Creating a Lab in the Field: Economics Experiments for Policymaking

Angelino C. G. Viceisza
Spelman College

August 26, 2013

Abstract

In this article I focus on the role that lab-like field experiments (LFEs), particularly when conducted in rural areas of developing countries, play in informing policymaking. Using specific examples, I identify four main purposes of LFEs: (1) to test theories or heuristic principles; (2) to identify and estimate parameters associated with characteristics; (3) to explore the structural nature of parameters derived from empirical methods including other types of experiments; and (4) to assess methodological difficulties associated with LFEs and how these can impact parameter estimates. I address the importance of generalizability for LFEs that are intended to inform policymaking and in the process, emphasize the complementary role between LFEs and other empirical methods, in particular other experiments. Finally, I discuss 19 basic principles and eight practical aspects to keep in mind when conducting LFEs; I also suggest three future directions. *JEL Codes:* C9, O1.

Keywords: lab-like field experiment, development, policymaking.

1 Introduction

The experimental methodology has gained extensive popularity in economics, particularly in recent decades (see for example discussions by Smith, 1987; Burtless, 1995; Heckman, 1995; Banerjee and Duflo, 2009; Levitt and List, 2009, and the references within). However, no fame goes without controversy. While experiments have been appreciated for their contributions to the research process (see former references as well as Falk and Heckman, 2009; Banerjee and Duflo, 2010; Imbens, 2010; Camerer, 2011; List, 2011), cautions have been placed on the parameters they (can) estimate as well as their generalizability (see for example Deaton, 2010; Heckman, 2010; Al-Ubaydli and List, 2012).

In this article, I focus on the role that lab-like field experiments (LFEs), particularly those conducted in rural developing country contexts, can play in the scientific and policymaking process. To be more concrete what we mean by LFEs, consider the taxonomy proposed by Harrison and List (2004), which covers the spectrum of experiments from the laboratory to the field. They describe conventional *laboratory* (lab) experiments as those that employ a standard subject pool of students, an abstract framing, and an imposed set of rules and classify *field* experiments into (1) artefactual field experiments (AFEs), which are the same as lab experiments except that they draw non-student participants from the field environment of interest; (2) framed field experiments (FFEs), which are the same as AFEs except that they have field context in the commodity, task, stakes, or information set; and (3) natural field experiments (NFEs), which are the same as FFEs in which subjects make decisions in their day-to-day environment, but different because subjects do not know that they are in an experiment. LFEs comprise AFEs and those FFEs that are identical to AFEs except for the task being framed in a field context. So, LFEs maintain the nature of lab experiments by having subjects perform laboratory tasks, but are field experiments in the sense that they draw subjects from and create a laboratory in the field.[1] In the terminology of Charness et al. (2013), I focus on the so-called "extra-laboratory experiments".

I focus exclusively on LFEs, particularly those conducted in developing country contexts, because of three primary reasons.

First, it seems that traditionally conventional lab experiments and field experiments that study behavior in naturally occurring settings (that is, certain FFEs such as RCTs and NFEs) have received more attention than LFEs. Consider two examples that illustrate this point. At the conceptual level, recent debates on 'generalizability' seem to have mainly focused on lab experiments on the one hand (see for example Levitt and List, 2007; Falk and Heckman, 2009; Camerer, 2011) and on RCTs on the other hand (see for example Banerjee and Duflo, 2009, 2010; Imbens, 2010; Deaton, 2010; Heckman, 2010).[2] At the more practical level, while how-to's for conducting lab experiments and RCTs exist (see for example Davis and Holt, 1993; Friedman and Sunder, 1994; Duflo et al., 2007), to my knowledge, no equivalent reference exists for LFEs, other than my own recent discussion (Viceisza, 2012).

Second, the use of LFEs to explore research questions, particularly in development contexts, seems to be growing relatively fast. As will become clearer throughout the article, LFEs are not only being used for their stand-alone contribution to the research process, but also because of their complementary role in informing the scientific and policymaking processes. In fact, this illustrates a broader point that has been made in the literature regarding the complementary role of different classes of experiments (see for example Camerer, 2011;

Al-Ubaydli and List, 2012). So, now seems like a good time to discuss some issues related to LFEs and formulate a forward-looking perspective.

Third, one could claim that LFEs conducted in developing countries, more so than when conducted in other contexts, are targeted towards informing policymaking. So, certain issues such as generalizability can be argued to be of greater concern than when dealing with a conventional lab experiment.

The remainder of the article is organized as follows. Next, I discuss how LFEs can contribute to the scientific and policymaking processes. While I support the discussion by means of examples, the article is not intended to be a complete review of the literature. The set of examples is selected to illustrate certain key arguments. I then address the extent to which generalizability (sometimes also referred to as external validity) is a concern for findings derived from LFEs. I continue with a discussion of 19 basic principles and eight practical aspects to keep in mind when conducting these types of experiments. I conclude by providing a forward-looking perspective on the direction in which the literature is/should be headed.

2 Contributions of lab-like field experiments

In this section, I discuss four non-mutually exclusive purposes for conducting LFEs. The discussion is based on a review of selected studies that report LFEs conducted in developing country contexts. As shall become evident, LFEs are not only useful because of their stand-alone contribution to the research and policymaking process, but also because of how they complement other types of approaches, in particular other types of experiments.

Table 1 provides an overview of some studies that report LFEs. Four primary, non-mutually exclusive purposes for conducting these types of experiments emerge:[3]

Purpose 1 To test explicitly developed models (theories) or heuristic principles.[4]

Purpose 2 To elicit, or more precisely, identify and estimate parameters associated with, characteristics that may have traditionally been considered 'unobservable' such as preferences, beliefs, and social norms.

Purpose 3 To explore the 'structural' nature of parameters derived from other types of empirical methods, particularly RCTs, NFEs, and LFEs conducted with purposes 1 and 2. Some parts of the development literature have also termed this 'to explore (behavioral) mechanisms underlying treatment effects' or 'to assess heterogeneous treatment effects'.

Purpose 4 To identify and resolve the methodological complexities associated with conducting LFEs and where relevant, assess how methodology impacts parameters obtained from LFEs.

Below, I elaborate on each of these purposes by first explaining their main features in the context of the all-causes model (Heckman, 2000; Al-Ubaydli and List, 2012). Then, I support this conceptual discussion by means of some of the studies in table 1.

In the all-causes model, Y is a random variable, denoted the dependent variable, whose realizations are in $S_Y \subseteq \mathbb{R}$; X is a random variable, denoted the explanatory variable of interest, whose realizations are in $S_X \subseteq \mathbb{R}$; and Z is a random vector, denoted the additional explanatory variables, whose realizations are in $S_Z \subseteq \mathbb{R}^k$. Further, Z, part of which may be unobservable, contains all the explanatory variables (apart from X) that have an impact on Y. Let (X, Y, Z) be related according to the function $f : S_X \times S_Z \to S_Y$. Then, each $(x, x', z) \in S_X \times S_X \times S_Z$ is denoted a causal triple. The causal effect of changing X from x to x' on Y given $Z = z$ is described by the function $g : S_X \times S_X \times S_Z \to \mathbb{R}$, where $g(x, x', z) = f(x', z) - f(x, z)$. I will call upon some of these components when discussing the different purposes of LFEs further below.

Before I continue, it is important to note, and thus I reiterate, that the above purposes of LFEs are not mutually exclusive. Many studies do not fall neatly into just one category. I have done my best to identify each study according to its main contribution. In addition, these purposes are not necessarily unique to LFEs, or even experiments more broadly. Much of what I discuss applies to experimental (including LFEs), quasi-, and non-experimental approaches alike.

2.1 Testing theory and heuristics

Three key features make LFEs a great tool for testing models and heuristics:

Feature 1 The ability to create a robust counterfactual.

Feature 2 The relative ease and low cost of conducting them.

Feature 3 The ability to collect outcome variables and unpack mechanisms at a relatively detailed and refined level.[5]

The first feature applies to any type of experiment. One of the primary reasons why experiments have gained so much popularity in recent decades is because they enable us to test 'comparative static' predictions by constructing a proper counterfactual scenario for establishing causal (treatment) effects. In light of the all-causes model, recall that in order for a researcher to get a magnitude for $g(x, x', z)$–that is, the causal (treatment) effect– she must observe $f(., z)$ both at x and x'. However, any given unit (say, an individual) is typically not observed at both $X = x$ *and* $X = x'$. By construction, if a unit is observed under condition $X = x$, it is not under condition $X = x'$. This is the so-called 'problem of a missing counterfactual'. LFEs, and (controlled) experiments more generally, represent the most convincing method of creating this counterfactual, since they tend to do so via randomization. In other words, LFEs tend to construct 'comparable' comparison groups by assigning random subsamples of a population to x and x'. So, the researcher will typically sample Y repeatedly at $(X, Z) = (x, z)$ and $(X, Z) = (x', z)$ and use this to obtain an estimate of $g(x, x', z)$. *This* is the estimate of the causal (treatment) effect in which we are interested.[6]

Examples of this causal effect can be found in any of the purpose 1 studies cited in table 1. While I discuss three of these studies in further detail below (notably, Hill and Viceisza, 2012; Giné et al., 2010; Attanasio et al., 2012), table 2 summarizes the primary dependent

variables Y, the primary independent variables X as well as the estimates of the causal effects $\hat{g}(x, x', z)$ for all of the purpose 1 studies. Not all of these estimates are treatment effects; nor are all of them obtained through a fully randomized design. Nonetheless, in all cases the LFE nature of the design enables the researcher to measure a dependent variable that can be combined with other experimental and/or non-experimental survey-based independent variables to get an estimate of the main (causal) effect.

The second feature is particularly relevant when we compare LFEs to experiments in which behavior is closer to the naturally occurring environment such as RCTs and NFEs. LFEs enable the researcher to assess the impact of changes in preferences, institutions, or policies in a fairly costless and easy manner. Specifically, there will be circumstances in which conducting an RCT or NFE is practically impossible or too costly. So–setting aside possible issues with generalizability, which we return to in section 3–an LFE can be an ideal way to proceed.

To make these claims more concrete, let us consider Hill and Viceisza (2012). We created a laboratory in rural Ethiopia to conduct a framed field experiment aimed at testing the seminal hypothesis that insurance induces farmers to take greater, yet profitable risks (along the lines of Sandmo, 1971). As summarized in table 2, the outcome (dependent) variable Y in our experiment was the purchase/application of bags fertilizer; the main treatment (independent) variable X was whether an individual farmer was insured ($X = x_2$, the so-called 'treated') or not ($X = x_1$, the so-called 'control/baseline'); and the estimate of the causal treatment effect, $\hat{g}(x_1, x_2, z)$, suggests that insurance has some positive effect on fertilizer purchases.

Why explore this research question using an LFE as opposed to for example an RCT or NFE? As Hill and I argue, an LFE provided the ideal opportunity to address this question. The literature on insurance for development has been quite active in recent years (see for example Hill and Torero, 2009, and the references within) and studies have shown that take up of innovative insurance products (such as weather-index based contracts) in more naturally occurring settings (such as RCTs and NFEs) has been relatively low (see for example Cole et al., 2009). As a result, using data from these settings to assess the impact of insurance on riskier, more profitable investments has been nearly impossible due to limited statistical power (a more recent exception is Cole et al., 2012). So, an LFE seemed like an ideal starting point since it enabled us to study lab experimental risk taking in the presence of mandated insurance on a random subset of subjects and as such, contribute to the literature and policy discussion at a time when other, more general, types of experiments could not.

To illustrate the third feature of LFEs, I call upon two additional studies: Giné et al. (2010) and Attanasio et al. (2012).

The first study illustrates the usefulness of LFEs to help unpack refined behavioral mechanisms. Given the relative ease and low cost of conducting these types of experiments, Giné et al. (2010) were able to conduct ten different treatments, that is $X = \{x_1, ..., x_{10}\}$. They created a laboratory in a large urban market in Lima, Perú, and conducted variants of LFEs to unpack microfinance mechanisms in a systematic way. Their baseline treatments were an individual liability (IL) and a joint liability (JL) microfinance contract and their main treatments were variants of the JL contract. The remaining treatments introduced dynamic incentives. This wide range of treatments allows them to test refined hypotheses on the behavioral aspects of microfinance arrangements.

They find that risk-taking broadly conforms to theoretical predictions, with dynamic incentives strongly reducing risk-taking even without group-based mechanisms. Group lending increases risk-taking, especially for risk-averse borrowers, but this is moderated when borrowers form their own groups. Group contracts benefit borrowers by creating implicit insurance against investment losses, but the costs are borne by other borrowers, especially the most risk averse. So, their findings have implications for the design of microfinance arrangements and suggest factors that policymakers and/or NGOs should take into account when designing these schemes, particularly if the risk averse are also more likely to be otherwise deprived.

The second study illustrates the usefulness of LFEs to collect refined outcome (dependent) variables of interest Y. Attanasio et al. (2012) conduct LFEs in 70 Colombian communities to investigate who pools risk with whom. They have subjects play a standard lottery choice risk game (à la Binswanger, 1980), which is then followed by a risk pooling game. In the latter game, subjects play the lottery choice risk game again (that is, for a second round), but prior to making their choices in private they get to form "sharing groups". Within sharing groups, winnings from this second round are pooled and shared equally. However, in their private meetings, after seeing the outcome of their gambles, each participant is given the option to withdraw from their sharing group, taking their own winnings with them, but forfeiting their share of the other members' winnings. They use the second round to create a dependent variable Y_{ij} that captures whether individual i chooses to pool risk with individual j. Furthermore, they use the lottery choice game and survey based social network data to construct the main independent variables of interest X–risk attitudes and the relationship between subjects respectively. These allow them to explore the main effects of interest.

They find that close friends and relatives group assortatively on risk attitudes and are more likely to join the same risk pooling group. Meanwhile, unfamiliar participants group less and rarely assort. These findings indicate that where there are advantages to grouping assortatively on risk attitudes, those advantages may be inaccessible when trust is absent or low. Notice that this study does not identify a treatment effect, as many experimental studies do. Their study shows two potential advantages of LFEs relative to more general types of experiments such as RCTs and NFEs.

First, collecting risk pooling data in a more naturally occurring environment would have entailed a substantially costlier approach. The process of day-to-day group formation for purposes of research has been illustrated by some microfinance RCTs and NFEs (see for example Feigenberg et al., 2012).

Second, since constructing proxies for risk preferences at a naturally occurring level is complex, LFEs can fill the void. In order for a researcher to use naturally occurring behavior to identify risk attitudes, she would have to design very refined experiment treatments. Naturally occurring behavior can be seen as a 'reduced form' outcome of a plethora of factors, among which are preferences and beliefs. As such, risk attitudes tend to be intertwined with other types of preferences such as time. In fact, I posit that the inability to pin down components of preferences exactly using decisions made at a more naturally occurring level is why an increasing number of researchers are combining RCTs/NFEs with LFEs. This may enable them to assess the behavioral mechanisms that underlie causal (treatment) effects more carefully and potentially test theory in a more refined manner. We return to this complementary role for LFEs when discussing purpose 3 in section 2.3.

2.2 Eliciting characteristics

As indicated in the previous section, LFEs have also been conducted to elicit characteristics that have traditionally been considered 'unobservable' such as preferences, beliefs, and social norms. A substantial number of LFEs have been conducted with this purpose in mind.

Table 1 distinguishes between three broad topics: (1) those conducted to elicit risk, time, and ambiguity preferences; (2) those conducted to elicit aspects of social capital (preferences, norms, trust and reciprocity); and (3) those conducted to understand gender differences in preferences and decisionmaking. Despite the distinct foci of these LFEs, the common thread that connects them–in the context of the all-causes model–is that they enable researchers to measure characteristics of interest, either as the main dependent (outcome) variable Y or as an independent variable X or Z, at a refined level that is not achievable by other more general types of experiments.

I illustrate this purpose of LFEs by means of four studies in table 1, Harrison et al. (2010), Mahajan and Tarozzi (2011), Hill et al. (2012), and Gneezy et al. (2009).

Harrison et al. (2010) review experimental evidence collected from risky (lottery) choice experiments conducted in Ethiopia, India and Uganda. They find that just over 50% of the sample behaves in accordance with expected utility theory (EUT) and that the rest subjectively weight probability according to prospect theory (PT). Their results show that inferences about risk aversion are robust to whichever model (EUT or PT) is adopted when estimated separately. However, when both models are allowed to explain portions of the data simultaneously, they infer risk aversion for subjects behaving according to EUT and risk-seekingness for subjects behaving according to PT.

This example not only illustrates how LFEs can be used to characterize people's risk attitudes, but it also demonstrates how these types of LFEs are tied to theory (recall purpose 1 and the claim that the purposes are not mutually exclusive). In fact, an increasing number of studies (see for example Akay et al., 2012; de Brauw and Eozenou, 2011; Tanaka et al., 2010) are using lottery choice procedures to characterize people's risk and ambiguity attitudes in developing country contexts–a tradition that started, as far as I know, with Binswanger (1980). In doing so, a researcher must typically establish an underlying model for decisionmaking under uncertainty. This immediately gives rise to the link.

Mahajan and Tarozzi (2011) use LFEs in a way that is likely to become more popular in the future, as complements to non-experimental structural approaches. They develop a dynamic discrete choice model with unobserved types and time varying utilities, and provide identification results for all time preference parameters. They overcome the previous problem of lack of identification reported in the literature by (1) adding more information in the form of elicited beliefs about state occurrences and elicited responses to (hypothetical) time preference questions (that is, the LFEs) and (2) designing a product appealing to particular types of agents and offering it for sale in a field intervention. So, the LFEs not only enable them to characterize agents' time preferences using a field experimental version of the structural approach proposed by Andersen et al. (2008), but they also enable them to identify the structural, dynamic discrete choice model.

They first show that the model is identified and then estimate it to test several hypotheses of interest. They find that while per-period utilities do vary across agent types, they are not substantively important in explaining outcomes in their sample. Second, they estimate that

approximately 40% of the population from which the sample is drawn are time consistent, while 50% are "naïve" inconsistent and the remaining 10% are "sophisticated" inconsistent. Further, they find that "sophisticated" agents are considerably more present-biased than "naïve" agents. This finding is possible because they show identification for separate hyperbolic parameters for each type. In particular, they find that "naïve" agents have a hyperbolic parameter close to 1 and that in a set of counterfactual simulations, "naïve" agent choices are similar to those made by consistent agents. Finally, they find that commitment products are not particularly appealing to "sophisticated" agents and that the purchase of these products is in fact higher among wealthier (and even "naïve") households.

In Hill et al. (2012), we use the dichotomous choice trust game and slight modifications therefrom to primarily explore the robustness of reciprocity in a sample drawn from rural Perú. Our baseline treatment ($X = x_1$) is a twice repeated trust game and our main treatment ($X = x_2$) is a modification in which (1) first movers are provided with *information* about the rate at which second movers chose not to share in a dichotomous choice dictator game played days prior and (2) certain first movers are provided with *personal information* that the second mover they are paired with chose not to share in the same dictator game. This information is provided in the first repetition of the game. We use this design not only to get a sense of baseline levels of trust and reciprocity, but also to assess whether the second mover's decision to reciprocate is influenced by the observed reciprocity of others.

In documenting the impact of the experimentally induced external shock to observed reciprocity, by means of the (personal) information shock to first movers, we show that small increases in non-reciprocal behavior can result in an unraveling of the norm of reciprocity. Survey data are used to explore mechanisms and the results are not found to be consistent with learning effects, suggesting that preferences may be changed by observing others deviating from a norm of reciprocity. Our results suggest that investing in encouraging trustworthy behavior can have large benefits in situations where individuals are observing each other's behavior, such as may be the case when exposing people to new (market) institutions or technologies.

Finally, Gneezy et al. (2009) implement an LFE across two types of societies–matrilineal and patriarchal–to explore the underpinnings of gender differences in competitive attitudes. They implement a controlled experiment with the Maasai in Tanzania and the Khasi in India. One unique aspect of these societies is that the Maasai represent a textbook example of a patriarchal society, whereas the Khasi are matrilineal. Similar to the widespread evidence drawn from experiments executed in Western cultures, they find that Maasai men opt to compete at roughly twice the rate of Maasai women. However, this result is reversed among the Khasi, where women choose the competitive environment more often than Khasi men, and even choose to compete weakly more often than Maasai men. Their results provide insights into the factors hypothesized to be determinants of the observed gender differences in selecting into competitive environments. Specifically, they suggest that nurture (social context) may be a stronger determinant of competitive inclination than nature.

2.3 Unpacking the black box

LFEs also serve a complementary role vis-à-vis other experimental, in particular RCTs and NFEs, as well as quasi- and non-experimental methods. Table 1 primarily focuses on studies

that combine LFEs with RCTs and NFEs since this approach has become increasingly popular in development economics. However, as the discussion of Mahajan and Tarozzi (2011) illustrated, LFEs can also be combined with other non-experimental empirical methods.[7]

To place this third purpose in the context of the all-causes model, suppose a researcher is interested in the estimate of a causal effect, that is $\hat{g}(x_1, x_2, z)$. She obtains this estimate by experimentally manipulating X between x_1 and x_2 and comparing the outcome variable Y across these experimental conditions. The researcher may also be interested in the extent to which $\hat{g}(x_1, x_2, z)$ varies with Z. That is, suppose she also observes $\hat{g}(x_1, x_2, z)$ at z_1 and z_2. Then, she may be interested in the estimate $\hat{d}(x_1, x_2, z_1, z_2) = \hat{g}(x_1, x_2, z_1) - \hat{g}(x_1, x_2, z_2)$. There are different methods to get at this estimate. In this section, I am interested in studies that measure Z by means of an LFE.

Table 1 summarizes some of these studies. Among them are ambiguity experiments (for example Engle Warnick et al., 2011), coordination experiments (for example Bernard et al., 2012), dictator experiments (for example Jakiela et al., 2012), public goods experiments (for example Barr et al., 2012), and risk experiments (for example Berge et al., 2011; Jamison and Karlan, 2011). Two additional studies, by Karlan (2005) and Ashraf et al. (2006), are mentioned in table 1. I discuss these in greater detail below due to their seminal nature.

Karlan (2005) combines decisions in among others a trust game with behavior in a group lending arrangement to assess the extent to which behavior in the game predicts naturally occurring behavior. Interestingly, he finds that second movers identified as trustworthier in the game are more likely to repay their loans one year later, as one might expect. On the other hand, first movers identified as more "trusting" save less and have higher repayment problems. This finding calls into question whether first-mover behavior in the game is driven by trust or merely a propensity to gamble.

Ashraf et al. (2006) combine hypothetical time preference experiments with a commitment savings product to assess the extent to which time (in)consistency predicts take-up of the product as well as the impact of the treatment on savings. They find that women with a lower discount rate for future relative to current trade-offs, and hence potentially a preference for commitment, are significantly more likely to open the commitment savings account. Furthermore, after twelve months, average savings balances increase by 81 percentage points for those clients assigned to the treatment group relative to those assigned to the control group. This treatment effect, however, does not seem to vary with time (in)consistency.

These two examples illustrate two main features of these types of LFEs:

Feature 1 To the extent that decisions in the LFE correlates with/predicts naturally occurring behavior in a comparable environment, such decisions can lead to construction of intermediary outcome measures (supplementary Y variables, if you will) that can be used to assess causal (treatment) effects.[8] This naturally leads to a discussion of the broader issue of generalizability, which we turn to in section 3.

Feature 2 Decisions in the LFE can lead to construction of measures Z that can be used to assess the heterogeneity of causal (treatment) effects. This enables the researcher to get a better sense of the behavioral mechanisms that underlie such effects.

Having discussed these two features, I would like to briefly address the title of this section. Recently, critiques have been raised about the behavioral mechanisms that underlie causal

(treatment) effects obtained by means of RCTs and NFEs, particularly when conducted in developing country contexts (for example, see the discussions by Deaton, 2010; Heckman, 2010; Imbens, 2010). Specifically, RCTs and NFEs have been criticized for being too reduced form ("black-boxy"), removed from theory, and as such, lacking generalizability. The types of LFEs discussed in this section potentially enable researchers to better understand the behavioral foundations underlying the causal (treatment) effects obtained from RCTs and NFEs. As such, they can enable a researcher to use RCTs and NFEs to better (1) test hypotheses derived from theory (for example, with regard to risk preferences) and (2) formulate policy recommendations.

For example, Ashraf et al. (2006) cannot only say something about the impact of the type of treatment on savings, but the LFE data also enable them to discuss whether this impact is uniform or differential across the treated. Specifically, they are able to say whether or not the more time-(in)consistent participants respond differently to treatment. To the extent that a well-developed model suggests certain Z measures that impact the main causal (treatment) effect and they have additional data for a sample that is beyond their treatment and control, they may be able to make out-of-sample predictions and as such, "generalize" certain findings. In the absence of this model and/or these additional data, they may still be able to "hypothesize" on the direction of these impacts.

There is a final, ex-ante role for these types of LFEs. Consider the following example. Suppose a researcher is interested in studying take-up of a new production technology. She may wish to conduct an LFE ex ante to get a sense of the types of preferences and beliefs that are present among the target population. For example, suppose she conducts an LFE to obtain a ranking according to risk and ambiguity aversion, which she then uses to stratify eligibility assignment. The LFE fulfills an ex-ante design role as opposed to an ex-post data analysis role.

In order to take this approach, the researcher should be convinced–in principle by means of a model, but possibly as a result of previous findings or intuition/heuristics–that the measure elicited by means of the LFE correlates with the main dependent variable in a significantly meaningful way. This brings us back to feature 1.

All in all, the main take-away is that LFEs can serve as complements to other types of empirical methods, in particular RCTs and NFEs, when conducting research and formulating policy recommendations.

2.4 Methodological advances

The final purpose for conducting LFEs is to identify methodological complexities and their potential effects on the estimates of the causal (treatment) effects, $\hat{g}(x_1, x_2, z)$. Given the relatively heuristic nature of research questions aimed at identifying methodological difficulties, LFEs conducted with this purpose tend to be driven by intuition, more so than explicitly developed models. As such, somewhat different from what was discussed in section 2.1, these LFEs tend to inform the development of new models (the role for experiments to inform theory as well as theory to inform experiments has been discussed more generally by for example Samuelson, 2005).

Table 1 summarizes four studies that have focused on methodological questions using LFEs. I discuss two of them further below: Charness and Viceisza (2013) and Cilliers et al.

(2012).

In Charness and Viceisza (2013), we compare responses across three risk elicitation instruments: (1) the multiple price list proposed by Holt and Laury (2002, 2005); (2) a binary mechanism pioneered by Gneezy and Potters (1997); and (3) a non-incentivized willingness-to-risk scale implemented by for example Dohmen et al. (2011). We find a low level of understanding with the Holt-Laury task and an unlikely-to-be-accurate pattern with the willingness-to-risk question. Our analysis indicates that the simple binary mechanism seems to have more predictive power than does the Holt-Laury mechanism. Our study is a cautionary note regarding utilizing either relatively sophisticated mechanisms or non-incentivized questions to elicit risk attitudes in rural developing country contexts.

Perhaps more importantly, our findings also suggest avenues for future work. (1) Given the similarities in how the Holt-Laury task and the Gneezy-Potters task were presented, what might explain the difference in performance? (2) Given the simplicity of the non-incentivized willingness-to-risk question, why do the data show a non-standard distribution; is it because the term 'risk' is not well defined or the instrument is not incentivized? (3) How do these complexities impact our ability to characterize risk preferences? For example, in the Holt-Laury task over 50% of our sample is inconsistent and thus, should be discarded. If one were to use the number of safe lotteries chosen as a measure of the respondents' risk aversion instead, what type of bias does this introduce? This latter question in particular suggests not just areas for further experimentation, but also areas for further economic and statistical modeling. For example, to what extent are respondent's deviations from initial (theoretical) predictions considered trembles or mistakes? What is the appropriate benchmark and what does a model of 'deviations' look like? These latter questions reiterate the feedback from experiments to theory.

Cilliers et al. (2012) quantify what I termed the 'mzungu effect' (see Viceisza, 2012, page 75). Specifically, they ask whether the presence of white foreigners influence behavior measured by means of LFEs in developing countries. They experimentally vary foreigner presence across LFEs conducted in 60 communities in Sierra Leone, and assess its effect on standard measures of generosity.

They find that foreigner presence substantially increases player contributions in dictator games, by an average of 19 percent. Using household and village level survey data, they show that the treatment effect–what I call the mzungu effect–is smaller for players who hold positions of authority, suggesting that perceived power differentials between players and the experimenter, based on identity, plays a role in mediating this effect. They also find that subjects from villages with greater exposure to development aid give substantially less, and are more inclined to believe that the LFEs were conducted to test them for future aid. These findings suggest that behavioral responses to researcher identity are in part related to expectations regarding development assistance. More generally, their findings have implications for measuring generosity and the design and administration of LFEs in developing countries.

One way in which researchers have attempted to mitigate this and other types of unintended effects, at least when identifying causal treatment effects, is by maintaining an across-treatment design. In such a design, the researcher randomly assigns subjects to condition $X = x_1$ (treatment 1, possibly a control) and condition $X = x_2$ (treatment 2), and identifies/estimates $g(x_1, x_2, z)$ by subtracting $f(x_1, z)$ from $f(x_2, z)$. The findings reported

by Cilliers et al. (2012) suggest that the presence of a white foreigner (the mzungu) in principle increases $\hat{f}(x,z)$ to $\hat{f}(x,z)+\beta$. But, notice that in an across-treatment design, this will only be problematic *if* β is specific to one condition, but not another. So, if one is willing to assume that the effect β is uniform and orthogonal to x, $\hat{g}(x_1,x_2,z)$ will not be impacted by the mzungu effect since $\hat{g}(x_1,x_2,z) = (f(x_2,z)+\beta) - (f(x_1,z)+\beta) = f(x_2,z) - f(x_1,z)$, which is the same as before.

Having said this, unintended effects remain problematic if the causal effect is not identified across-treatments or if the researcher has reason to believe that β is *not* uniform and orthogonal to x. In either of these cases, β persists and further steps are required to get a sense of its magnitude and correct for it as necessary.

3 Can't we all just 'generalize' along?

I now turn to the so-called 'hot topic' of generalizability, or what some typically refer to as external validity. Crudely speaking, generalizability asks to what extent can the findings in one context generalize to a different (in this case, broader and less stylistic) context? Concretely, if we move from an LFE to an RCT, to an NFE, or to a field context that is not necessarily experimental, will the estimates we report hold up? Generalizability is not unique to experiments (as Al-Ubaydli and List, 2012, and others have indicated); it just so happens that in non-experimental contexts, the relevance of generalizability is trumped by the more pressing matter of identification.

Generalizability has recently been the topic of much debate in the experimental literature in particular (see for example the discussions by Levitt and List, 2007; Falk and Heckman, 2009; Camerer, 2011). Having followed the aforementioned debate and collected my own thoughts over the years, my opinion is that of the typical 'two-handed' economist. On the one hand, generalizability can be quite relevant, for example when an LFE is conducted to have direct policy implication. On the other hand, if an LFE is conducted to for example 'stress-test' a particular hypothesis derived from a game-theoretic model, we may not be concerned about generalizability, at least in the short run. So, it depends on the purpose of the research question under consideration.

The extent to which findings generalize in contexts where we care about generalizability is an empirical question that is tied to theory. On the empirical side, only years of running experiments will shed light on the conditions in which generalizability may or may not be problematic (Camerer, 2011, takes a step in this direction by reviewing some of the literature). On the theoretical side, it is necessary to formalize what it means to generalize. Specifically, what is the comparable, more general context that serves as the benchmark for the generalization? This can be used to create empirical tests.

3.1 Formalizing generalizability

Al-Ubaydli and List (2012) formalize generalizability by taking the all-causes model discussed in section 2.1 as the starting point.

They build on the previously discussed components as follows. Let $T \subseteq S_X \times S_X \times S_Z$ be the target space that describes the set of causal triples in which a researcher is interested.

Typically, the researcher wants to know the exact value of of the causal effect, $g(x, x', z)$, of each element of T. Let $h : S_X \times S_X \times S_Z \to \mathbb{R}$ be a function that captures the aspect of a causal effect in which the researcher is interested. Before embarking upon a new empirical investigation, a researcher has a prior $F^0_{x,x',z} : \mathbb{R} \to [0, 1]$ about the value of $h(x, x', z)$ for each $(x, x', z) \in T$. The prior is a cumulative density function based on existing theoretical and empirical studies, as well as the researcher's introspection.

An empirical investigation is a dataset $D \subseteq S_X \times S_X \times S_Z$. D and T may be disjoint, and both may be singletons. Let the results $R \subseteq D \times \mathbb{R}$ be the set of causal effects obtainable from the dataset D making no parametric assumptions: $R = \{(x, x', z, g(x, x', z)) : (x, x', z) \in D\}$. After seeing the results, R, the researcher updates her prior $F^0_{x,x',z}$ for each $(x, x', z) \in T$, forming a posterior $F^1_{x,x',z}$. The generalizability debate is concerned with formation of the posterior, especially for elements of $T \backslash D$. The posterior is the conclusion of the empirical investigation.

Given a set of priors $\mathfrak{F}^0 = \{F^0_{x,x',z} : (x, x', z) \in S_X \times S_X \times S_Z\}$ and results R, the generalizability set $\Delta(R)$ is the set of causal triples outside the dataset where the posterior $F^1_{x,x',z}$ is updated as a consequence of learning the results, that is:

$$\Delta(R) = \{(x, x', z) \in \{S_X \times S_X \times S_Z\} \backslash D : F^1_{x,x',z}(\theta) \neq F^0_{x,x',z}(\theta) \text{for some } \theta \in \mathbb{R}\}.$$

Results are generalizable when the generalizability set is non-empty ($\Delta(R) \neq \emptyset$) and a researcher is said to generalize when the generalizability set intersects with the target space ($\Delta(R) \cap T \neq \emptyset$). Specifically, they distinguish between three types of generalizability:

1. Given prior beliefs $\mathfrak{F}^0$, a set of results R has *zero* generalizability if its generalizability set is empty, $\Delta(R) = \emptyset$. This is the most conservative empirical stance.

2. Given prior beliefs $\mathfrak{F}^0$, a set of results R has *local* generalizability if its generalizability set contains points within an arbitrarily small neighborhood of points in D, that is:

   $$(x, x', z) \in \Delta(R) \Rightarrow (x, x', z) \in B_\varepsilon(\overline{x}, \overline{x}', \overline{z}) \text{ for some } \varepsilon > 0, (\overline{x}, \overline{x}', \overline{z}) \in D.$$

   The simplest way to obtain local generalizability is to assume $h(x, x', z)$ is continuous (or only has a small number of discontinuities), since continuity implies local linearity and therefore permits local extrapolation.

3. Given prior beliefs $\mathfrak{F}^0$, a set of results R has *global* generalizability if its generalizability set contains points outside an arbitrarily small neighborhood of points in D, that is:

   $$\exists (x, x', z) \in \Delta(R) : (x, x', z) \notin B_\varepsilon(\overline{x}, \overline{x}', \overline{z}) \text{ for some } \varepsilon > 0, \text{for all} (\overline{x}, \overline{x}', \overline{z}) \in D.$$

   At the core, global generalizability is about assuming that a large change in (x, x', z) does not have a large effect on h.

3.2 Lab and field: Complements, not substitutes

Al-Ubaydli and List (2012) use this setup to discuss the advantages of field and lab experiments. The bottom line of their discussion is that, once viewed through the lens of

their model, lab and field experiments are more likely to serve as complements rather than substitutes.

Assuming that (A1) a causal effect is investigation-neutral (that is, unaffected by the fact that it is being induced by a scientific investigator *ceteris paribus*) and (A2) as economists we are more interested in behavior in a natural setting (that is, a triple (x, x', z) that can plausibly exist in the absence of academic, scientific investigation), they discuss three propositions:

Proposition 1 Under a liberal stance (global generalizability), neither field nor laboratory experiments are demonstrably superior to the other.

Proposition 2 Under a conservative stance (local generalizability; or if the researcher is confident that $h(x, x', z)$ is continuous), field experiments are more useful than laboratory experiments.

Proposition 3 Under the most conservative stance (zero generalizability), field experiments are more useful than laboratory experiments because they are performed in one natural setting.

Some of these considerations follow immediately from assumption A2 above. By their very nature lab-like experiments represent an environment that could only come about as the result of a scientific investigation. As such, according to the above definition, they are not completed in natural settings. If we are interested in behavior in natural settings, they are at a disadvantage. This may be particularly relevant if there are factors that affect behavior in the non-natural setting that are otherwise not present. This can typically be a problem when participants know they will be or are being studied, as tends to be the case in lab experiments, AFEs and FFEs.

For example, there may be selection as to whom chooses to participate in the experiment. As Al-Ubaydli and List (2012) discuss, while this need not be a problem for internal validity (as one typically randomizes treatments over those who choose to participate), we must be careful in generalizing findings (see their pages 7, 17-18 and what they call 'treatment specific selection bias'). Another example is that the participant may adapt her behavior because she knows she is being studied. As Cilliers et al. (2012) discuss, this may very well be related to participants' expectations as to how the results of the experiments may inform subsequent policy. As discussed in section 2.4, similarly to the above, this type of 'adaptation bias' (β) need not pose a threat for internal validity if one maintains an across-treatment design to identify the causal effect and if one assumes that β is uniform across and orthogonal to treatment conditions (x, x').

Despite propositions 1 through 3 above, the model discussed by Al-Ubaydli and List (2012) shows that there is a critically important advantage of lab experiments over field experiments and in particular, of LFEs over certain FFEs and NFEs. Their main support for this claim goes back to an issue I indicated previously when discussing the first purpose for LFEs, in particular when illustrating the rationale for the experiments conducted by Hill and Viceisza (2012). Many causal triples (x, x', z) are inestimable in field settings due to ethical, feasibility, or cost reasons. As they discuss, the range of causal triples that cannot be directly estimated in an NFE and that lie outside the local generalizability set

of estimable causal triples is so large that in many environments, lab and field experiments become natural complements. They also allude to ease of replication as a clear advantage of lab-like experiments over pure field experiments.

So, despite the debate, it appears to me that particularly in the case of LFEs intended to inform (development) policy, we should care about generalizability. Interestingly, as several researchers have argued (see for example Falk and Heckman, 2009; Camerer, 2011; Al-Ubaydli and List, 2012), the complementary role of lab, lablike field, and pure field experiments suggests running more experiments at all levels. This is not only an important step in helping us gain insight into the conditions in which findings generalize, but it is also likely to help us formulate policies more confidently. Indeed, there is an increasing trend in the literature to conduct different types of experiments in a complementary manner, as signaled by the discussion in section 2.3.

4 Basic principles

Having discussed the primary purposes of LFEs and the extent to which generalizability is important, in this section I turn to some basic, general principles a researcher may want to consider when conducting an LFE, specifically in a rural area of a developing country. These principles are likely to be necessary, but not sufficient. A broader discussion with additional supporting references can be found in Viceisza (2012). These principles are complementary to and sometimes overlapping with those discussed by List (2011).

Principle 1 Clearly define the research question and purpose of interest.

Principle 2 Develop a theoretical or conceptual framework that informs those components of an experiment that need to be controlled.

Principle 3 Depending on the purpose of the experiment, consider parallelism and generalizability. Design the experiment considering the most relevant components of the agents' day-to-day decisionmaking environment.

Principle 4 Identify the focus variables–that is, the dependent/outcome variable Y and the main explanatory/treatment variable X–and the nuisance/additional explanatory variables Z.

Principle 5 Vary the treatment variables independently. Some focus variables may be held constant–this is a special case of a 'treatment.'

Principle 6 Use randomization or another more elaborate experimental design as an ex ante tool to indirectly control unobservable characteristics (part of Z), specifically nuisance variables, and thus rule out potential confounding of treatment effects. Use additional, possibly stated-preference, data collected after the experiments to test for possible confounding effects.

Principle 7 Submit your experiment protocol to the Institutional Review Board (IRB) for review. When conducting experiments in rural areas of developing countries,

this may include consultation of local IRBs or local authorities such as local governments.

Principle 8 Keep the experiment protocol as simple as possible given the research question under consideration.

Principle 9 Decide whether your experimental environment will be paper based or computer based. The main deciding factors should be (1) whether your subjects, enumerators, or both can handle computer-based tasks and (2) whether it is feasible to have the computer hard- and software deployed in the field.

Principle 10 Decide whether the experiment will maintain neutral or loaded framing. If neutral, make sure to test subjects' understanding more than if framed.

Principle 11 Determine how much information should be concealed from your subjects. Deception should be used only if absolutely necessary, in which case the protocol should be carefully scrutinized by the IRB.

Principle 12 At a minimum, do 'communal' exercises to test subjects' understanding. Be wary of those subjects who are likely to fall through the cracks and of dazed looks!

Principle 13 Decide on your payment protocol (single or double blind), and communicate that to your subjects via the instructions so they understand the level of privacy involved in the experiment protocol. It is important to communicate this tactfully in order not to cause paranoia among the subjects.

Principle 14 Identify a target population, and draw a sample according to a multistage procedure, taking into account sampling and nonsampling errors as well as analytical domains. Make sure that the sample size is based on careful power calculations.

Principle 15 Formulate a proper recruitment strategy to ensure maximum attendance. This entails having a proper invitation (possibly in the form of an official letter supported by an additional letter from a 'trusted' party) that provides necessary, relevant information that entices subjects to attend the experiment. Make sure all recruiters are trained collectively and are aware of the proper protocol for recruitment (that is, what information to provide, whether to obtain the consent of the subjects, how to deal with nonresponses, and so on).

Principle 16 Hire a team of collaborators that comprises a field coordinator (who coordinates the listing, recruitment, and show-up of participants at the experiment), a main translator (who will either do line-by-line translation or conduct the experiment herself), and if necessary an assistant experimenter (who will perform the necessary calculations behind the scenes, provided the experiment is paper based), additional assistants (who will facilitate the procedures of the experiment), and enumerators (who will conduct any necessary surveys). It is important for you to be convinced that your team of collaborators is well equipped to perform the tasks at hand.

Principle 17 Secure an 'ideal' locale (that is, a spacious one with tables, chairs, and a board, and relatively easily accessible) to serve as the laboratory for your experiment. Make sure the assistant experimenter has a private area or room in which to perform his tasks during the experiment. Buy dividers locally if possible; these could vary from voting boxes to sturdy cartons to standard boxes.

Principle 18 Have money available to pay your subjects after the experiments. Take appropriate measures to ensure the security of the field team (in relation to the money). Try not to publicly display cash. Use tact in dealing with cash; be prepared for potential crowds as word travels. Train the field team on how to deal with such potential situations as well.

Principle 19 Design and conduct surveys to complement your experimental data. Typically, these surveys will be of two types: (1) a relatively short postexperiment survey of 30 to 60 minutes intended to test subjects understanding, gain a better understanding of their rationale for decisionmaking, and collect specific individual-level characteristics that may not be collected as part of the household survey and (2) an extensive household survey intended to collect observable characteristics of the household as a whole (to be used to capture heterogeneity across the subject pool). Train enumerators collectively, and make sure that each enumerator reports to a lead enumerator, who in turn reports to the field coordinator, who in turn reports to you.

Furthermore, I would suggest to keep the following practical aspects in mind:

Aspect 1 Conduct a pilot experiment. Pilot experiments are very informative and shed light on many complications that may arise during the actual experiment.

Aspect 2 Randomize entry to and seating in the experimental laboratory.

Aspect 3 Plan for attrition among experiment subjects. One way to prepare for this is to recruit regular and alternate subjects.

Aspect 4 Plan for (curious) nonparticipants. Members of the larger community who are not subjects might still be interested in the experiment; they should be treated with courtesy and tact.

Aspect 5 Beware of long sessions. Do not address too many issues in a single experiment. Keep things simple, and have realistic expectations.

Aspect 6 Be aware of religious and cultural sensitivities. Subjects might be sensitive to particular colors, symbols, or concepts (such as usury, in the case of certain Muslim communities).

Aspect 7 Be prepared for distractions. A certain level of distraction (from children accompanying a subject, for example) might be unavoidable, so plan for this; holding multiple sessions, as suggested below, can help.

Aspect 8 Hold two or more sessions for each treatment variable you test. This allows you to identify and allow for session-level differences such as nuisances or distractions that might affect all individuals participating in the session.

5 Ways forward: Experimenting with experiments

In this article, I reviewed some of the literature that reports LFEs, particularly when conducted in rural areas of developing countries. I discussed four main purposes of these types of experiments. First, to test theories or heuristic principles. Second, to identify and estimate parameters associated with characteristics. Third, to explore the structural nature of parameters derived from empirical methods that may or may not be experimental. Fourth, to assess methodological difficulties associated with LFEs and their potential impact on parameter estimates. I also addressed the importance of generalizability for LFEs intended to inform policymaking and emphasized the complementary role between LFEs and other empirical methods, in particular other experiments. I concluded with a discussion of 19 basic principles and eight practical aspects to consider when conducting LFEs.

Having discussed these main issues and specifically, after having identified the four purposes of LFEs, I foresee–and to some extent hope–that the literature will continue in at least the following three, nonmutually exclusive directions.

First, I believe the literature would benefit from a continued refinement of experimental methodologies. In general, but perhaps even more so in the context of LFEs intended to inform policy, it is important to know what to infer from experimental parameters. Thus, as has been a mainstay of (lab) experimental economics, it will be worthwhile to continue designing experiments aimed at addressing methodological questions. Among these are topics that address (1) framing, experimenter, design and elicitation effects; (2) the interplay between experimental and non-experimental (quasi and structural) data; (3) potential benefits and costs (statistically and physically) from maintaining modes of data collection beyond the standard paper- or electronic-based methods in the field such as virtual experiments (see for example Fiore et al., 2009); and (4) the types of data collected in the field beyond standard choices such as neuro- or bio-economic data (see for example Giné et al., 2012).

Second, as signaled previously, the generalizability 'debate' is partly an empirical question that will only be resolved over time. As such, I foresee (and also hope) that the literature will continue to conduct LFEs in conjunction with other types of experiments and empirical methods. This will not only enable us to better understand the complementary role of different types of experiments, but it is also likely to lead to more careful and precise policy recommendations. Furthermore, since experiments also inform the development of new or alternative models, this process is also likely to contribute to the formalization of generalizability.

Finally, while implicit in the former, I foresee that LFEs will continue to be conducted together with other types of empirical methods, in particular structural approaches. Not only can experiments enable estimation of structural parameters (such as risk and time preference parameters), but they also allow for identification of broader and more complex structural models (as done by Mahajan and Tarozzi, 2011, for example).

Overall, I find that LFEs have an important role to play in informing policymaking,

particularly in rural areas of developing countries.

References

Akay, A., P. Martinsson, H. Medhin, and S. Trautmann (2012). Attitudes toward uncertainty among the poor: An experiment in rural Ethiopia. *Theory and Decision 73*, 453–464. http://dx.doi.org/10.1007/s11238-011-9250-y.

Al-Ubaydli, O. and J. A. List (2012, March). On the generalizability of experimental results in economics. Working Paper 17957, National Bureau of Economic Research.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström (2008). Eliciting Risk and Time Preferences. *Econometrica 76*(3), 583–618.

Ashraf, B. N. (2009). Spousal Control and Intra-Household Decision Making : An Experimental Study in the Philippines. *The American Economic Review 99*(4), 1245–1277.

Ashraf, N., I. Bohnet, and N. Piankov (2006). Decomposing trust and trustworthiness. *Experimental Economics 9*, 193–208.

Ashraf, N., D. Karlan, and W. Yin (2006). Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines. *The Quarterly Journal of Economics 121*(2), 635–672.

Attanasio, O., A. Barr, J. C. Cardenas, G. Genicot, and C. Meghir (2012, April). Risk pooling, risk preferences, and social networks. *American Economic Journal: Applied Economics 4*(2), 134–67.

Banerjee, A. V. and E. Duflo (2009). The experimental approach to development economics. *Annual Review of Economics 1*, 151–178.

Banerjee, A. V. and E. Duflo (2010). Giving credit where it is due. *Journal of Economic Perspectives 24*(3), 61–80.

Barr, A. (2003). Trust and expected trustworthiness: Experimental evidence from Zimbabwean villages. *The Economic Journal 113*(489), 614–630.

Barr, A., T. Packard, and D. Serra (2012). Participatory accountability and collective action: Experimental evidence from Albanian schools. Working Paper.

Berge, L. I. O., K. Bjorvatn, and B. Tungodden (2011). Human and financial capital for microenterprise development: Evidence from a field and lab experiment. Working Paper.

Bernard, T., L. Sène, A. Viceisza, and F. Wouterse (2012). Let's coordinate! Experimental evidence from farmer groups in Senegal. Working Paper.

Binswanger, H. P. (1980). Attitudes Toward Risk, Experimental Measurement in Rural India. *American Journal of Agricultural Economics 62*(August), 395–407.

Bursztyn, L. and L. C. Coffman (2012). The schooling decision: Family preferences, intergenerational conflict, and moral hazard in the Brazilian favelas. *Journal of Political Economy 120*(3), pp. 359–397.

Burtless, G. (1995). The case for randomized field trials in economic and policy research. *Journal of Economic Perspectives 9*(2), 63–84.

Camerer, C. F. (2011). The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List. *SSRN eLibrary*.

Card, D., S. DellaVigna, and U. Malmendier (2011, September). The role of theory in field experiments. *Journal of Economic Perspectives 25*(3), 39–62.

Carter, M. (2008). Inducing innovation: Risk instruments for solving the conundrum of rural finance. Working Paper.

Charness, G., U. Gneezy, and M. A. Kuhn (2013). Experimental methods: Extra-laboratory experiments-extending the reach of experimental economics. *Journal of Economic Behavior & Organization 91*(0), 93 – 100.

Charness, G. and A. Viceisza (2013). Comprehension and risk elicitation in the field: Evidence from rural Senegal. Working Paper.

Cilliers, J., D. Oeindrila, and B. Sidiqqi (2012). 'White man's burden'? A field experiment on generosity and foreigner presence. Working Paper.

Cole, S. A., X. Giné, J. Tobacman, P. B. Topalova, R. M. Townsend, and J. I. Vickery (2009). Barriers to household risk management: Evidence from India. Harvard Business School Finance Working Paper No. 09-116.

Cole, S. A., X. Giné, and J. I. Vickery (2012). How does risk management influence production decisions? Evidence from a field experiment. Working Paper.

Davis, D. D. and C. A. Holt (1993). *Experimental Economics*. Princeton, NJ: Princeton University Press.

de Brauw, A. and P. Eozenou (2011). Measuring risk attitudes among Mozambican farmers. *Harvest Plus Working Paper Series 6*.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature 48*(2), 424–55.

Delavande, A., X. Giné, and D. McKenzie (2011). Eliciting probabilistic expectations with visual aids in developing countries: How sensitive are answers to variations in elicitation design? *Journal of Applied Econometrics 26*(3), 479–497.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequencesmen. *Journal of the European Economic Association 9*(3), 522–550.

Duflo, E., R. Glennerster, and M. Kremer (2007). Using randomization in development economics research: A toolkit. Volume 4 of *Handbook of Development Economics*, pp. 3895 – 3962. Elsevier.

Engle Warnick, J. C., J. Escobal, and S. C. Laszlo (2011). Ambiguity aversion and portfolio choice in small-scale peruvian farming. *The B.E. Journal of Economic Analysis & Policy 11*(1), 68.

Falk, A. and J. J. Heckman (2009). Lab experiments are a major source of knowledge in the social sciences. *Science 326*(5952), 535–538.

Feigenberg, B., E. Field, and R. Pande (2012). The economic returns to social interaction: Experimental evidence from microfinance. Working Paper.

Finan, F. and L. Schechter (2012). Vote-buying and reciprocity. *Econometrica 80*(2), 863–881.

Fiore, S. M., G. W. Harrison, C. E. Hughes, and E. E. Rutström (2009, January). Virtual experiments and environmental policy. *Journal of Environmental Economics and Management 57*(1), 65–86.

Friedman, D. and S. Sunder (1994). *Experimental Methods: A Primer for Economists*. Cambridge, UK: Cambridge University Press.

Giné, X., J. Goldberg, D. Silverman, and D. Yang (2012). Revising commitments: Field evidence on the adjustment of prior choices. Working Paper.

Giné, X., J. Goldberg, and D. Yang (2012, September). Credit market consequences of improved personal identification: Field experimental evidence from Malawi. *American Economic Review 102*(6), 2923–54.

Giné, X., P. Jakiela, D. Karlan, and J. Morduch (2010, September). Microfinance games. *American Economic Journal: Applied Economics 2*(3), 60–95.

Gneezy, U., K. L. Leonard, and J. A. List (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica 77*(5), 1637–1664.

Gneezy, U. and J. Potters (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics 112*(2), pp. 631–645.

Guiteras, R. (2012). Eliciting and utilizing willingness to pay: Evidence from field trials in Northern Ghana. Working Paper.

Harrison, G. W., S. J. Humphrey, and A. Verschoor (2010). Choice under uncertainty: Evidence from Ethiopia, India and Uganda. *The Economic Journal 120*(543), 80–104.

Harrison, G. W. and J. A. List (2004). Field experiments. *Journal of Economic Literature 42*(4), 1009–1055.

Heckman, J. J. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives 9*(2), 85–110.

Heckman, J. J. (2000). Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics 115*(1), 45–97.

Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature 48*(2), 356–98.

Hill, R. and A. Viceisza (2012). A field experiment on the impact of weather shocks and insurance on risky investment. *Experimental Economics 15*, 341–371. http://dx.doi.org/10.1007/s10683-011-9303-7.

Hill, R. V., E. Maruyama, and A. Viceisza (2012). Breaking the norm: An empirical investigation into the unraveling of good behavior. *Journal of Development Economics 99*(1), 150 – 162.

Hill, R. V. and M. Torero (2009). *Innovations in insuring the poor*, Volume 17 of *2020 Vision Focus Brief*. Washington, DC: International Food Policy Research Institute (IFPRI).

Holt, C. A. and S. K. Laury (2002). Risk aversion and incentive effects. *The American Economic Review 92*(5), 1644–1655.

Holt, C. A. and S. K. Laury (2005). Risk aversion and incentive effects: New data without order effects. *The American Economic Review 95*(3), 902–904.

Imbens, G. W. (2010). Better late than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature 48*(2), 399–423.

Jakiela, P., E. Miguel, and V. te Velde (2012). You've earned it: Combining field and lab experiments to estimate the impact of human capital on social preferences. Working Paper.

Jakiela, P. and O. Ozier (2012). Does Africa need a rotten kin theorem? Experimental evidence from village economies. Working Paper.

Jamison, J. and D. Karlan (2011). Measuring preferences and predicting outcomes. Working Paper.

Karlan, D. S. (2005). Using experimental economics to measure social capital and predict financial decisions. *The American Economic Review 93*, 1688–1699.

Levitt, S. D. and J. A. List (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives 21*(2), 153–174.

Levitt, S. D. and J. A. List (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review 53*(1), 1–18.

List, J., S. Sadoff, and M. Wagner (2011). So you want to run an experiment, now what? some simple rules of thumb for optimal experimental design. *Experimental Economics 14*, 439–457. http://dx.doi.org/10.1007/s10683-011-9275-7.

List, J. A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives 25*(3), 3–16.

Mahajan, A. and A. Tarozzi (2011). Time inconsistency, expectations and technology adoption: The case of insecticide treated nets. Working Paper.

Samuelson, L. (2005). Economic theory and experimental economics. *Journal of Economic Literature 43*(1), 65–107.

Samuelson, P. (1947). *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Sandmo, A. (1971). On the theory of the competitive firm under price uncertainty. *The American Economic Review 61*(1), 65–73.

Schechter, L. (2007, September). Theft, gift-giving, and trustworthiness: Honesty is its own reward in rural Paraguay. *American Economic Review 97*(5), 1560–1582.

Smith, V. L. (1982). Microeconomic Systems as an Experimental Science. *The American Economic Review 72*(5), 923–955.

Smith, V. L. (1987). Experimental methods in economics. In J. Eatwell, M. Milgate, and P. Newman (Eds.), *The New Palgrave: A Dictionary of Economics* (First ed.). New York, NY: Palgrave MacMillan.

Tanaka, T., C. F. Camerer, and Q. Nguyen (2010, March). Risk and time preferences: Linking experimental and household survey data from Vietnam. *American Economic Review 100*(1), 557–71.

Viceisza, A. C. G. (2012). *Treating the Field As a Lab: A Basic Guide to Conducting Economics Experiments for Policymaking*, Volume 7 of *Food Security in Practice Technical Guide Series*. Washington, DC: International Food Policy Research Institute (IFPRI).

Table 1: Overview of selected studies reporting LFEs in developing/emerging country contexts

Study	Finding	Country
	Purpose 1: Testing theory and/or heuristics	
Ashraf (2009)	Info & comm interact with household control to produce mutable gender-specific outcomes.	Philippines
Attanasio et al. (2012)	Advantages to grouping assortatively on risk may be inaccessible when trust is absent or low.	Colombia
Bursztyn and Coffman (2012)	Intergenerational conflicts in schooling decisions & lack of parental control & observability.	Brazil
Giné et al. (2010)	Group lending increases risk taking and group contracts benefit borrowers.	Perú
Hill and Viceisza (2012)	Insurance weakly positively affects fertilizer purchases while weather shocks also have an impact.	Ethiopia
Schechter (2007)	Evidence for a model of theft is found: giving ↑ when trust is lower & the threat of theft is greater.	Paraguay
	Purpose 2: Eliciting characteristics	
	Risk, time and ambiguity preferences	
Akay et al. (2012)	Strong risk and ambiguity aversion were found suggesting their potential importance for agriculture.	Ethiopia
Binswanger (1980)	Risk attitudes are elicited & at high payoff levels, virtually all subjects are moderately risk averse.	India
de Brauw and Eozenou (2011)	Risk attitudes are elicited & 75% of the sample is consistent with rank dependent utility.	Moçambique
Giné et al. (2012)	Revisions of prior choices based on measures of time suggest self-control problems.	Malawi
Harrison et al. (2010)	Risk attitudes are elicited & 50% of the sample is consistent with expected utility (prospect) theory.	Ethiopia, India, Uganda
Mahajan and Tarozzi (2011)	Time preferences (TP) are elicited & used to identify parameters and time-inconsistent agents.	India
Tanaka et al. (2010)	Using measures of risk/time, people in higher-income villages are less loss-averse/more patient.	Vietnam
	Social capital (preferences, norms, trust and reciprocity)	
Ashraf et al. (2006)	Dictator and risk experiments show that mainly expectations of trustworthiness explain trust.	Russia, S. Africa, USA
Barr (2003)	While trustworthiness is important for trust, so are non-expectational desires to 'community build'.	Zimbabwe
Hill et al. (2012)	Trust games show that small increases in non-reciprocal behavior result in an unraveling reciprocity.	Perú
Jakiela and Ozier (2012)	Observability is varied and women seek to conceal experimental income to avoid pressures to share.	Kenya
	Gender differences in preferences and decisionmaking	
Gneezy et al. (2009)	Gender differences may be due to nurture (social context) rather than nature.	Tanzania, India
	Purpose 3: Unpacking the black box	
Ashraf et al. (2006)	Women with a lower discount rate (TP) are more likely to open a commitment savings account.	Philippines
Barr et al. (2012)	Public goods game data correlate with survey data on voluntary participation in institutions.	Albania
Berge et al. (2011)	LFE data (e.g. risk, time) & RCTs show that human capital is a barrier to business training.	Tanzania
Bernard et al. (2012)	Coordination games & RCTs are used to assess the impact of training on collective behavior.	Senegal
Engle Warnick et al. (2011)	Ambiguity aversion ↓ the likelihood that farmers plant more than one variety of their main crop.	Perú
Finan and Schechter (2012)	Trust game & vote-buying survey data show that politicians target reciprocal individuals.	Paraguay
Jakiela et al. (2012)	Dictator games & RCTs show that education may impact social preferences, norms & institutions.	Kenya
Jamison and Karlan (2011)	Risk & time data predict outcomes from RCTs similarly to demographic survey data.	Uganda
Karlan (2005)	"Trustworthy" individuals in a trust game are less likely to default on their microfinance loans.	Perú
	Purpose 4: Methodological advances	
Charness and Viceisza (2013)	A simple binary risk elicitation mechanism outperforms Holt-Laury and a willingness-to-risk scale.	Senegal
Cilliers et al. (2012)	Foreigner presence ↑ contributions in dictator games. Effect driven by perceived power differentials & aid.	Sierra Leone
Delavande et al. (2011)	Responses are robust to variations in three facets of a belief elicitation methodology.	India
Guiteras (2012)	The BDM mechanism is compared to the take-it-or-leave-it method for eliciting willingness to pay.	Uganda

Table 2: Estimates of causal effects for purpose 1 studies

Study	Y	X	$\hat{g}(x, x', z)$[a]
Ashraf (2009)[b]	$Y = 1$ if gift certificate for self	x_1 = public	$\hat{g}(x_1, x_2, z) = -0.33$
	$Y = 0$ if deposit into own/joint account	x_2 = private	$\hat{g}(x_2, x_3, z) = 0.28$
		x_3 = negotiation	$\hat{g}(x_1, x_3, z) = -0.05$
Attanasio et al. (2012)[c]	$Y = 1$ if risk pooling group formed	x_1 = close family/friends	$\hat{g}(x_1, x_2, z) = 0.30$
	$Y = 0$ if risk pooling group not formed	x_2 = other relationship	
Bursztyn and Coffman (2012)[d]	Y_1=parent's preference for conditionality	x_1=baseline	$\hat{g}(x_1, x_2, z) = -0.62$
	Y_2=child's preference for conditionality	x_2=text message	$\hat{g}(x_1, x_3, z) = -0.64$
		x_3=don't tell	
		x_4=nonclassroom	
Giné et al. (2010)[e]	Y_1 = risky project choice rate	x_1 = individual liability (IL)	$\hat{g}(x_1, x_2, z) = -0.02$
	Y_2 = repayment rate	x_2 = joint liability (JL)	$\hat{g}(x_2, x_3, z) = 0.02$
		x_3 = JL + monitoring (M)	$\hat{g}(x_3, x_4, z) = -0.07$
		x_4 = JL + M + communication (C)	$\hat{g}(x_4, x_5, z) = -0.01$
		x_5 = JL + M + C + partner choice	
		$x_6, ..., x_{10}$ = Above + dynamic incentives	†
Hill and Viceisza (2012)	Y = bags of fertilizer	x_1 = no insurance	$\hat{g}(x_1, x_2, z) = 0.273$
		x_2 = insurance	
Schechter (2007)[f]	Y_1=theft	x_1 = no one passes field	$\hat{g}(x_1, x_2, z) = -0.60$
	Y_2=gift giving	x_2 = someone passes field	$\hat{g}(x_3, x_{3'}, z) = 1.33$
	Y_3=trust	x_3 = number of stealable crops	

[a] I do not report the estimation procedure or the statistical significance of these effects. Refer to the original articles as necessary.

[b] The original article reports gender disaggregated effects. The estimates reported here are for males.

[c] The estimates reported are for social relationships. See table 5 (page 147) of the original article for further detail.

[d] The estimates reported are for the parent's choices.

[e] The estimates reported are for the rate of risky project choice.

† See table 4 (page 77) of the original article for further detail.

[f] The estimates reported are for gift giving. See table 5 (page 1578) of the original article for further detail.

Notes

[1]Notice that LFEs exclude randomized controlled trials (RCTs), which have been of recent interest in some subfields of economics such as development.

[2]I say 'mainly' because some of the discussion in the aforementioned and other sources (such as List et al., 2011; List, 2011; Al-Ubaydli and List, 2012) apply broadly to all types of experiments including LFEs.

[3]There is also a pedagogical role for LFEs to help explain complex economic concepts as alluded to by Carter (2008) and Hill and Viceisza (2012) for the case of insurance contracts. Due to the minimal role that these types of LFEs have played in the literature, I do not discuss this purpose further (Viceisza, 2012, has a slightly longer discussion).

[4]Using data to test models (theories) is one of the basic premises of empirical research. As discussed by Samuelson (1947) among others, it is important to derive *operationally meaningful theorems*–hypotheses about empirical data that could conceivably be refuted, if only under ideal conditions. As such, one of the primary uses for LFEs, and empirical methods more broadly, has been theory testing. Smith (1982) and Card et al. (2011) discuss the role for theory (and heuristics) when conducting experiments.

[5]This feature is closely related to the third purpose of LFEs. Having said this, as I will discuss further below, the third purpose is more general in that it makes the case for LFEs to help us explore the structural nature of all types of (experimental) parameters, not just those obtained through LFEs.

[6]Harrison and List (2004) have a similar discussion, but use alternative notation and at times, terminology.

[7]A slightly more general, yet consistent argument is made by Heckman (2010) regarding the linkages between reduced form (program evaluation) and structural approaches.

[8]This is part of the argument made by Camerer (2011).