive aid plan which would restore their lost capitals. The second one, championed by Easterly (2006), argues against this position, maintaining that international aid is in fact responsible for the slow development with its many perverse effects.⁴ Banerjee and Duflo (2011) argue that evidence is needed to settle this debate.⁵ Hence, J-PALs researchers aim at offering a third way to approach this debate: to determine through RFEs whether or not international aid is efficient in fighting poverty. The RFEs that are most likely to generate internally valid results has to be the starting point of evidence-based development economics, and then theory should follow from the experimental results (i.e. the evidence).⁶ An innovative character of RFEs on this view is that they can sometimes demonstrate a significant effect of those interventions unexpected or inexplicable from theoretical perspectives (and conversely that they can show that theoretically expected effects do not manifest in the field). This is all encouraging because evidence-based development economics might find effective interventions to alleviate poverty without getting involved in endless theoretical debates.

However, several methodological criticisms of J-PAL's experiments have been raised from a wide range of fields, such as development economics (Acemoglu, 2010; Barrett & Carter, 2010; Basu, 2014; Deaton, 2010; Ravallion, 2009; Rodrik, 2008), experimental economics (Harrison, 2011, 2013), econometrics (Heckman, 1992; Heckman, Ichimura, Smith, & Todd, 1998; Heckman & Smith, 1995; Heckman, Smith, & Clements, 1997; Leamer, 2010), and philosophy of science (Cartwright, 2007, 2009, 2010; Cartwright & Hardie, 2012; Davis, 2013; Guillin, 2013; Teira, 2013; Teira & Reiss, 2013). Although they target different methodological, practical, and technical aspects of RFEs, they all reiterate the need to identify the mechanisms underlining the results of the experiments to facilitate their effective applications in policy. We agree with the gist of this literature, which we call the mechanism- or process-based criticism. However, while the literature has converged on the criticism, its positive proposals have not, and it remains unclear how RFEs should be conducted differently in practice. This is partly because we currently lack a detailed examination of actual RFE studies in which the methodological and theoretical problems figure, as well as a careful evaluation of J-PAL's recent proposals in response to the mechanism-based criticism (e.g. Banerjee, Chassang, & Snowberg, 2017).

We bridge these gaps between the mechanism-based criticism of RFEs and the contemporary practices of RFEs by providing a constructive and concrete methodological strategy, based on a careful case study. This scrutiny points to two closely related problems that J-PAL's approach faces, one methodological and the other theoretical. Both concern the external validity of RFEs. We discuss the external validity issue as reincarnated in the development of Dupas's experiments (Cohen & Dupas, 2010; Dizon-Ross, Dupas, & Robinson, 2017; Dupas, 2009, 2011, 2014; Dupas & Robinson, 2010, 2012, 2013a, 2013b), and consider the two recent proposals to address the external validity issue of RFEs, namely *machine learning modeling* (Banerjee, Chassang, Montero, & Snowberg, 2017) and *structured speculation* (Duflo, 2018). Although these proposals make evidence-based development economics more sophisticated, we argue that the theoretical and methodological problems can be addressed more up-front by realizing the complementarity between RFEs and other types of field experiments developed in behavioral and experimental economics. Our proposal is motivated by the rise of 'field experiments' (Harrison & List, 2004), or 'extra-laboratory experiments' (Charness,

Gneezy, & Kuhn, 2013) in these sub-fields. We argue that the systematic complementary use of lab and field experiments, rather than using RFEs in isolation, will more effectively provide knowledge of the underlying mechanisms of behavior of the poor, which will be of use for development policy. Our proposal reflects the economists' emerging attempts to create synergies between behavioral, experimental and development economics, which have not been noticed by the philosophical critics of randomized field experiments.

We proceed as follows: Section 2 summarizes the much-discussed issue of external validity of RFE-based policy recommendations, both in general and in specific contexts of our case study. This section also examines two prominent reactions to the critics of RFEs from J-PAL's researchers. Section 3 proposes our methodological strategy (from behavioral and experimental economics) that aims to illuminate the mechanisms underlying the economic phenomena that development economists are interested in. Section 4 concludes.

2. From 'black box' to plausible mechanisms: external validity issues of randomized experiments in development economics

The goal of this section is two-fold. First, we want to extract the gist of the disparate critiques as issues of external validity in RFEs (Section 2.1). Second, we want to move this debate forward by evaluating J-PAL's researchers and associates' recent responses to the criticism. We will discuss two prominent kinds of responses: a promotion of *machine learning modeling* to detect the relevant heterogeneity underlying the treatment effects of RFEs (Chernozhukov et al., 2018; Duflo, 2018) (Section 2.2); and a formulation of a standardized guideline for conjectures, or *structured speculation* on the mechanisms and generalizability of the observed treatment effects (or the lack thereof) (Banerjee, Chassang, & Snowberg, 2017) (Section 2.3).⁷ We emphasize that although these responses address some of the problems in RFEs they miss a more promising pathway, which we will take in the next section.

Focusing on concrete cases of RFEs is pertinent to these goals: it illustrates concrete practices of evidence-based development economics, against which some of the abstract critiques can be assessed. Indeed, RFEs's criticisms from economics or philosophy of science often appear to deal with RFEs in general (e.g. they are not discipline-specific) and thus they tend to offer alternatives that can be seen as abstract. Moreover, focusing on concrete cases of RFEs also helps us evaluate the recent methodological proposals by the J-PAL's researchers since some of them are already manifest there in practice, suggesting that J-PAL's researchers are aware of the external validity concerns. In this sense, their work provides an ideal case in which both the criticism towards and the responses from evidence-based development economics can be evaluated. In addition, drawing from concrete cases enables us to highlight how RFEs in development economics tackle the development aid debate.

2.1. External validity issues: from the abstract to the concrete

The crucial questions are: Can policy-makers use results from RFEs in order to implement efficient development programs? And if so, how? These questions concern *external validity*⁸ – more specifically, the validity of the inference to the effect that the same or similar interventions as in an RFE will produce similar results in another non-experimental context. This leads to two more specific questions. Imagine a policy maker in Kenya benefiting from a very promising RFE's result (e.g. showing the efficacy of a specific development program for one specific region of her country). First, can this policy maker implement with efficacy such a development program in another region of her country? This refers to what we call local external validity. What is at stake is for the RFE's result to travel from one specific region to another. In other words, to go from local to local. Second, can this policy maker scale up the program to the whole of Kenya? It refers to what we call global external validity, the issue being to implement the development program on a global scale. If the size of the program implementation is changed, other effects might interact with the

one of the program, leading to substantial changes of its previous effects. As stated above, a growing body of literature in both philosophy of science and economics has challenged the external validity of RFEs in development economics. To put it somewhat categorically, the philosophical literature concentrates on the local external validity, whereas the economic literature focuses on the global external validity issues.

For instance, the philosophical literature (Cartwright, 2007, 2009, 2010; Cartwright & Hardie, 2012; Guillin, 2013) highlights that RFEs do not demonstrate causal capacities but causal laws instead. That is to say, the RFEs's results offer a proper isolation of a causal effect. The identified causal inference is reliable, although of no use *per se*. In order for RFEs' results to be used, one needs to identify *how* and *why* such causal inference occurs and in which context. Let us go back to the Kenyan policy maker. The RFE's result tells her in a reliable way that in one specific region of her country one specific program is efficient. However, it does not tell her how to use such an result, if she aims to implement such a program in another city. According to Cartwright and Hardie (2012), in order to reply to this question, first one needs to precisely (and in the abstract) define the causal inference at play when the program is implemented. Second, the goal is to highlight the support factors at play that allows such a causal capacity to happen. The first step is what Cartwright and Hardie (2012) call a horizontal search and the second one is a vertical search. These elements are factors that support the program tested. Indeed, if children are not in good health they would have troubles to follow in class, diminishing their learning; whereas if they are in good health, they would follow more easily and might increase their learning. The horizontal search aims to define the causal role of the program. For instance, if the program tested was an educational program that aims to increase the learning of students, the causal role would be: increasing the learning. The vertical search defines the support factors. For example, concerning the increase of learning, the kids might need to be in good health, to be present at school, to benefit from enough textbooks or materials to study, etc.

The literature in economics (Acemoglu, 2010; Barrett & Carter, 2010; Basu, 2014; Deaton, 2010; Ravallion, 2009; Rodrik, 2008) also highlights the need to understand *why* a development program is working or not. This literature insists on the theory that is needed to pursue such a task. In particular, Acemoglu (2010), Deaton (2010), Harrison (2011, 2013), Heckman (1992), Heckman and Smith (1995), Heckman et al. (1998, 1997), Leamer (2010) emphasize the importance of structural estimation in econometrics. Indeed, the econometric estimation drawing from RFEs in development economics is a reduced form, again isolating one causal effect. Such a causal effect is thus completely detached from other factors, and their interactions are not defined. In a nutshell, these criticisms concentrate on the fact that RFEs produce a result termed: 'black-box test of "what works"' (Deaton, 2010, p. 451). In other words, RFEs' results would not illuminate the mechanisms behind such results, making their possible evidential use difficult or even impossible. For instance, Acemoglu (2010) shows how, in first defining a theoretical framework, and second in writing the econometric equation, the Kenyan policy maker would be able to determine the effect of the RFE's result on a global scale. Heckman (2010) defines through the 'Marshallian maxime' an econometric way that would allow minimizing the number of hypotheses (guarantying the internal validity of the RFE) but that would also be a structural estimation (allowing for the causal channel at play and enabling the identification of the global scale effect). Following Heckman (2010), Deaton (2010) emphasizes the need of a theoretical framework.⁹ Put simplistically, these economic criticisms and alternatives deal with global external validity, although, the role theory should play in explaining RFEs as well as structural econometric estimations are meant to account for local external validity as well.

The problem is that these critiques from different sub-fields of economics and philosophy of science are rarely brought together and thus remain disparate. Deaton and Cartwright (2018a, 2018b) are a notable recent exception that attempts to bridge this gap, although their focus is principally on warning that the apparent strong *internal* validity of RFEs is not as strong as J-PAL's researchers would claim, highlighting that there is no methodological gold standard.¹⁰ What is needed is thus to bring together these criticisms and alternatives. Moreover, in order to attempt to offer some highlights to the Kenyan policy maker, it is necessary to scrutinize what the RFEs's

results are in practice, in order to illustrate them. As stated above, most of the criticisms of RFEs seem to be dealing with RFEs in general and rarely scrutinize how a series of RFEs are motivated and conducted in practice. Doing so enables us to identify precisely *what* the results of RFEs are, and *whether* and *how* they could concretely guide a policy maker. In addition, this allows us to bring together the RFEs's criticisms from philosophy of science and economics, as well as assess the J-PAL's responses to these criticisms.

For these purposes, we focus on a series of RFEs conducted by Pascaline Dupas and colleagues (Cohen & Dupas, 2010; Dupas, 2009, 2011, 2014; Dupas & Robinson, 2010) as a case study (see Box 1). Those experiments are an emblematic case of evidence-based development economics in that they aim to offer the needed evidence to settle the development aid debate and to guide policy makers of developing countries. As RFEs they are implemented at a local scale, focusing on a miniature of the development aid debate: whether bednets have to cost something (Easterly's view) or be fully subsidized (Sachs' point of view). Harrison (2011, p. 636) dubs this the 'sell them cheap or give them away' debate.

What should the policy maker do based on these results? Should she invest massively on bednets and fully subsidize them? Should she use the existing infrastructure of her country to deliver the bednets? Moreover, should she complement her investment in bednets with the implementation of a safebox in order to favor saving among members of her population? Would this saving on a global scale have a negative impact on the economy? How to understand the stronger saving barriers for women? These experiments, although highlighting key effects related to the development aid debate such as the elasticity of the demand for bednets, or the willingness to pay for bednets after having benefited from them for free, do not clearly identify the processes behind these effects. To make reliable inferences from these experimental results to different populations or on different scales, the processes have to be systematically investigated and better understood. The external validity criticisms from economics or philosophy of science relate to this. In the following, we will use this set of experiments (in Box 1) as a reference point, and keep coming back to these experiments to discuss how to investigate the processes underlying RFEs. This case enables us to examine in more detail how J-PAL's researchers have tried to find out the mechanisms behind the effects of RFEs, in these experiments and beyond.

Box 1. From 'sell them cheap or give them away?' to the emerging new puzzle

Cohen and Dupas (2010) studied how the ways of distributing insecticide-treated bednets (ITNs) to fight malaria affect their uptake (see Deaton, 2010; Harrison, 2013). The experiment was set up in 2007 in Kenya, where malaria is endemic.¹¹ The experimental results, in highlighting the elasticity of the demand for ITNs and thus suggesting the need to subsidize such preventive health products in line with the World Health Organization (2009) recommendation, offer evidence on a vivid debate within development economics. In addition, Dizon-Ross et al. (2017) shows that the most cost-effective way to deliver health products such as bednets is through the already existing infrastructure.

Going one step further from the development aid debate, Dupas (2014) aims to determine if benefiting from free bednets will decrease the willingness to pay for such bednets in the future. This question is relevant because the user needs to restore the insecticide on an ITN every six or twelve months.¹² Here also, the experiment aims to provide evidence that will answer how short-run subsidies of a health product affect future purchases of the health product. The experiment shows that benefiting from free bed-nets actually increases the willingness to pay for bednets in the long run. In addition, Dupas (2009, p. 228) highlights that the cheaper the bednets, the more likely it is that children would sleep under it. Dupas (2009, p. 230) concludes this study with what she calls a 'puzzling fact'. Individuals do not invest in LLINs, while LLINs are intensively used when obtained. This puzzling fact reframes the development aid debate. The question is no longer (at least for the population of the experiment) whether bednets should be given for free or not. The question becomes: why don't the poor invest in health products that are proven effective? Is the problem credit constraint, present-biased, or something else?

Dupas and Robinson (2013a, 2013b) aim to tackle this 'puzzling fact'. First, they investigate the savings constraints of the poor. According to Dupas and Robinson (2013b), the common reason explaining the puzzle is that the poor have credit constraints.¹³ However, Dupas and Robinson (2013b) consider the cost for such products to be affordable, and therefore they reason that it would be possible for the poor to save a little in order to buy a bednet. In this experiment they thus intend to identify 'the key barriers to saving' (Dupas & Robinson, 2013b, p. 1139). They conclude that the 'most striking result is that getting access to a safe and designated storage technology as simple as a *Safe Box* can have large and lasting impacts on health savings and investment behavior' (Dupas & Robinson, 2013b, p. 1163). Moreover, Dupas and Robinson (2013a) emphasize that women face stronger savings barriers than men.

2.2. Are there any mechanisms revealed?: modeling and heterogeneity of treatment effect

One option to investigate the mechanisms behind the result of an RFE is to build an econometric model of the data-generating process of the treatment effects, as the economic criticisms highlighted. Following on her experiments, Dupas and her colleagues use two such models, both of which only concern the first experimental block on the cost-sharing of bednets. The question is whether or not such modeling in defining inferences would help to promote externally valid inferences from the RFEs. As already emphasized, their first experiment on the cost sharing of bednets (Cohen & Dupas, 2010) leads to what Dupas (2011, p. 428) calls a ‘puzzling’ fact: individuals do not invest in preventive health products although they are less costly than curative products, suggesting that these individuals are credit and/or savings constrained. In this experiment they use a model to determine which parameters the experiment should estimate. The model defines two uses of bednets: a health use (to hang a bednet) and a non-health use (to keep one unhung)¹⁴. They, then, define two utilities for these two uses. The model is thus defined by the addition of these two utilities minus the difference between the cost and the subsidies of the bednets for hung and non-hung bednets, in addition to the positive externalities that they generate. This modeling allows them to define the ‘efficient subsidy level’ for bednets. This optimal level of subsidy is based on the Pigou-Dalton principle: between two utility profiles, the one which decreases inequality between the agents should be chosen. The experiment helps to identify by how much the number of hung bednets decreases when the price increases. Note that this model and the experiment do not directly address the puzzling fact. It is thus difficult to be confident that the optimal level of subsidy can be determined by this model alone. It does not offer a clear answer for the Kenyan policy maker.

Perhaps, as suggested by the economic literature, if the model was structural rather than a simple, reduced-form one, some proper mechanisms could have been identified. Dupas (2014) thus uses a structural model when assessing the willingness to pay after benefiting from a free bednet. The model defines two forms of externalities for bednets: health spillovers and information spillovers. The households are supposed to be myopic (they do not anticipate the externalities), but are well aware of the disease burden. The model also defines two periods. In period 1, the households obtain the bednets and update the beliefs about their quality and private returns. A year later, in period 2, the bednets have a new price. The disease risk is different than in period 1 and the price of period 1 acts as a reference point. Unlike period 1, the spillover effect cannot be estimated. The model identifies the learning from experimentation and the reference dependence parameter, in order to determine if there is an anchoring effect. Therefore, the model links together learning from product use and the willingness to pay for it in another period with the possibility to anchor oneself to the original price. The model enables the identification of a causal interaction between the learning effect and the anchoring one with the willingness to pay. However, the model estimation is still dependent on the sample population. In the words of Cartwright and Hardie (2012), this model, by highlighting a causal interaction, might help in the horizontal search but not in the vertical one. We still do not know whether the results from the estimation of the model through randomized experiments would be generalizable to another population. Some ingredients are still missing. In other words, it does not help with external validity issues.

For that, in addition to the model, characteristics about the population of the experiments are needed. These characteristics enable researchers to disaggregate the heterogeneous responses to the experimental treatment, correlating individual-level treatment effects and their characteristics, e.g. rich households responding to a given intervention more than poor households. If we have some background knowledge as to why this difference exists, this knowledge can be then utilized in inferring the effects of a similar intervention in another population (e.g. by looking at the composition of rich and poor households of that population). Some statistical and econometric techniques help identify such correlations between known individual characteristics (such as gender, age, income, etc.) and the impact of an intervention (see for example Crump, Hotz, Imbens, & Mitnik, 2008 and Athey & Imbens, 2017).

Recently, *machine learning* (ML) modeling has also been proposed as a new approach to the problem of heterogeneous treatment effect (e.g. Duflo, 2018).¹⁵ ML modeling has been developed in computer science and the related fields as a new method, or a method to improve existing methods, to analyze large data sets. It is impossible to give an overview of ML and its use in economics here, but it suffices to note that the main idea of combining it with RFEs is to let algorithms classify individuals *ex post* according to their levels of responses to the intervention, and further investigate their possible differences in the measured characteristics. Knowing such classification can help understand the mechanisms behind the experimental data. Duflo (2018) emphasizes that machine learning modeling has already been used to guide predictions (e.g. Davis & Heller, 2017; Hussam, Rigol, & Roth, 2016).

In line with the original theory-agnostic motivation of the new development economics movement, Chernozhukov et al. (2018) and Duflo (2018) highlight the virtue of a machine learning approach as making as few assumptions as possible:

In this paper we take an agnostic view. We neither rely on any structured assumptions, which might be difficult to verify or believe in practice, nor impose conditions that make the ML estimators consistent. We simply treat ML as providing proxy predictors for the objects of interest. (Chernozhukov et al., 2018, p. 5)

Duflo (2018) adopts the same agnostic attitude in ML modeling and characterizes its goal as ‘to give up on estimating all the possible heterogeneity but focus on a limited number of core features (is there heterogeneity? what are the characteristics of those with the largest treatment effect?)’ (Duflo, 2018, p. 35).

A first application (Crépon, Devoto, Duflo, & Parienté, 2015), according to Duflo (2018), assesses the impact of a micro-credit in Morocco. The main results are that only 13% of the treatment group took a micro-credit and no one in the control group did. This micro-credit had no impact on income and consumption, but it increased profit and investments. Through machine learning modeling, it is possible to distinguish between, for example, the 10% most and least affected by the program. Then, accessing some of their characteristics, such as age, agricultural employment, and other borrowing, might help to find out key characteristics that correlate with the strong or weak treatment effects. Therefore, ML modeling might help the researcher understand why some heterogeneity exists. Doing so might help in conducting a vertical and horizontal search. For the Kenyan policy maker, being able to identify within the experimental sample for example, for which subgroups the demand elasticity for bednets was the smallest might be of importance. Moreover, identifying for which subgroups the safebox had the larger impact on saving might also be useful. This access to the distribution is a key for the understanding of the mechanisms behind RFEs’ results.

However, ML modeling as such offers no clear guidelines regarding which dimension(s) of heterogeneity of treatment effect one should focus on. In addition, the agnostic approach, in focusing only on key features, may lead researchers to leave out a relevant heterogeneity they did not think about. Just like other statistical approaches, unmeasured but potentially relevant characteristics (e.g. subjective beliefs and risk preferences) cannot be utilized, either. A crucial question remains: how to determine the key factors for heterogeneity? This tends to diminish the power that ML modeling could offer in helping to discover potential mechanisms. ML is a technical innovation within the paradigm of theory aversion, i.e. systematic search for patterns in the data while limiting the number of assumptions. As such, it cannot provide the researcher with reasons to believe that some mechanisms, rather than others, are at work in the data-generating process. This is the main limitation of ML in improving RFEs’ external validity.

2.3. Only plausible mechanisms?: possible explanations and structured speculations

The second proposal of the J-PAL’s researchers might be seen as offering a set of guidelines in the search for mechanisms. Banerjee, Chassang, and Snowberg (2017) develop a decision-theoretic approach for experiments to address the external validity issue. A model was first developed in

Banerjee, Chassang, Montero, et al. (2017) to formalize the decision-making process of experimenters choosing between randomizing or not. If the audience is skeptical, the experimenters would prioritize internal validity and would thus randomize.¹⁶ In addition to such formalization, Banerjee, Chassang, Montero, et al. (2017) address external validity concerns in their approach. They define external validity as ‘the following problem: after running an experiment in environment z_e , the experimenter is asked to make a recommendation for the external environment z_p ’ (Banerjee, Chassang, & Snowberg, 2017, p. 24). They then conclude that the recommendation for z_p reflects subjective beliefs of the experimenters. In other words, external validity is essentially a subjective matter.¹⁷ The goal then is to render such subjective judgements as honest and truthful as possible: ‘What we should aim to do is extract the experimenter’s honest beliefs about the efficacy of the treatment in different environments’ (*Ibid*). Banerjee, Chassang, and Snowberg (2017, p. 26) offer three guidelines: (1) experiments have to speculate about the external validity of their findings; (2) this speculation should be written in a separate section in their academic papers; and (3) such speculations have to be falsifiable. These guidelines ‘structure’ the speculation of the experimenters on the mechanisms and generalizability of the results from RFEs. Hence the term *structured speculation* for this approach. It is supposed to (i) guarantee a means to transmit information that would otherwise be lost; (ii) encourage formulation of precise and falsifiable hypotheses; and (iii) give incentives to the experimenters to design experiments that can address the external validity issues.

Banerjee, Chassang, and Snowberg (2017) illustrate their call for structured speculation with examples, mostly from Dupas’s work. For example, Dupas (2014) in assessing the willingness to pay after benefiting from a free bednet devotes a section to the external validity issues of her results. Dupas (2014, p. 222) asks three questions related to such issues: (1) Did the experimental design limit the scope of the anchoring effect?; (2) Are there cross-product entitlement effects of subsidies? Such an effect means that individuals, who benefited from a subsidy for a product, would expect that similar products would be subsidized as well, and thus do not acquire this last product waiting for it to be subsidized; and (3) What types of contexts and health products would lead to the same results? Dupas (2014, p. 223) concludes with three ‘conjectures’. First, the learning effect would be stronger for water filters. However, a one-time subsidy of chlorine would generate a small learning effect. Second, there is no cross-product entitlement effect of subsidies. Finally, a one-time subsidy of deworming would generate a negative learning effect and thus reduce its adoption. Such conjectures do not really highlight the underlining mechanisms; but, very interestingly, such conjectures do connect several experiments done by J-PAL’s researchers (such as Miguel & Kremer, 2004 on deworming, and Dupas, 2014 on water filters, and cook stoves). Nonetheless, as Dupas (2014) emphasizes, learning by doing might be an important channel for preventive health-product adoption. These exemplifies how structural speculations can be made after RFEs.

Moreover, Cohen and Dupas (2010) speculate on different effects, the selection effect and the psychological effect, neither of which were observed. As Banerjee, Chassang, and Snowberg (2017) emphasize, Cohen and Dupas (2010) focus on another aspect of external validity: the possibility to scale up the program. For that purpose, Cohen and Dupas (2010) conduct an analysis to identify sub-groups in the distribution. That might be done as the previous sub-section highlighted through ML modeling or, in a simpler version, with help of the observable distribution of the characteristics, or a baseline survey conducted prior to the experiments that allows stratifying the sample into sub-groups. For example, from April 2007 to October 2007, before her experiment, Dupas (2014) conducted a baseline survey on the households’ size, socio-economic status, access to electricity, etc. Such a survey may be extremely helpful to unpack the RFEs results. We will develop this point more in the next section.

Although Banerjee, Chassang, and Snowberg (2017) mention only Dupas (2014) and Cohen and Dupas (2010), structured speculation can also be found in Dupas’s other experiments. For instance, Dupas and Robinson (2013b) have a paragraph on mechanisms. In this paragraph they aim to define the mechanism behind their main result: devices like a safe box had a large effect on saving behavior. According to Dupas and Robinson (2013b), such simple devices have large effects

because they marked the money for health expenditures, making it easier for individuals to say no to relatives who asked for money, since the money is marked.¹⁸ Dupas and Robinson (2013a) also have a section on 'potential mechanisms'. In this section they raise two main questions: (1) Why do women need a savings account, when they could reinvest immediately in their business?; (2) Why is the private return higher for women than for male taxi drivers? Concerning the first question, they offer three possibilities. First, the investment can be bumpy. Second, the profit might vary. Finally, holding the capital in the business might make its use more difficult if a shock occurs; therefore, having the money in a bank account would be preferred to counteract against potential shocks. Concerning the private returns of savings for women, Dupas and Robinson (2013a) argue that women might be present-biased, or that having the money in a bank account prevents women from lending it to their husbands or relatives.

These speculations suggest potential mechanisms relevant for inferring whether or not a similar effect will obtain in RFEs and policy applications. One problem however is that it is not clear where these speculations about mechanisms come from. The idea of structured speculation implies that they are necessarily subjective. This is true in the trivial sense that inductive or abductive inferences are underdetermined by evidence. However, this does not mean that all the experimenter can do is make her speculations explicit and falsifiable, per experiment. In addition, she can ask where her prior beliefs come from: are they motivated by some well-established theory, or some related experimental findings?

As it turns out, a closer examination of our case reveals that the importance of these questions are recognized by the experimenters, who try to do something more than structured speculation in practice. Specifically, Dupas and Robinson (2010, 2012) aim to investigate the effects of the political crisis in Kenya¹⁹ in a series of experiments. In their first study, the experiment is rather instrumental: it allows isolating the effects of the political crisis on income and expenditures, which Dupas and Robinson (2010) refer to as 'descriptive evidence', whose mechanisms can be studied in further experiments. This opens the door to another potential use of evidence than that usually promoted by J-PAL's researchers. Furthermore, their second experiment (Dupas & Robinson, 2012) focusing on sex workers and their coping strategies refers to a theoretical model (the permanent income hypothesis) to guide the experiment. This theory is considered as a benchmark to estimate the relationship between income, consumption, and savings. This is the only experiment in our case study that emphasizes the feedback between experimental data and theory. This leads Dupas and Robinson (2012) to conclude that the coping strategies were not sufficient for sex workers. That is why, for the period, we observed an increase of risky transactional sex to make up for their lost income. This experiment is perhaps the only one offering a clear *explanation* and therefore an explicit mechanism behind the experimental result, allowing Dupas and Robinson (2012) to conclude that this could have larger implications for developing countries where formal insurance is rare.

In sum, we identified several promising lines of development in our case study, for instance (1) to design multiple RFEs in temporal phases, (2) to use results from RFEs as descriptive evidence to investigate more specific underlying mechanisms, (3) to use a theoretical model as an analytic framework and an empirical benchmark, and (4) to use baseline surveys (*ex ante*) to unpack the heterogeneity of treatment effects, in addition to *ex post* machine learning modeling of the data. We can interpret these strategies as a range of efforts to identify what Cartwright and Hardie (2012, p. 25) call *support factors* – INUS conditions (Insufficient but Necessary parts of a condition which is itself Unnecessary but Sufficient) for a given intervention to have an intended effect not only in z_e but also in z_p . We suggest that there is a discrepancy between the two prominent proposals (machine learning modeling and structured speculations) on the one hand, and these actual methodological developments in practice, on the other. That is, the former represents a refinement within the paradigm of theory aversion, with which new development economics started, while the latter represents a need to go beyond that paradigm and utilize the theoretical and methodological resources from elsewhere than from RFEs alone. In particular, (3) and (4) suggest that something more than

structured speculation is needed to guide mechanistic investigations. In the next section, we propose two resources to guide such investigations.

3. Unpacking the processes: integrating RFEs with experimental and behavioral economics

The aim of this section is to present two kinds of resources to effectively complement evidence-based development economics and its search for evidence. Unlike Banerjee, Chassang, Montero, et al. (2017), who propose a subjectivist decision-theoretic framework to make the experimenter's claims both honest and internally unbiased, we highlight the importance of the sources of the prior subjective beliefs of the experimenter. Unlike Duflo (2018), who advocates machine learning modeling as an *ex post* technique to address the heterogeneity of treatment effects in RFEs, we propose *ex ante*, a more proactive and integrative use of other types of experimental and theoretical resources. Our methodological proposals are motivated by some of the new methodological developments in RFEs reviewed in the last section, which suggest paths different from structured speculations and machine learning. Specifically, we propose to explicitly align RFEs with two other sub-fields, experimental economics and behavioral economics, to inform the underlying processes of the treatment effects. Experimental economics offers types of field experiments different from those RFEs that complement the latter. Behavioral economics provides explicit hypotheses regarding the mechanisms underlying behavior, specifically beliefs and preferences and their potential relations to institutions.

We will first introduce the field experiments in the tradition of experimental economics and why they have become popular. We then go one-step back and provide a general methodological analysis of economic experiments as *exhibits* (Section 3.1). We then argue that behavioral development economics (BDE) complements RFEs by providing insights into the behavior of the poor (Section 3.2). Finally, we illustrate how some of the recent hybridization of field experiments – a combination of RFEs and lab-like field experiments (LFEs) – are helping to unpack the black box of RFEs in practice (Section 3.3).

3.1. Economic experiments as regularity refinement

In experimental economics, independently from evidence-based development economics, there has been an increasingly popular trend for conducting experiments 'in the field', as opposed to 'in the lab' (Carpenter, Harrison, & List, 2005; Harrison & List, 2004). These experiments are called field experiments, but their origins are distinct from randomized field experiments [citation deleted]. An important question that drives this trend concerns whether the behavior observed in a controlled laboratory setting is a reliable indicator of the behavior in a naturally occurring setting outside the lab. This is a variation of the problem of external validity discussed in the last section, where the experimental setup (z_e) is in the lab, while the target environment of interest (z_p) is a non-experimental setup outside the lab.²⁰ Field experiments aim at addressing external validity thus specified by extending conventional laboratory experiments, adding one or more extra components that represent the natural settings in which economic decisions are made. As a reference for discussion we draw on Harrison and List (2004), who provide the widely adopted taxonomy, which can be schematically compared to a more recent taxonomy by Charness et al. (2013) (see Table 1).²¹

What are these kinds of field experiments good for? To answer this question, we first explain what economic experiments are good for in general, drawing on Bardsley et al. (2010), who provide a very useful methodological analysis of economic lab experiments as *devices to systematically search for, refine, and measure empirical regularities, as well as to explain them*. Bardsley et al. (2010, p. 156, 160) focus on empirical regularities captured by an experimental design, which they call an *exhibit*: 'a replicable experimental design that reliably produces some interesting result', or 'a mechanism that induces some specific *regularity* (or "effect," or "phenomenon") in human behavior'. Classic examples of exhibits include Public Goods games with Voluntary Contribution Mechanism, which

Table 1. Taxonomies of field experiments.

Labels			Variables			
Ours	Charness et al.	Harrison & List	Subject	Reference	Location/situation	Subjects' Awareness
F0	Conventional Lab		standard (students)	imposed	artefactual	yes
F1	Extra-lab	Artefactual field	nonstandard	imposed	artefactual	yes
F2		Framed field	nonstandard	field	artefactual	yes
F3	Field	n/a	nonstandard	field	natural	yes
F4		Natural field	nonstandard	field	natural	no

induce a steady decline of contributions in repeated rounds, or the Ultimatum games that show more-than-minimum offers by the Proposers and the Receivers' rejections of what they perceive as unfair offers. Note that a 'mechanism' here is defined operationally, that is, not in terms of the underlying *process* of the phenomenon, but in terms of the procedure to create it. This appears unhelpful because what we want to know in RFEs, as well as in other experiments, is the mechanism in the former sense. How does the study of the latter contribute to the understanding of the former?

Bardsley et al. (2010) provide a range of strategies. We highlight three of them that are most relevant in our context. First, a *robustness check* increases the reliability of the regularities by modifying a range of design variables of the original experiment. We follow the definition of robust experimental regularities of Bardsley et al. (2010, p. 179):²²

To say that a regularity observed in one experimental design is *robust* is to say that similar regularities are found in similar designs: minor differences in parameter values, instructions, incentives, subject pools, and so on do not make major differences to the result.

Robustness is tested both in undirected and directed ways. The undirected robustness check is done by contingent variations of the ways different experimental researchers conduct similar experiments with different 'house styles'. Despite those differences, which are impossible to eliminate in practice, some regularities have emerged, such as the ones mentioned above, enabling experimenters to cluster classes of experiments as exhibits. An analogous clustering can happen in RFEs as well, e.g. the under-investment in preventive health products has emerged as a regularity through experiments using different referent goods, such as bednets, deworming pills, and so on, in different locations. This can be seen as an example of an undirected robustness check. The more directed a robustness check is motivated by a skeptical hunch. For example, an apparently robust phenomenon may be caused by a particular aspect of the experimental design, such as inadequate incentives, framing of the task description, the elicitation procedure, the presence of the experimenter, etc. The variation of field experiments shown in Table 1 organizes these skeptical hunches as systematic ways of varying experimental designs. Specifically, the experimental variables, such as subject pools, reference goods, location or situation of the decision-making, and the subjects' awareness that they are under experimental investigation, pick up most salient dimensions that may affect the result and can be modified experimentally. Although the salience of these dimensions are not directly coming from formal theories, such salience guides decisions about what kind of experiments need be carried out in succession. The idea of the robustness check makes it clear that field experiments in this tradition are continuous with lab experiments, as an attempt to expand the 'library' of exhibits-phenomena developed in the former (Guala & Mittone, 2005).

Once some exhibit is identified, it invites more attention and research from the experimenters, as its robustness signals its significance to theory or its potential relevance to devising effective policy interventions. In particular, it invites more hypotheses regarding its mechanisms and generate competition between them. This is what Bardsley et al. (2010) calls *decomposition* or refinement, the second method of exhibit analysis. The Ultimatum game, for example, was modified by removing the Receiver's vetoing power (resulting in the Dictator game), to see whether the higher-than-predicted offers by the Proposers are caused by their fear of rejection (belief-dependent preferences),

or (belief-independent) preference for fair distributions. These are two distinct but potentially overlapping mechanisms that can be teased out by the variations of the two games.

An analogous case is found in the way Dupas (2014) tests two distinct mechanisms that may underlie the ‘puzzling’ fact from the RFEs that, although households become much less likely to buy LLINs as they become slightly costly, they are intensively used once obtained. The first one, *the anchoring effect hypothesis*, states that individuals anchor themselves to the low or no price, becoming more unwilling to pay in the future (Kőszegi & Rabin, 2006). The second one, *the learning effect hypothesis*, states that individuals have learned about the benefits of the health products by receiving them for free for the first time, becoming more willing to pay for them in the future (Ashraf, Berry, & Shapiro, 2010; Chassang, Padró I Miquel, & Snowberg, 2012).²³ To determine which effect is dominant, the experiment proceeds in two phases. The first phase is concerned with the adoption of the health products in the short-run. The second phase focuses on the long-run and assesses the willingness to pay, of those individuals who participated in phase one. The experiment was implemented in Kenya as was the previous experiment (e.g. Cohen & Dupas, 2010). The participating 1120 households were randomly assigned to 6 different levels of subsidies that varies from 100% to 40%. Twelve months after phase one, follow-up visits were conducted in phase two. During this visit, all households were offered a voucher with the same amount worth \$2.30. Dupas (2014) was thus able to compare the results from phase one and phase two. This comparison permits identifying the willingness to pay, of those households which had been exposed to high subsidies, and those which had been exposed to lower subsidies. Since, in phase two, everyone faces the same price, Dupas (2014) could determine the impact of different subsidy levels on the willingness to pay. The take-up of LLNs turns out to be higher for the groups who benefited from the higher level of subsidy (i.e. lower price), suggesting that the learning effect dominates the anchoring effect. Dupas’s (2014) structural model measures the magnitude of both effects and confirms the importance of the learning effect. The two-phase design of the experiments allows Dupas to identify learning as part of the puzzling fact: present-bias and credit constraints may still cause under-investment in bednets, but a one-time subsidy may counteract this as a remedy by enabling learning about the efficacy, rather than increasing the future unwillingness to pay in the case of bednets.

While both the robustness check and decomposition attempt to identify various mechanisms of exhibits (why a particular design induce a particular regularity), the third approach uses exhibits as devices to *measure* the variation of behavior in the field-sample population, possibly correlated with some demographic characteristics, such as gender and income. Some experimental designs (of F1 type in our Table 1) have established a status as portable measurement devices (Guala & Mittone, 2010). For example, the Dictator game is often used as a measure of belief-independent other-regarding preferences. The Public Goods game can be used as a measure of free-riding as well as conditional preferences for cooperation. The standard risky choice experiments or inter-temporal choice experiments are used as a measure of risk aversion and temporal discounting, respectively.²⁴ To the extent that they are reliable and refined enough, the use of exhibits as a measure of behavioral characteristics may provide important information about the heterogeneity of the subject pool which, in turn, may explain the heterogeneity of the treatment effect of RFEs. This is analogous to the use of *ex ante* surveys of household characteristics in Dupas (2014), but with a more explicit focus on unobservable characteristics discernible by experiments. It can also reveal relevant cross-country, cross-regional or cross-sectorial behavioral differences if two or more such experiments (type F1 in our Table 1) with representative samples show systematic variations (e.g. Rustagi, Engel, & Kosfeld, 2010), informing the relation between institutions and preferences, which is still not well understood (Cardenas & Carpenter, 2008). We will discuss a study by Jakiela and Miguel (2015) that illustrates such an experiment in the next sub-section.

In sum, experiments as exhibits look for the process-mechanisms underlying phenomena by systematically varying the procedure-mechanisms (experimental designs). In addition, certain established and refined exhibits can be used as a measure of behavioral characteristics, which can be used to measure the correlation between the heterogeneity of treatment effects of RFEs and the

variation of those characteristics, in addition to other demographic characteristics. As we have shown, the similar approaches can be observed in what we consider some of the best practices in RFEs we highlighted from our case study. This suggests that our methodological analysis of economic experiments, following Bardsley et al. (2010), captures them well descriptively, and also provides them with methodological justification, as well as promising future directions. In contrast, Banerjee, Chassang, and Snowberg's (2017) *structured speculation*, which aims to guide the subjective aspect of the experimenter's reasoning in a decision-theoretic framework, does not describe those approaches in RFEs well, nor does it provide as rich and detailed a guidance as the exhibit analysis of experimental economics.

3.2. Behavioral economics as sources of models of choice

Another set of resources that RFEs can mobilize more systematically are behavioral economics, which have been developing a range of models of how beliefs and preferences produce choice behavior. We argue that this tradition complements RFEs.

The literature called 'behavioral development economics' (Cardenas & Carpenter, 2008; Kremer, Rao, & Schilbach, 2019) specifically focuses on *getting behavioral hypotheses right* as a precondition for the effective reduction of poverty. Cardenas and Carpenter (2008), for example, review findings from (mostly) field experiments with subjects from developing countries (and 'small-scale societies'),²⁵ focusing on three types of preferences, that is, (*pro*-)social, time, and risk preferences. These experiments use standard experimental designs such as binary lottery choice task and delayed choice task, and economic games such as Public Goods, Dictator, Ultimatum, and Trust games to measure the beliefs and preferences of the poor, using as a benchmark the well-established findings from these experiments conducted in developed countries (F0 of Table 1 on p. 15).

What is distinct about behavioral development economics is its continuity with behavioral economics that have been developing various alternative models of choice over the last 40 years or so. In the case of risk preferences, for example, Expected Utility Theory assumes that an individual's risk attitude is constant over the entire domain of income, while Prospect Theory (Kahneman & Tversky, 1979) allows it to depend on the reference point. It also allows subjective probability to be weighted, or equivalently nonlinearly transformed. In the case of time preferences, the standard model assumes that the future outcomes are discounted exponentially, while an alternative model assumes quasi-hyperbolic discounting, implying dynamic choice inconsistency (Laibson, 1997). Models of social preferences incorporate agents' non-monetary preferences such as concerns for fairness and equity (Fehr & Schmidt, 1999), as well as norm-sanctioned conditional preferences (Bicchieri, 2006). These theories of choice from behavioral economics can offer the explicit hypotheses concerning the processes underlying the behavior of the poor, which can be tested and applied in specific contexts such as the adoption of preventive health products.

Some advocates of RFEs have been skeptical about the role of behavioral economics in development economics. This is precisely because these alternative models of choice proliferated; for instance, Banerjee (2005) describes the 'behavioral economics revolution' in this skeptical light, suggesting that it would be more appealing to bypass the debates and focus on what works (See quote in endnote 3). However, as we have seen in the last subsection, the recent practices of RFEs rely on a lot of behavioral models (anchoring, loss-aversion, learning, sunk-cost fallacy, etc. see Kremer et al. (2019) for a review) and even sometimes actively investigate these mechanisms underlying the treatment effects. Moreover, construing experiments as exhibits, as we did in the last subsection, enables us to see that the proliferation of behavioral models represents progress rather than an obstacle: it shows not so much the failure of conversion to one true theory of everyone's behavior in every situation as the success of reliable operational isolation of different phenomena. What we observe is that 'holding back from theory,' despite its manifestation in the advocacy of machine learning modeling and structured speculation, is becoming increasingly violated in practice, as why people behaved as they did is explicitly modeled and sometimes investigated. If this is the case,

then it would be more effective to move from assuming or speculating about behavioral hypotheses in the model estimation to directly investigating them in the target population of interest to support the choice of particular behavioral assumptions. In the next subsection we review a pair of examples to illustrate how this can be done in practice, showing that our proposal is consistent with and helps to understand some of the recent developments in the field.

3.3. Unpacking the black box by hybrid field-experimental design

Since the publication of Cardenas and Carpenter (2008), interests in behavioral foundations have gradually grown in development economics. Viceisza (2016) provides a useful overview of recent 'lab-like field experiments' (LFEs henceforth) conducted in developing countries, which corresponds to F1 and F2 in our Table 1, or 'extra-lab experiments' (Charness et al., 2013) in our Table 1.²⁶ Viceisza (2016, pp. 836–837) identifies four distinct purposes of LFEs: (1) testing theory; (2) eliciting behavioral characteristics (preferences, beliefs, and social norms); (3) unpacking the black box, or the heterogeneity of treatment effects of RFEs; and (4) methodological advances. Since we have already discussed the importance of the first two purposes in the last subsection, we illustrate some experiments with the third purpose that directly concerns the use of other types of field experiments to unpack the black-box of RFEs.²⁷

Studies that combine RFEs with LFEs have recently become popular (Viceisza, 2016, Section 2.3). Here we highlight an LFE that exploits an existing RFE, and another that nests an LFE inside its own RFE. Both are motivated to understand the processes through which interventions on human capital (primary education and business training, respectively) mediate or are mediated by preferences.

Jakiela and Miguel (2015) studied causal impacts of human capital on respect for others' earned 'property rights'. They conducted a modified Dictator game with the students who participated in Kremer, Miguel, and Thornton' (2009) RFE, which had randomly assigned primary schools in Busia District, western Kenya, into the treatment group (receiving the scholarship program) and the control group (not receiving it). This RFE found that the program improved the treatment group's standard academic test scores by 0.2–0.3 standard deviation, relative to the control group. Exploiting these randomized samples, Jakiela and Miguel (2015) investigated whether human capital (in this case increased academic performance) nurtures a particular type of reciprocal social preferences, namely respect for others' earned income. To operationalize this dependent variable, they modified the Dictator game such that the dictator divides the dictatee's earnings from an effortful task, instead of her own, as in the popular protocol. Jakiela and Miguel (2015) found that (i) the students from the RFE's treatment group have a higher level of academic performance (as measured in school exit exams), and that (ii) those students (in the dictator's role) allocate significantly more to the other. Jakiela and Miguel (2015) also elicited students' beliefs (in the dictatee's role) about the allocation by the dictator, and rejected the alternative hypothesis that the increase in the dictator's allocation is due to the change in beliefs rather than preferences. This study thus provides an understanding of the process, through which school education may support the institution of property rights in development contexts. As such, this study thereby contributes to the larger body of existing literature on modernization theory studying the impact of formal education on social values and norms that support development. This study is innovative (i) because it actively exploits the existing RFE as part of the study design, and (ii) because it is informed by experimental and behavioral economics both methodologically (e.g. building on the variations of the Dictator game, use of strategy methods, elicitation of beliefs) and theoretically (use of models of social preferences and social norms). Berge, Bjorvatn, and Tungodden (2015) studied effects of human capital (the entrepreneur's education, business skills and mind-set) and financial capital (in particular business grants) on microenterprise performance in Tanzania. Their main design is an RFE with two waves of interventions. First, the microentrepreneurs, who are registered users of Promotion of Rural Initiatives and Development Enterprise (PRIDE) – the largest microfinance institution in Tanzania – completed a baseline survey (644) and were randomly assigned to the treatment group (half-year weekly business

training: 319) and the control group (no training: 325). Second, both groups received another random assignment into the business grant comparable to the cost of the training (126 for each group), and no grant. The follow-up survey collected various kinds of data, including the main dependent variable of (self-reports on) business profit. The main finding from the RFE is that the training increases profit by 20 % only for male entrepreneurs, while the grant has no impact on both genders. Why does the human capital intervention improve the business performance of male but not female entrepreneurs? The question becomes even more puzzling when one notices that the knowledge (measured in a post-test) increases in both genders after the training. The clue comes from an incentivized LFE, which the experimenters included in the main RFE to elicit indicators of a business-related mind-set: risk and time preferences, as well as willingness to compete, and confidence.²⁸ Berge et al. (2015) found (i) no gender nor treatment effect on time preference; (ii) the closing gender differences in risk preference and confidence were due to the training;²⁹ but (iii) a stark contrast in male and female willingness to compete: males are more willing to compete than females, which the training did not change. Berge et al. (2015) thus infer that this particular mind-set – willingness to compete – mediates the training, improving the business performance of males but not of females. The authors consider female entrepreneurs' relative unwillingness to compete as a relevant 'internal' constraint, which is as important as the external constraints such as gender inequalities in household responsibilities and decision-making power.³⁰

These hybrid approaches – conducting an LFE on the subjects of an existing RFE or designing both at the same time – have the benefit of experimental control, uncovering the processes underlying the heterogeneity of the treatment effects utilizing well-established experimental paradigms and behavioral hypotheses. Behavioral hypotheses can be directly tested with the same population in question, and then included in a structural modeling of the data-generating process of an RFE. This is a viable alternative or complement to machine learning modeling and structured speculations, utilizing methodological and theoretical resources that already exist in other fields of economics. These experiments are still rare in development contexts, but we propose that the hybridization of RFEs and LFEs, as well as other types of field experiments and non-experimental studies, should be more consciously conducted to base policy recommendations on more externally valid inferences (cf. Deaton, 2010; Falk & Heckman, 2009).

4. Conclusion

Our detailed examination of Dupas and her colleagues' experiments allowed us to illustrate J-PAL's research program in development economics as well as some of the key theoretical and methodological issues at stake when using RFEs in development economics. Dupas's work highlights the vivid theoretical debate on the role of aid dividing contemporary development economics. Scrutinizing Dupas's work advances the issue of external validity in evidence-based development economics, in at least two ways. First, we attempted to bring together (from philosophy of economics, development economics, and experimental economics) the problems with external validity in development economics. In so doing, we showed that J-PAL's researchers are well aware of those problems and even propose answers, which so far have not been evaluated by the critics. Second, in bringing together the J-PAL research program (through Dupas's work), the criticism it faces and its methodological reactions, we identified a hybridization of RFEs and LFEs as a promising methodological approach that would improve the external validity of inferences from the RFEs to policy. The two answers of J-PAL's researchers for improving external validity – the use of machine learning modeling and the development of structured speculation – are a refinement within the tendency to hold back from theory, and as such has blind spots, namely experimental and behavioral economics that are increasingly addressing two overlapping issues of development. First, experimental economics offers a way to see field experiments, including RFEs, as a part of systematic effort to search for empirical regularities and their refinement and explanations. Second, behavioral economics offers a range of behavioral hypotheses that can be tested and estimated in hybrid field experiments.

We do not deny the epistemic benefits of the refinement inside the RFE paradigm as such, but we suggest that methodological low-hanging fruit can be reaped by the marriage of the sub-fields. Our unique methodological contribution is to highlight some of the rationales and the existing seeds of such a marriage, when practicing economists in different experimental paradigms naturally tend to compete with each other, trying to advance their own paradigms. This competition may facilitate methodological innovations in the short-run but it can slow down the progress of the field as a whole in the long-run, especially after each methodological paradigm has matured to a certain level. Our integrative approach is also contrasted to the existing philosophical discussions, which tend to critically analyze randomized field experiments in isolation; we have instead brought together the practices of different sub-fields of economics and shown how they can complement each other.

Coming back to the debate around the aid, our detailed analysis of Dupas's work identified some possible key mechanisms underlying the effect of free and subsidized bednets. A set of mechanisms including the credit constraints and present-bias tend to prevent people from buying bednets. Another set including learning and positive spill-over effects tend to encourage people to buy them, once they are adopted. How many of these mechanisms are present and how they are distributed in a particular population of interest can be investigated by systematically using methods and models from experimental and behavioral economics, together with RFEs. This will ultimately support the Kenyan policy maker's decision whether, when, and how to subsidize bednets.

Notes

1. For a discussion on the gold standard nature of RFEs, see for example Cartwright (2007), Cartwright and Hardie (2012) and Duflo (2004).
2. This follows exactly the same trend that first appeared in medicine with the rise of evidence-based medicine, and then more broadly in social science with the movement of evidence-based policy. See, for example, Cartwright and Hardie (2012) for an analysis of such a movement.
3. For Instance, Banerjee (2005, p. 4343) claims:

The fallout of the behavioral economics revolution in economics is that we are no longer particularly sure of what the right theory ought to look like, especially inasmuch as decision problems are concerned. In particular, we are no longer sure in the presumption that utility functions and cost functions are somehow more stable and more universal than behavioral rules. (Duflo, 2009)

makes a similar claim.

4. Although Sachs and Easterly are the two main figures of this popularized debate, other economists have claimed similar positions. For instance, Singer (2009) and Collier (2007) obviously share the position of Sachs, while Moyo (2009) sides with Easterly. Since J-PAL's researchers focus on the Sachs-Easterly debate, we follow this in our arguments.
5. 'Whom should we believe? Those who tell us that aid can solve the problem? Or those who say that it makes things worse? The debate cannot be solved in the abstract: We need evidence' (Banerjee & Duflo, 2011, p. 4).
6. It is instructive to see a similar 'forget about theory and collect data' idea resonate in the popular discourse on big data science (e.g. Anderson, 2008). For critical assessment of this idea in climate science, see Faghmous and Kumar (2014). We thank Miles MacLeod for pointing out this similarity.
7. Another type of answer focuses on the statistical aspects of RFEs in order to unpack the distribution of the treatment effect. See for example Athey and Imbens (2017) for the promotion of RFEs and Bareinboim, Lee, Honavar, and Pearl (2013) for the promotion of data fusion. Since our concern about external validity of RFEs in development economics is more general than a purely statistical one, we do not discuss this type of answer here.
8. We use the term external validity in the remainder of the paper to refer to the validity of inferences, not experimental design or results. See Jiménez-Buedo (2011) for a review of a range of things external validity can be about in experimental social science.
9. Deaton (2010) takes the example of Keynesian investment. He shows that in a Keynesian model, investment is both part of the income explanation as well as the consumption, which one would thus need to theoretically explain. The Keynesian explanations relate on the entrepreneurs' optimism or pessimism.
10. Deaton (2019) insists on that last point as well as the ethical concerns that RFEs in development economics encounter. Barrett and Carter (2010) also scrutinize this point, referring to a driving licence RFE in India.
11. In 2001, Kenya's health ministry led an important program to fight malaria by distributing 3 million ITNs for free. The Global fund fighting HIV, malaria, and tuberculosis financed this Kenyan program up to 17 million dollars.

After this program, Guyatt, Ochola, and Snow (2002) assess whether the ITNs reached their destinations and show that they did. In other words, neither waste nor corruption was detected, which were the main concerns of the massive aid opponents.

12. Dupas (2014) assesses this willingness to pay with a new health product: *long-lasting insecticide-treated nets*, or LLIN, in which the insecticide needs to be restored every four or five years.
13. Kremer et al. (2019) offers a large overview of such a puzzle. They enlarge it to several domains, such as savings, health, technology adoption, labor,... They refer to this enlarged puzzle as the Euler's puzzle.
14. For instance, Easterly (2006) claims that bednets are often used as fishing nets or wedding veils.
15. The title of Duflo's presentation at the NBER (Duflo, 2018), 'Machinistas meet randomistas', is a direct response to Deaton (2010), who with Ravallion (2009) coined the term randomistas to refer to J-PAL's researchers.
16. This approach also encourages re-randomization. Randomization might not fully balance the sample into the two groups, therefore re-randomizing might improve such balance. However, re-randomizing might also threaten the internal validity of the experimental result. That is why Banerjee, Chassang, and Snowberg (2017) define a threshold for such re-randomization.
17. 'Proposition 3 formalizes the natural intuition that external policy advice is unavoidably subjective. This does not mean that it needs to be informed by experimental evidence, rather, judgment will unavoidably color it' (Banerjee, Chassang, & Snowberg, 2017, p. 25).
18. Furthermore, Dupas and Robinson (2013b) have another section in which they attempt to 'rule out alternative explanations'. The three alternative mechanisms all concern the role of a Rotative Savings and Credit Associations (ROSCAs). In this peer-to-peer credit system, members make cyclical contributions of the same amount of money to a common fund (called the 'pot') at each meeting, and a lump sum of the pot of money is given to one member in each cycle. First, Dupas and Robinson (2013b) claims that the ROSCA might have been perceived as a reminder to save. Second, seeing people save at the ROSCA might give individuals reasons to save as well. Third, Dupas and Robinson (2013b) question the future of the ROSCA, since their device implies less commitment than the one of the ROSCA. However, the ROSCA also plays a social and even moral role in the community, which is implicit but not questioned by Dupas and Robinson (2013b).
19. When Mwai Kibaki was elected president of Kenya in December 2007, his election encountered strong opposition. From January to February 2008, Kenya faced social violence. As a result, Kenyans' incomes as well as expenditures dropped. As well-discussed in the literature (e.g. Barro, 1991), politically unstable countries tend to have lower growth than stable ones. The main question, however, is which causes which: does slow growth lead to political instability or vice versa? To answer this question, Dupas and Robinson (2010) use three data sets coming from three different experiments. The first data set is the one from Dupas and Robinson (2013a). The second one is from an experiment in Kenya conducted by Robinson and Yeh (2009) on transactional sex and risky behavior. The last one is from an experiment by Robinson and other colleagues at the J-PAL: Kremer, Lee, Robinson, and Rostapshova (2016) on local shop owners in Kenya.
20. Note however that z_p is not necessarily an environment to which the experimenter recommends a policy intervention, as in Banerjee, Chassang, Montero, et al. (2017), but rather an environment to which she applies an inference of a causal mechanism – similar to that at work in the lab. We emphasize this point because the recommendation of a similar policy requires much more information than the causal similarity.
21. Harrison and List (2004) discuss other types of experiments such as thought experiments, social experiments and natural experiments, which we shall not discuss here. Another thing to note is that, logically, there would be more types of experiments than five. But being exhaustive in this respect is not necessary for our purposes here.
22. Robustness in this experimental/operational sense is related to but distinct from the robustness much discussed in the philosophical literature since Wimsatt (2007), focusing on robust theoretical modeling as triangulation through independent theoretical derivations.
23. There is another closely related hypothesis, *the spillover effect*, which is important in Dupas (2014), but we will not discuss this to make our exposition simple.
24. Depending on the experimental design and the specification of the estimation model, the experimenters can also measure the non-standard behavior such as loss-aversion and present-biasedness, and their importance relative to other more 'rational' behavior. In general, to measure preferences is to identify and estimate parameters associated with those preferences in a model (or models) of those preferences.
25. That is, F1-type experiments in Table 1 on p. 15. Although Cardenas and Carpenter (2008) use the term 'field lab' to characterize those experiments, we will stick to the terminology of Table 1 here to avoid confusion.
26. Cardenas and Carpenter (2008) call them 'field labs'. Gneezy and Imas (2017) calls the use of this type of field experiments the 'lab-in-the-field' methodology.
27. The last purpose concerns the validity of different elicitation and other experimental methods. Although they are also relevant for external validity of experimental results in general, we will not discuss this issue in order to focus on the complementarity of RFEs and LFEs.
28. The first two types of preferences were measured in common choice tasks. The latter two types were elicited as a willingness to participate and confidence in an incentivized knowledge quiz competition.
29. Male entrepreneurs have become more risk averse and female entrepreneurs have become more confident.

30. This contrast between internal and external constraints does not have to imply that the women are inherently unwilling to compete relative to the men. It is equally likely that this mind-set is an endogenous response to the gender structure in the society.

Acknowledgments

We thank two anonymous reviewers and Pascaline Dupas for their insightful and detailed comments on earlier versions of this paper. We also benefited from the comments from the audiences at a TINT brown bag seminar and the meetings of the following societies: SPSP (Society for Philosophy of Science in Practice), INEM (International Network for Economic Method), and ESHET (European Society for the History of Economic Thought). In particular we thank Glenn Harrison, Marriá Jiménez Buedo for valuable comments. Sofia Blanco Sequeiros helped the project as a research assistant. Lisa Muszynski improved the language and style of the paper. The usual caveats about any shortcomings of the work being our own apply.

Disclosure statement

Q2 No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Academy of Finland in the context of the project “Model-building across disciplinary boundaries: Economics, ecology, and psychology” (No. 294545).

Notes on contributors

Judith Favereau is an associate professor in philosophy of economics and history of economics thought at the pluridisciplinary laboratory TRIANGLE in the University Lyon 2. She is also affiliated with the Centre for Philosophy of Social Science (TINT). Her topics of interest are development economics, experimental economics, philosophy of economics, and evidence-based policy. Her research focuses on how development economics, experimental economics, and evidence-based policy interact together to fight poverty.

Michiru Nagatsu is an associate professor at the Helsinki Institute of Sustainability Science, and Practical Philosophy, the University of Helsinki. He runs *Economics and Philosophy Lab* and *HELSUS Methodology Lab*. His research uses a range of empirical approaches – including experimental philosophy, collaborations with scientists, interviews, integrated history and philosophy of science – to study conceptual and methodological questions in the philosophy of science.

References

- Acemoglu, D. (2010). Theory, general equilibrium, and political economy in development economics. *Journal of Economic Perspectives*, 24(3), 17–32.
- Q3** Anderson, C. (2008). *The end of theory: The data deluge makes the scientific method obsolete*.
- ▲ Angrist, J. D., & Pischke, J. -S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2), 3–30.
- Ashraf, N., Berry, J., & Shapiro, J. M. (2010). Can higher prices stimulate product use? Evidence from a field experiment in zambia. *American Economic Review*, 100(5), 2383–2413.
- Athey, S., & Imbens, G. (2017). The econometrics of randomized experiments. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of field experiments*, volume 1 of *Handbook of Economic Field Experiments* (chapter 3, pp. 73–140). North-Holland.
- Banerjee, A. V. (2005). New development economics’ and the challenge to theory. *Economic and Political Weekly*, 40(40), 4340–4344.
- Banerjee, A. V., Chassang, S., Montero, S., & Snowberg, E. (2017). *A theory of experimenters* (Working Paper 23867). National Bureau of Economic Research.
- Banerjee, A. V., Chassang, S., & Snowberg, E. (2017). Decision theoretic approaches to experiment design and external validity. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of field experiments*, volume 1 of *Handbook of economic field experiments* (chapter 4, pp. 141–174). North-Holland.
- Q4** Banerjee, A. V., & Duflo, E. (2011). Poor economics: A radical rethinking of the way to fight global poverty. Public Affairs.
- ▲ Bardsley, N., Cubitt, R., Loomes, G., Moffat, P., Starmer, C., & Sugden, R. (2010). *Experimental economics: Rethinking the rules*. Princeton University Press.
- Q5**

- Bareinboim, E., Lee, S., Honavar, V., & Pearl, J. (2013). Transportability from multiple environments with limited experiments. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* 26 (pp. 136–144). Curran Associates, Inc.
- Barrett, C. B., & Carter, M. R. (2010). The power and pitfalls of experiments in development economics: Some non-random reflections. *Applied Economic Perspectives and Policy*, 32(4), 515–548.
- Barro, R. J. (1991). Economic growth in a cross section of countries. *The Quarterly Journal of Economics*, 106(2), 407–443.
- Basu, K. (2014). Randomisation, causality and the role of reasoned intuition. *Oxford Development Studies*, 42(4), 455–472.
- Berge, L. I. O., Bjorvatn, K., & Tungodden, B. (2015). Human and financial capital for microenterprise development: Evidence from a field and lab experiment. *Management Science*, 61(4), 707–722.
- Bicchieri, C. (2006). *The grammar of society*. Cambridge: Cambridge University Press.
- Cardenas, J. C., & Carpenter, J. (2008). Behavioural development economics: Lessons from field labs in the developing world. *The Journal of Development Studies*, 44(3), 311–338.
- Carpenter, J. P., Harrison, G. W., & List, J. A. (Eds.). (2005). *Field experiments in economics*. Research in Experimental Economics. Oxford: Elsevier-JAI.
- Cartwright, N. (2007). Are rcts the gold standard? *BioSocieties*, 2(1), 11–20.
- Cartwright, N. (2009). What is this thing called efficacy? In C. Mantzavinos (Ed.), *Philosophy of the social sciences: Philosophical theory and scientific practice* (chapter 7, pp. 185–206). Cambridge: Cambridge University Press.
- Cartwright, N. (2010). What are randomised controlled trials good for? *Philosophical Studies*, 147(1), 59.
- Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford: Oxford University Press.
- Charness, G., Gneezy, U., & Kuhn, M. A. (2013). Experimental methods: Extra-laboratory experiments — extending the reach of experimental economics. *Journal of Economic Behavior & Organization*, 91(0), 93–100.
- Chassang, S., Padró i Miquel, G., & Snowberg, E. (2012). Selective trials: A principal-agent approach to randomized controlled experiments. *American Economic Review*, 102(4), 1279–1309.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Dufo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68.
- Cohen, J., & Dupas, P. (2010). Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment. *The Quarterly Journal of Economics*, 125(1), 1–45.
- Collier, P. (2007). *The bottom billion: Why the poorest countries are failing and what can be done about it*. Oxford University Press.
- Crépon, B., Devoto, F., Dufo, E., & Parienté, W. (2015). Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in Morocco. *American Economic Journal: Applied Economics*, 7(1), 123–150.
- Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *Review of Economics and Statistics*, 90(3), 389–405.
- Davis, J. (2013). Economics imperialism under the impact of psychology: The case of behavioral development economics. *Economia*, 3(1), 119–138.
- Davis, J. M., & Heller, S. B. (2017). *Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs* (Working Paper 23443). National Bureau of Economic Research.
- Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48(2), 424–55.
- Deaton, A. (2019). Randomization in the tropics revisited: A theme and eleven variations. In I. G. Florent Bédécarrats & R. François Roubaud (Eds.), *Randomized controlled trials in the field of development: A critical perspective*. Oxford University Press. Forthcoming.
- Deaton, A., & Cartwright, N. (2018a). Reflections on randomized control trials. *Social Science & Medicine*, 210, 86–90. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue.
- Deaton, A., & Cartwright, N. (2018b). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2–21. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue.
- Dizon-Ross, R., Dupas, P., & Robinson, J. (2017). Governance and the effectiveness of public health subsidies: Evidence from Ghana, Kenya and Uganda. *Journal of Public Economics*, 156, 150–169.
- Dufo, E. (2004). Scaling up and evaluation. In F. Bourguignon & B. Pleskovic (Eds.), *Annual world bank conference on development economics* (pp. 341–369). The World Bank.
- Dufo, E. (2009). Experiments, science, and the fight against poverty. In *Inaugural lecture at College de France 1*.
- Dufo, E. (2018). Machinistas meet randomistas: Useful ML tools for empirical researchers.
- Dupas, P. (2009). What matters (and what does not) in households' decision to invest in malaria prevention? *American Economic Review*, 99(2), 224–230.
- Dupas, P. (2011). Health behavior in developing countries. *Annual Review of Economics*, 3(1), 425–449.
- Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica*, 82(1), 197–228.
- Dupas, P., & Robinson, J. (2010). Coping with political instability: Micro evidence from kenya's 2007 election crisis. *American Economic Review*, 100(2), 120–124.
- Dupas, P., & Robinson, J. (2012). The (hidden) costs of political instability: Evidence from kenya's 2007 election crisis. *Journal of Development Economics*, 99(2), 314–329.

- Dupas, P., & Robinson, J. (2013a). Savings constraints and microenterprise development: Evidence from a field experiment in Kenya. *American Economic Journal: Applied Economics*, 5(1), 163–192.
- Dupas, P., & Robinson, J. (2013b). Why don't the poor save more? Evidence from health savings experiments. *American Economic Review*, 103(4), 1138–1171.
- Easterly, W. (2006). *The white man's burden: Why the West's efforts to aid the rest have done so much ill and so little good*. Penguin.
- 955 Q7 Faghmous, J. H., & Kumar, V. (2014). A big data guide to understanding climate change: The case for theory-guided data science. *Big Data*, 2(3), 155–163.
- Falk, A., & Heckman, J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science (New York, N.Y.)*, 326(5952), 535–538.
- 960 Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817–868.
- Gneezy, U., & Imas, A. (2017). Lab in the field: Measuring preferences in the wild. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of field experiments*, volume 1 of *Handbook of Economic Field Experiments* (chapter 10, pp. 439 – 464). North-Holland.
- Guala, F., & Mittone, L. (2005). Experiments in economics: External validity and the robustness of phenomena. *Journal of Economic Methodology*, 12(4), 495–515.
- 965 Guala, F., & Mittone, L. (2010). Paradigmatic experiments: The dictator game. *The Journal of Socio-Economics*, 39(5), 578–584.
- Guillin, V. (2013). De quoi les essais contrôlés randomisés sont-ils capables? évaluation, mécanismes et capacités en sciences sociales. *Cahiers philosophiques*, 133(2), 79–102.
- Guyatt, H. L., Ochola, S. A., & Snow, R. W. (2002). Too poor to pay: Charging for insecticide-treated bednets in highland Kenya. *Tropical Medicine & International Health*, 7(10), 846–850.
- 970 Harrison, G. W. (2011). Randomisation and its discontents. *Journal of African Economies*, 20(4), 626–652.
- Harrison, G. W. (2013). Field experiments and methodological intolerance. *Journal of Economic Methodology*, 20(2), 103–117.
- Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4), 1009–1055.
- Heckman, J. J. (1992). Randomization and social policy evaluation. In C. F. Manski & I. Garfinkel (Eds.), *Evaluating welfare and training programs* (chapter 5, pp. 201–230). Harvard University Press.
- 975 Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature*, 48(2), 356–98.
- Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66(5), 1017–1098.
- Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, 9(2), 85–110.
- 980 Heckman, J. J., Smith, J. A., & Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4), 487–535.
- Hussam, R., Rigol, N., & Roth, B. (2016). Targeting high ability entrepreneurs using community information: Mechanism design in the field. Retrieved from <http://economics.mit.edu/grad/benroth>.
- Jakiela, P., & Miguel, E. (2015). You've earned it: Estimating the impact of human capital on social preferences. *Experimental Economics*, 18(3), 385–407.
- 985 Jiménez-Buendo, M. (2011). Conceptual tools for assessing experiments: Some well-entrenched confusions regarding the internal/external validity distinction. *Journal of Economic Methodology*, 18(3), 271–282.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk*. *Econometrica*, 47(2), 263–292.
- Kőszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences*. *The Quarterly Journal of Economics*, 121(4), 1133–1165.
- 990 Kremer, M., Lee, J., Robinson, J., & Rostapshova, O. (2016). *Rates of return, optimization failures, and behavioral biases: Evidence from Kenyan retail shops* (Working paper).
- Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics*, 91(3), 437–456.
- Kremer, M., Rao, G., & Schilbach, F. (2019). Behavioral development economics. In B. D. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of behavioral economics – foundations and applications* (Vol. 2, chapter 5, pp. 345–458). Amsterdam: Elsevier.
- 995 Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2), 443–478.
- Leamer, E. E. (2010). Tantalus on the road to asymptopia. *Journal of Economic Perspectives*, 24(2), 31–46.
- Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1), 159–217.
- Q8 Moyo, D. (2009). *Dead aid: Why aid is not working and how there is a better way for Africa*. Macmillan.
- Q9 Ravallion, M. (2009). Should the randomistas rule? *The Economists' Voice*, 6(2).
- 1000 Robinson, J., & Yeh, E. (2009). *Transactional sex as a response to risk in western Kenya* (Policy Research working paper WPS 4857). Washington, DC: World Bank.

Rodrik, D. (2008). *The new development economics: We shall experiment, but how shall we learn?* (HKS Working Paper No. RWP08-055). Revised, October 2008.

Rustagi, D., Engel, S., & Kosfeld, M. (2010). Conditional cooperation and costly monitoring explain success in forest commons management. *Science (New York, N.Y.)*, 330(6006), 961–965.

Q10 Sachs, J. (2005). *The end of poverty: How we can make it happen in our lifetime*. Penguin.

1005 Singer, P. (2009). *The life you can save: How to do your part to end world poverty*. London: Random House.

Teira, D. (2013). Blinding and the non-interference assumption in medical and social trials. *Philosophy of the Social Sciences*, 43(3), 358–372.

Teira, D., & Reiss, J. (2013). *Causality, impartiality and evidence-based policy* (pp. 207–224). Dordrecht: Springer.

Viceisza, A. C. G. (2016). Creating a lab in the field: Economics experiments for policymaking. *Journal of Economic Surveys*, 30(5), 835–854.

1010 WHO (2009). Global tuberculosis control: Epidemiology, strategy, financing: WHO report 2009. World Health Organization.

Q11

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, MA: Harvard University Press.

1015

1020

1025

1030

1035

1040

1045

1050