CHAPTER 10

Lab in the Field: Measuring Preferences in the Wild

U. Gneezy[*,§,1], **A. Imas**[¶]

[*]University of California, San Diego, La Jolla, CA, United States
[§]University of Amsterdam, Amsterdam, Netherlands
[¶]Carnegie Mellon University, Pittsburgh, PA, United States
[1]Corresponding author: E-mail: ugneezy@ucsd.edu

Contents

1. Theoretically-Relevant Populations 441
2. Using Lab-in-the Field for Collecting Covariates as Complements to RCTs and for Targeting Policy 445
3. Comparing Between Contexts and Cultures 452
4. External Validity 456
5. Conclusion 459
References 461

Abstract

In this chapter, we discuss the "lab-in-the-field" methodology, which combines elements of both lab and field experiments in using standardized, validated paradigms from the lab in targeting relevant populations in naturalistic settings. We begin by examining how the methodology has been used to test economic models with populations of theoretical interest. Next, we outline how lab-in-the-field studies can be used to complement traditional randomized control trials in collecting covariates to test theoretical predictions and explore behavioral mechanisms. We proceed to discuss how the methodology can be utilized to compare behavior across cultures and contexts, and test for the external validity of results obtained in the lab. The chapter concludes with an overview of lessons on how to use the methodology effectively.

Keywords

Experimental economics; Field experiments; Laboratory experiments; Risk preferences; Social preferences; Time preferences

JEL Codes

B40; C91; C92; C93; D01; D03

Lab experiments and field experiments differ on several core dimensions. Lab experiments are typically conducted in environments that attempt to abstract from the naturalistic setting where individuals typically make their decisions. Factors orthogonal to the theoretical problem being studied such as context and background are removed so

© 2017 Elsevier B.V.
All rights reserved.

that the experimenter can maintain tight control and eliminate potential confounds from the study. These experiments are typically conducted on university campuses with convenient populations of students who are aware that their actions are being studied. While the high level of experimenter control has benefits such as reducing noise and ease of replicability, abstracting from the naturalistic setting and using student populations brings into question whether students in the lab making abstract decisions are a good representation of the types of decisions made by individuals actually relevant to the economic theory.

We have learned quite a lot from carefully designed experiments that impose a strict structure on decision-making. Yet, it is important to explore how individuals' preferences in theoretically relevant settings shape behavior. When studying performance under different incentive schemes, output on a real-effort task could be a more appropriate measure than an induced value design.[1] Similarly, manipulating incentives for charitable giving with actual donors to study social preferences may yield more insightful results for charities than the same manipulation in an anonymous giving game in the lab.[2]

Field experiments are conducted in naturalistic environments and typically use a nonstudent population that is not aware that their decisions are being studied. By targeting a population of theoretical interest in its natural environment, the experimenter can be more confident that the results are applicable to the relevant context. However, field experiments sacrifice experimenter control that may inject noise into the data and introduce potential confounds that bias the results. It is also harder to replicate results from field experiments as they are often inherently situation specific. Replicating a dictator game in the university lab with student participants is easier than replicating the same game with a tribe in a remote area. This situation-specific element also makes it difficult to generalize the results and make direct comparisons to other environments and populations.

In this chapter, we discuss a methodology termed "lab-in-the-field" and argue that by combining elements of both lab and field experiments, it provides researchers with a tool that has the benefits of both, while minimizing the respective costs. We define a lab-in-the-field study as one conducted in a naturalistic environment targeting the theoretically relevant population but using a standardized, validated lab paradigm. Targeting the relevant population and setting increases the applicability of the results. Employing a standardized paradigm permits the experimenter to maintain tight control, while allowing for direct comparisons across contexts and populations. Importantly, the use of lab-in-the-field is an important additional tool in understanding preferences in the wild that could be employed alongside traditional field work.

[1] See Fehr et al. (1998) and Gneezy and List (2006) for the qualitative difference in effort and reciprocity depending on methodology used.

[2] See Andreoni and Miller (2002) and Karlan and List (2007) for qualitative differences in price sensitivities in giving depending on the methodology used.

In some cases, the method that we consider is similar to that of Harrison and List (2004). According to their taxonomy, lab-in-the-field is a type of field experiment. What we call lab-in-the-field using nonstandard populations is similar in spirit to what they termed an artefactual field experiment, which they define as "…the same as a conventional lab experiment but with a nonstandard subject pool." As an example of such artefactual field experiment, Harrison and List (2004) discuss the paper by Harrison et al. (2002) who use a standard lab experiment but instead of running it at a university run it in hotels in order to be able to attract a representative sample of the Danish population. Following Charness et al. (2013), we argue that the physical location of the lab is not what defines a method, and laboratory experiments that are run outside of the university are not best described as field experiments.

While there is no clear cut difference here between lab experiments and lab in the field, we argue that the population itself does not make a study a field experiment. For example, Cappelen et al. (2015) conducted a social preference experiment with a representative population in their lab. In our classification, that would be a lab experiment. According to Harrison and List (2004), this would be an artefactual field experiment.

1. THEORETICALLY-RELEVANT POPULATIONS

One of the limitations of standard experiments in the lab is the use of a narrow set of participants, typically university students, with similar cognitive abilities, low variance in age, education, income, etc. A natural concern is whether results obtained in this specific population would be representative of behavior in a more relevant population. Henrich et al. (2010) argue that participants in laboratory experiments are typically drawn from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies, and results obtained from such studies may not generalize to other populations and settings. For example, in fairness and social preference experiments, WEIRD subjects tend to be more generous and make fewer income maximizing offers than participants drawn from other societies.

The issue of generalizability is particularly important for using experimental data for informing economic theory. For example, economic models of financial decision making such as of asset pricing and household consumption and saving were often developed to capture the behavior of market participants like finance professionals (e.g., traders) and individuals investing to save for retirement. Experiments to test these models in the lab typically used a convenient sample of undergraduates and implicitly assumed that behavior in the lab would generalize to the relevant population of experienced traders and financial market participants.

Locke and Mann (2005) discuss the applicability of studying behavior of nonprofessional traders in the context of information cascades and herd behavior in financial

decisions, stating that individuals without experience in financial markets are too far removed from the price discovery process and may therefore behave differently than the population of market professionals. In the paper, the authors study the disposition effect—the tendency to hold on to losing stocks longer than winning stocks—in a population of professional traders and retail traders. Although they find that both groups display a pronounced disposition effect, the former group does not suffer financial losses as a result whereas the latter group does. This discrepancy in how a well-studied behavioral phenomenon effects different populations is taken as evidence for the importance of studying the theoretically relevant population rather than a convenient sample. Theorists examining herding and information cascades similarly argue that to examine herding behavior requires a population of individuals "who trade actively and act similarly" (Bikhchandani and Sharma, 2000).

Alevy et al. (2007) aimed to address this issue by comparing behavior of market professionals and undergraduate students in a paradigm typically used to study information cascades and herding (Anderson and Holt, 1997). In this setting, individuals make decisions based on a noisy private signal and a public signal based on the behavior of others who faced the same decision before them. Cascades are said to form when individuals ignore their private signal to follow the public signal and can be either statistically justified or not depending on the quality of the public and private signals. Students were recruited for a lab study on a university campus, while traders participated in the experiment at the Chicago Board of Trade (CBOT). The behavior in the experiment differed significantly between the two populations. Market professionals were more likely to use their private signal and were more sensitive to the quality of the public signal, making better use of it than the undergraduates. In turn, the professionals were involved in (weakly) fewer cascades overall and significantly fewer suboptimal cascades (reverse cascades).

But professionals do not always "fix" biases. In a similar vein to the information cascade study, Haigh and List (2005) compared the propensity of market professionals (traders on the CBOT) and students to exhibit myopic loss aversion. Myopic loss aversion, which combines two behavioral concepts of loss aversion and mental accounting, predicts that people will take on more risk over a sequence of gambles than when the same gambles are presented in isolation (Benartzi and Thaler, 1995). It has been proposed as an explanation for the equity premium puzzle, suggesting that the high risk premium on stocks is due to traders evaluating asset performance over too narrow of a frame. Using a standard laboratory paradigm from the myopic loss aversion literature (Gneezy and Potters, 1997), Haigh and List (2005) found that rather than displaying less myopic loss aversion than the students, traders were even more likely to take on greater risk when gambles were framed together rather than separately.

Both papers offer insight on the extent to which behavior of relevant populations differ from convenient populations typically used in lab experiments. The results of using the lab-in-the-field methodology in these cases suggest that the students were not

qualitatively different than the relevant population and offer a step in the direction of showing the degree to which behavioral phenomena were applicable outside of the student population.

Policy is often designed to target a specific population. For example, initiatives such as Medicare Part D are aimed at improving the healthcare outcomes of retirees while programs to increase student retention and the development of human capital are targeted towards young children and adolescents. For these policies to be effective, it is important to examine how the preferences of these populations differ from those assumed in standard economic theory.

In the tradition of developmental psychology, Harbaugh et al. (2001) examine the question of whether age and greater market experience mitigates behavioral phenomena such as the endowment effect—the gap in valuations of a good between buyers and sellers. If age and market experience brings behavior closer to the predictions of the neoclassical model, then adults are expected to show a lower gap than children. The participants in the experiment were kindergarten children and undergraduates enrolled in an introductory economics class. Using the standard paradigm of Knetsch (1989), participants were randomly endowed with one of two objects and then asked whether they would like to keep the object or trade it for the other. The school-aged students made choices between different goods than the college students: the former kept or traded toys and school supplies while the latter made choices over chocolates and coffee mugs. The main finding was no difference in the propensity to choose the endowed item between the age groups, suggesting that exposure to markets between kindergarten and college does not diminish this behavioral phenomenon.

In a paper titled "GARP for Kids," Harbaugh et al. (2001) further studied the relationship between age and rationality by presenting groups of children aged 7 and 11 and undergraduate students with a series of choices between bundles of goods while varying relative prices and budget. Andreoni and Miller (2002) have previously used this experimental paradigm in a standard lab setup to test whether preferences are transitive and consistent with the Generalized Axiom of Revealed Preference (GARP). The authors find that children as young as seven already display a high degree of choices consistent with GARP. By age 11 years, the choices of children appear just as consistent as those of adult undergraduates, suggesting that models of economic behavior can be applied to children as well as adults.

The study of social preferences is a rapidly growing literature in economics. Several models (see Charness and Kuhn, 2011 for a recent survey) aim to capture the systematic violations of the purely selfish, money-maximizing actor, which used to be a typical assumption in neoclassical economics. People have been observed to share money with strangers (Forsythe et al., 1994), sacrifice money by rejecting unfair offers in ultimatum bargaining games (Guth et al., 1982), and cooperate with others even in one-shot interactions (Andreoni, 1989).

However, an important question for both theory and policy is when such social preferences develop. Fehr et al. (2008) sought to answer this question by examining the allocation decisions of young children. Groups of children aged 3–4, 5–6, and 7–8 at local preschools and elementary schools participated in the study. Each child was paired with another and asked to make decisions on how to allocate candy between themselves and their partner in three games. In the prosocial game, the child chose whether to receive one candy and give nothing to their partner, (1, 0), or for both to receive one candy each, (1, 1). This game was designed to examine whether the child would be willing to benefit another at no cost to themselves. In the envy game, the child chose between an equal split of candy, (1, 1), or disadvantageous inequality, (1, 2). Since allocating an extra candy to their partner came at no cost to the child, the envy game aimed to measure participants' inequity aversion. Finally, in the sharing game, children chose between an equal split, (1, 1), or a selfish allocation of, (2, 0). The authors found that preferences for equal splits increased significantly with age. While young children 3–4 years of age preferred selfish allocations, a large fraction of children aged 7–8 years chose the equal split of (1, 1) in each of the three games. These results suggest that rather than being innate, preferences for outcomes consistent with norms such as fairness develop with exposure to culture. In related work, Almås et al. (2010) show that a significant development in morality happens in adolescence, as a response to culture, and Almas et al. (forthcoming) show that family background is crucial for understanding levels of competitiveness.

Dohmen et al. (2012) jointly elicit preferences of both children and their parents. Their goal was to examine the extent to which willingness to take risks and trust others are traits that children inherit from parents, the influence of positive assortative matching on this intergenerational transmission and whether the local attitudes in the environment affects preferences. Children and parent pairs were interviewed at their homes. In order to maintain control and avoid potential confounds, each child and parent were interviewed separately to ensure that each answers questions individually and independent of the others. By studying children and their parents in their homes instead of using a convenient population of undergraduates, the authors were able to gain access to all members in a family. Results suggest significant intergenerational transmission of risk and trust attitudes, which is strengthened by positive assortative matching between the parents. The prevailing attitudes in the environment also play a significant but independent role in shaping children's risk and trust preferences.

On the other end of the age spectrum, as life expectancy in the developed world increases, there is greater pressure to push forward the retirement age and for individuals to keep working later into their years. However, employers are often reluctant to hire older workers (Bendick et al., 1999) due to the notion that seniors are less productive than their younger counterparts. While this belief is common (Kovalchick et al., 2005), evidence for it has been lacking in the economics literature. Using a lab-in-the-field design, Charness and Villeval (2009) aimed to directly compare the preferences and behavior of older

individuals such as retirees to those of a younger population. Particularly, whether the two populations differed in their willingness to cooperate and compete with others.

The first set of experiments took place at two large French firm work sites. To measure cooperation, juniors (under 30) and seniors (over 50) were invited to participate in a team production game that was akin to a public goods game typically studied in lab experiments. In the game, participants were endowed with a private sum that they could choose to either contribute to the public good (cooperate), where it is multiplied and split evenly amongst the group, or to keep it. Given the potential of free riding on the contribution of others, the equilibrium of the game under the assumption of selfishness is to keep the entire endowment while the efficient outcome is for everyone to contribute the maximum amount. To measure competitiveness, juniors and seniors engaged in a real-effort task (solving anagrams) and could choose to either be paid at a piece-rate for every anagram solved or to compete with others in a tournament, where the one who solved the most anagrams would win a large prize and the others would win a much smaller prize. The choice of compensation scheme (piece rate versus tournament) served as the measure of competitiveness. Attitudes towards financial risk-taking were also collected.

Charness and Villeval (2009) found that both juniors and seniors responded strongly to competition and that seniors were more willing to cooperate than juniors. The groups did not differ in their willingness to engage in financial risk taking. Moreover, groups containing both juniors and seniors were better off than more homogeneous groups because seniors responded to the presence of juniors by being even more cooperative. The authors replicated these findings in a traditional lab experiment with a student population and retirees. These findings suggest that age diversity in the work place may potentially be beneficial for both employees and employers.

These experiments comparing decision-making in children and adults of different age groups can teach us about the origin of violations of standard models as well as the development of behavior policy makers may either want to encourage or prevent. By using a standardized experimental paradigm, the authors were able to maintain tight control over the study and make direct comparisons between the populations of interest.

More generally, the evidence reviewed here shows that it is important to elicit behavior and preferences with nonstandard populations that are closer to the theoretically relevant target population. The method of lab-in-the-field is a useful tool to achieve this aim since it is often difficult or impossible to get these nonstandard populations into the standard laboratory environments (Table 1).

2. USING LAB-IN-THE FIELD FOR COLLECTING COVARIATES AS COMPLEMENTS TO RCTs AND FOR TARGETING POLICY

One criticism of randomized control trials (RCTs) is that they are often limited in discerning the mechanism driving the observed results. Because researchers are often

Table 1 Theoretically relevant populations

Article	Population and setting	Study
Harbaugh et al. (2001)	125 Children in kindergarten, third grade, and fifth grade, and 38 undergraduates in classrooms	Testing whether endowment effect changes with age/market experience.
Harbaugh et al. (2001)	Seven-year and 11-year-old children and college undergraduates in classrooms	Testing whether age affects rationality and consistency of preferences in line with GARP.
Alevy et al. (2007)	Market professionals at the Chicago Board of Trade and college students in lab	Testing for differences in cascade behavior and herding between students and market professionals.
Dohmen et al. (2012)	Families—children and their parents—interviewed at their homes	Testing whether willingness to take risks and trust are inherited from parents.
Frijters et al. (2015)	Chinese migrants interviewed in hotel rooms and over the phone	Examining selection bias for lab in field studies conducted on representative population of migrants versus self-selected population of migrants.
Marette et al. (2011)	201 Households—with women between 25 and 35 years old, with at least one child under 15, who eat fish at least 2× a week. Interviews conducted in home and preferences elicited at market	Welfare effects of regulatory tools such as labels and/or taxes.
Grossman and Baldassarri (2013)	2597 Ugandan farmers in rural communities	Tested whether group attachment and relative position in social networks affects prosocial behavior towards in-group.
Gilligan et al. (2014)	Residents in conflict-plagued regions	Used incentivized behavioral activities to measure Nepal communities' social capital. Took advantage of Nepal's natural landscape to study communities which are exposed to uncertainty of violence.

Table 1 Theoretically relevant populations—cont'd

Article	Population and setting	Study
Spears (2010)	Informal day market laborers in Rajasthan, India	Studies whether poverty causes impulsive behavior through a "store" game and behavioral test. Test was designed to mimic analogous decisions in the real world.
Chandrasekhar et al. (2014)	Villagers in Karnataka, India	Studies how real-world social networks may substitute for formal contract enforcement by conducting experiments in villages. Imitate real-world relation network as subjects have real-world relationships with each other. These relationships were observable from available detailed social network data for each household in the village.
Attanasio et al. (2012)	Residents in Columbia	Studies how risk-sharing group formation is affected by pre-existing social network and individual's risk attitude. Real-world relations were studied as friendship and kinship already existed among participants, many of whom came from the same community.
Binzel and Fehr (2013)	Residents in Cairo, Egypt	Studies how prosocial behavior is influenced by people's social distance and anonymity by conducting dictator game in Cairo communities. Utilizes pre-existing social relations to mimic real-world social networks.
Alexander and Christia (2011)	Students from Mostar, Bosnia-Herzegovina	Studies the effect of ethnic diversity on cooperation. Subjects were drawn from populations that have historically been in conflict (Croats and Bosnians).
Charness and Villeval (2009)	Juniors (under 30) and seniors (over 50) at two large French firms	Examine differences in competitiveness and cooperation amongst younger and older individuals.

limited in the number of treatments they can run due to costs or access to the required sample size, RCTs frequently cannot identify a particular theoretical model while ruling out alternative explanations on their own (Viceisza, forthcoming). In this section, we build the case for using lab-in-the-field methodology to collect explanatory covariates that can be used in conjunction with RCTs to inform researchers about the mechanism leading to the observed behavior. In general, lab-in-the-field can be used in two complimentary ways when conducting an RCT. First, lab-in-the-field could be used as a part of the RCT baseline, which allows the researcher to study whether treatment effects depend on behavior measured in experiments. Second, the researcher can conduct a lab-in-the-field as a part of the outcome of the RCT, studying whether the intervention affected related behavior, such as overconfidence or competitiveness.

Ashraf et al. (2006) ran an RCT exploring the effectiveness of commitment devices on savings behavior. The authors offered some bank clients in the Philippines commitment savings account with limited access to deposits. The account was designed to be appealing to clients who are sophisticated about their self-control problems, and in turn, would like to restrict access to their savings in order to limit impulsive purchases. Of the 710 clients who were offered this account, 202 (28.4%) opted in. Moreover, the intervention successfully increased savings: for those clients offered the commitment savings account, the average savings balances increased by 81% in the following year relative to clients who were in the control group and not offered the account.

While the nonzero take-up rate of the commitment savings device is interesting in its own right, it is not sufficient evidence for the hypothesis that people chose the account in order to overcome their self-control problems. In order to provide support for this mechanism, Ashraf et al. (2006) conducted a time discounting experiment before the RCT. They report that clients who exhibited lower discount rate for future relative to current trade-offs in the experiment were, consistent with the hypothesis, more likely to choose the commitment savings account.

Jakiela et al. (2015) sought to examine the causal impact of an education intervention on cultural values, norms, and social preferences. Although many have claimed that human capital gains lead to more equitable attitudes and support for democratic institutions, causal evidence for this relationship was largely absent. A Dutch NGO ICS Africa introduced a scholarship program to a random sample of sixth grade girls in Western Kenya. The program led to significant improvements in performance on academic tests relative to the control group (Kremer, 2009). Jakiela et al. (2015) ran a lab-in-the-field experiment with girls who were in the treatment group of the Girls Scholarship Program (GSP) and girls who were in the control group, and hence did not receive access to the scholarship program. Both groups participate in a modified dictator game where the dictator decides how to allocate money between themselves and another individual. Critical to the design, the sum to be divided was earned by

the latter individual through a real effort task, creating an informal "property rights" over the money.

The authors found that participants in the treatment group exhibited superior academic performance relative to the control group. Moreover, the GSP group allocated significantly more to the other, with a greater shift towards 50-50 splits, than girls who were not in the program. These results suggest that the randomized education intervention had medium to long-run effects not only on academic performance but on social preferences and cultural values as well.

In a study on microentrepreneurship, Berge et al. (2015a) examine the effects of business training and business grants on economic outcomes such as business performance, practice, and investment. The authors collaborated with an established microfinance institution in running the RCT which recruited small scale entrepreneurs in Dar es Salaam, Tanzania. Participants were randomized to either be enrolled in a business training course focusing on basic principles such as customer service, pricing, and accounting, or to receive a business grant equivalent to the cost of the training course; performance of both groups were compared to a control. In conjunction with the RCT, lab-in-the-field experiments were run to elicit participants' risk and competitiveness preferences, as well as their confidence and willingness to share information.

The authors found that business training had a significant and positive effect on outcomes such as sales, profit, and reported happiness—but only for the male entrepreneurs; business grants had no significant impact for either. Exploring this gender effect further using collected lab-in-the-field measures revealed that female participants were less willing to share income information with their spouse, suggesting a significant levy on their earnings that may make business expansion less worthwhile. Additionally, they were less competitive than their male counterparts, which the authors argue is an important factor in the entrepreneurial mind-set.

In order to optimally target policy and RCT interventions towards those who are most likely to experience positive impact, it is important to understand what traits and preference measures are associated with economic behavior. Burks et al. (2008) examined how elicited measures relate to and influence labor outcomes, specifically job attachment. A sample of 1000 trainee truckers at a company operated training facility took part in the study on how cognitive skills (CS) affect economic preferences and behavior. The authors elicited three measures of CS (IQ, planning ability, quantitative literacy) from each individual and examined the relationship between CS and standard measures of economic preferences (choice consistency, time and risk preferences). The lab-in-the-field method allowed them to examine how CS relates to actual economic behavior by linking the elicited measures to human resource records and the relationship between the measures and job attachment. CS was found to have a positive and significant correlation with patience and the willingness to take calculated risks.

Importantly, higher CS, particularly in the ability to plan, was significantly related to job attachment: participants who displayed better abilities to plan stayed at the job longer, which was profitable for the company. By using the lab-in-the-field methodology to link experimentally elicited measures to real-world behavior, these findings are able to inform policy by highlighting a series of traits relevant for labor market outcomes.

Another important use of the lab-in-the-field methodology is to examine how the environment and prior experiences of population shape their preferences in order to improve the targeting of policy measures and RCT interventions. Bchir and Willinger (2013) exploit natural variation in the potential for lahars (mudflows from volcanoes) in Arequipa, Peru to examine how living with greater ex ante background risk affects preferences for financial risk. The authors utilize a commonly employed method of eliciting risk preferences in the lab, a multiple price list over lotteries (Holt and Laury, 2002; see Charness et al., 2013 for review), to compare the preferences of individuals living in high-risk areas to those living with lower levels of background risk. In this method, individuals make a series of decisions between safer lotteries with smaller variances and riskier lotteries with greater variances; an individual's risk attitude is measured by the number of times he or she chooses the safer option. The authors find that, contrary to standard economic intuition, individuals living with greater background were more risk seeking than those in less exposed areas. However, this difference only held for low-income participants—there was no significant relationship between lahar exposure and risk preferences amongst those with higher incomes.

Eckel et al. (2009) document an analogous relationship between natural hazards and risk attitudes for individuals who experienced a natural disaster versus those who did not. Particularly, they elicited risk attitudes from a sample of individuals being evacuated from the aftermath of Hurricane Katrina and compared their responses to a similar group of people who did not experience the disaster. Risk preferences were measured using the Eckel and Grossman (2002) method which offered individuals a choice between six lotteries that differed in their expected return and variance; a given lottery choice could be used to classify the individual as risk-averse, risk neutral, or risk-seeking. Eckel et al. (2009) found that those who had experienced Hurricane Katrina were significantly more risk-seeking than the comparison group.

Similarly, Voors et al. (2012) examined how prior exposure to violence on the community level-shaped risk preferences. The authors identified communities in Burundi who had been exposed to violent conflict and matched them to comparable communities who were not exposed to the conflict. Individuals in both groups were asked to make choices between safe and risky lotteries in a multiple price list format. They found that, similar to exposure to natural disasters, exposure to violence also leads individuals to make riskier choices.

To summarize, collecting covariates using the lab-in-the-field methodology as part of an RCT helps in two ways. First, it can help identify the theoretical mechanism driving

Table 2 Lab-in-the-field as compliment to randomized control trials

Article	Population and setting	Study
Ashraf et al. (2006)	Banking customers in the Philippines	Significant portion of customers offered commitment savings vehicle limiting their access to cash opted in, especially those with higher elicited discount rates.
Berge et al. (2015a)	Microentrepreneurs in Tanzania	Evaluated effect of business training and business grants on performance. Business training had significant short and long-term effects for male entrepreneurs, business grants did not. Separately elicited measures suggest null effect for female entrepreneurs due to greater spousal levies and lower competitiveness.
Jakiela et al. (2015)	School-aged Kenyan women	Sixth-grade girls randomly assigned to girls scholarship program (GSP), explored whether intervention changed social preferences and norms. More educated girls found to be less likely to appropriate others' income and adhere to fair financial allocations.
Lahno et al. (2015)	Pairs of individuals in 30 villages in rural Uganda	Studies the external validity of elicited risk measures in predicting interpersonal conflict. Finds risk-aversion per se does not explain level of conflict but rather differences in risk attitudes strongly predict interpersonal conflict.
Voors et al. (2012)	Villagers from Burundi	Studies how exposure to violence affects risk preferences.

Continued

Table 2 Lab-in-the-field as compliment to randomized control trials—cont'd

Article	Population and setting	Study
Bchir and Willinger (2013)	Communities in Arequipa, Peru	Studies how differing exposure to background risk in the form of mudslides affects risk preferences.
Eckel et al. (2009)	Individuals who were evacuated after Hurricane Katrina	Studies how exposure to natural disasters affects risk attitudes.
Burks et al. (2008)	1000 Trainee truckers at company operated trainee facility.	Studies effect of cognitive skills on three tests of preferences, strategic behavior, and perseverance in the job.
Ward and Singh (2015)	Farmers in rural India	Examines how elicited risk preferences, loss, and ambiguity aversion relate to the propensity to adopt new risk-reducing farming technologies.
Liu (2013)	Farmers in rural China	Studies how elicited risk preferences, loss aversion, and probability weighting relate to adoption of new agricultural biotechnology.
Karlan (2005)	Individuals in rural Peru	Examines how elicited trustworthiness in a trust game predicts propensity to default on microfinance loans.

the success or failure of the program. Second, this data can assist policy makers in targeting future interventions to participants who are most likely to adopt/benefit from it (Table 2).

3. COMPARING BETWEEN CONTEXTS AND CULTURES

Another benefit of the lab-in-the-field methodology relative to other methods is the ability to make direct comparisons between different populations and contexts. This advantage is exemplified in studies examining the role of culture on decision-making. Henrich et al. (2006) study whether willingness to engage in costly punishment is universal amongst cultures, arguing that a possible mechanism for such cooperation could be the use of costly punishment of defectors. To test this conjecture, they compare

the use of costly punishment between industrialized (using a standard student population) and nonindustrialized populations.

A total of 1762 adults in 15 different societies participated in the experiment. Populations ranged from Western educated students at Emory University to nomadic adults in the Amazon. Each individual participated in three games aimed to capture willingness to engage in costly punishment and altruism. In the Ultimatum Game, one participant was endowed with a day's wage and chose how to split it with his or her partner. The partner could engage in costly punishment by rejecting allocations deemed too low—this would result in both players getting nothing. In the Third-Party Punishment game, participants observed the dictator game allocation decisions of another pair and could sacrifice part of his or her endowment to punish a greedy Dictator. Last, all participants played the dictator game where they decided how to split a sum of money between themselves and another participant (who did not have a choice).

Henrich et al. (2006) found substantial costly punishment in every culture. In the Ultimatum game, willingness to reject an offer decreased as the size of the offer increased from 0% to 50% of the endowed cash. Rejection rates differed substantially by population: in some societies only 15% were willing to reject a low offer while in others 60% were willing to reject. A similar pattern was found in the Third-Party Punishment game: all societies were willing to punish low offers to some extent, but this punishment rate ranged from 28% in Tsimane to 90% in Gusii. In each society, punishment rates in both games were highly correlated with each other as well as the measure of altruism in the dictator game.

Examining the data set of Herrmann et al. (2008) which used the standardized protocol of a public goods game across 16 subject pools and six distinct cultures, Gachter et al. (2010) analyze rates of contribution and cooperation between cultures in a public goods game with and without punishment. They find little variation in behavior amongst the subject pools within a culture. Consistent with prior findings (e.g., Gachter and Fehr, 2000), contributions were positive and dropped significantly at the end of the game. However, contribution rates as well as responses to the ability to punish differed significantly between cultures. Contributions in English-speaking cultures and Protestant Europe were higher than in Southern Europe and the Arab-speaking cultures. Additionally, English-speaking, Protestant Europe, and Confucian cultures contributed significantly more when players had the ability to punish free riders while those in Southern Europe, Arab-speaking, and ex-communist cultures did not respond to the potential to punish others.

By using the same experimental methodology across a variety of cultures, researchers were able to make direct comparisons between how social preferences developed in each of the societies studied. Despite similar social standing within their respective societies, individuals made vastly different choices in their willingness to cooperate with others, share resources, and punish defectors. This suggests that environmental factors and the

culture in which individuals develop have a critical influence on how they interact with others. Particularly, the presence of stable institutions and effective means of sanctioning violators of social norms appear to play a key role in people's willingness to engage in costly behavior that is beneficial for others. These findings have significant implications for the development of policy and interventions aimed at fostering such behavior.

In some cases, lab-in-the-field is useful to test a hypothesis regarding parameters that cannot be randomized in the lab. For example, Gneezy et al. (2009) examined whether culture influences the gender gap in willingness to compete or if the gap was due to innate differences in preferences. Gneezy et al. (2003) and Niederle and Vesterlund (2007) showed that women react less to competitive incentives and are significantly less likely to enter competitions than men even when their ability and performance would have allowed them to win.

This gender difference in preference with respect to competitiveness has been replicated many times in laboratory experiments (see Croson and Gneezy, 2009 for review). However, it is impossible to know from these experiments if the difference in preferences originated from innate biological differences between men and women ("nature") or due to the culture men and women are raised at ("nurture"). In order to disentangle the two explanations, Gneezy et al. (2009) examined gender differences in competitive preferences between a patriarchal society in Tanzania (the Maasai) and a matrilineal society in India (the Khasi). The Khasi tribe is special because it is organized around the women who own the property and make many of the substantive decisions. Participants in the experiment were asked to choose between a piece rate per success (landing a tennis ball in a basket 3 meters away) or compete with others on the number of successful tosses such that the winner would get three times more per success than in the piece rate payment and the losers would get nothing (in case of a tie both participants were paid the same as in the piece rate).

Results revealed that similar to gender differences in the west, Maasai men were significantly more likely to choose competition over piece rate than the women. However, this gap disappeared for the Khasi—women were just as likely to compete as men. The results were robust to a variety of controls including separately elicited risk attitudes. These findings suggest that culture could affect gender differences in preferences up to a point of eliminating them.

Hoffman et al. (2011) similarly examine the effect of culture on the gender gap in spacial ability. Voyer et al. (1995) demonstrate that women perform significantly worse than men on tasks requiring spacial reasoning. Spacial ability is related to performance on engineering and problem-solving tasks (Poole and Stanley, 1972), and the gender gap in these abilities has been used to explain the relative dearth of women in science jobs (Spelke and Pinker, 2005). Hoffman et al. (2011) tested whether the gender gap was

Table 3 Comparing between contexts

Article	Population and setting	Study
Henrich et al. (2006)	Random sample across 15 diverse populations around the world	Willingness to engage in costly punishment
Gachter et al. (2010)	120 Participants across 6 different cultures	Willingness to contribute and cooperate in public goods games with and without punishment
Herrman et al. (2008)	120 Participants across 6 different cultures	Willingness to engage in antisocial punishment in public goods games
Gneezy et al. (2009)	Members of the patrilineal Maasai tribe and the matrilineal Khasi tribe	Whether gender gap in competitive preferences is due to nature versus nurture
Hoffman et al. (2011)	Members of the patrilineal Karbi tribe and the matrilineal Khasi tribe	Whether gender gap in spacial ability is due to nature versus nurture
Hui et al. (2004)	33 Nations for study 1, Canada and People's Republic of China for studies 2 and 3	How cultural perceptions of power moderates the effect of empowerment on job satisfaction
Jakiela et al. (2015)	Residents of rural villages in western Kenya	Social preferences governing distribution of earned versus unearned income, finds unlike with western student samples, village populations do not distinguish between earned and unearned income

due to nature versus nurture by having two genetically similar participant pools (the Khasi and the Karbi) complete a puzzle task involving special abilities. Importantly, as described above, the Khasi are a matrilineal tribe while the Karbi are patriarchal. The authors found a strong and significant gender gap amongst the Karbi where men were more successful in solving the puzzle than women. However, there was no significant gender gap in the Khasi. The results were robust to a variety of controls such as education and income.

By comparing performance on the same task across different cultures, these findings suggest that like the gender gap in competitiveness, the gap in spacial reasoning is largely influenced by nurture rather than nature. If the gap in performance and preference is due to cultural and environmental factors rather than innate differences between genders, this leaves room for policy and external interventions aimed at closing that gap (Table 3).

4. EXTERNAL VALIDITY

A common concern with traditional lab experiments is whether findings would generalize to the relevant environments and contexts. Take, for example, the gift exchange model of labor contracts first proposed by Akerlof (1982). In the model, firms pay wages above the market-clearing rate in expectation that workers will reciprocate the higher wages by putting in greater effort. Fehr et al. (1993) provided an early test of the model by randomizing participants into the role of employer and employee in the lab. The employer's earnings were based on an exogenously assigned profit function of the employee's chosen level of effort minus the wage paid to them. The employee's earnings were calculated as the wage offered by the employer minus the effort cost, which was also determined by an exogenous function. The task proceeded with the employer choosing a number corresponding to the wage and the employee responding by either accepting the wage and choosing a number corresponding to effort or rejecting the wage contract. The authors found that higher wage offers were reciprocated with higher choices of effort—suggesting evidence for gift exchange. A very large literature based on lab experiments replicated and extended these early findings (Charness and Kuhn, 2011)

Gneezy and List (2006) studied gift exchange by examining whether employees reciprocated higher-wage offers by putting in greater effort. However, unlike Fehr et al. (1993), the authors used a lab-in-the-field setting where employees were recruited to complete an assignment and chose how much real effort to exert for a certain wage. Employees were recruited to perform actual work on a task for a specified amount of time at a wage of $12 an hour. When the employees arrived to complete the task, one group was told that instead of being paid $12 an hour, they would instead be paid $20. A second group worked for the expected wage. The authors found that although employees in the first group started out working harder than the second, the effort of the two groups quickly converged. The employers in the experiment would have been better off paying the market-clearing wage rather than attempting to encourage reciprocity by offering a higher wage.

In order to explore the external validity of experimentally elicited risk attitudes, Hanoch et al. (2006) studied the domain specificity of willingness to take risk or how people's perception and chosen course of action in dealing with risk vary depending on the domain: a person may appear risk seeking in one domain (finance) but risk averse in another (sports). The type of risk spans across different domains—divers and bungee jumpers in the recreational domain, gym members from the health-conscious domain, smokers from health-risk domain, casino visitors from the gambling domain, and stock traders from investment domain.

Hanoch et al. (2006) elicited risk perception and likelihood of engaging in risky activity across these domains. The results suggest that the domain-specific elicitation method is externally valid since it correlates with actual risk-taking in that domain by the target population. Moreover, risk attitudes themselves appear domain specific: risk taking in one domain does not appear to be correlated with risk taking in another. For example, gamblers who are risk seeking in casinos are not necessarily risk seeking in the health and recreation domains. The authors conclude that a general measure of risk fails to capture people's behavior across domains, and as such both theory and experiments should utilize more domain-specific measures.

Dohmen et al. (2012) explore a similar question of what measure of risk is optimal for predicting and describing behavior. They study how risk-taking propensity is affected by various biological and socioeconomic factors such as gender, age, height, and family background, and examine the stability of elicited risk attitudes across domains of real-life behavior.

The data used was from a national survey, the German Socio-Economic Panel (SOEP), which collects data from a large, representative sample. The survey asked general risk questions about people's willingness to take risk and recorded information on savings, investment behavior, health expenditures, etc. Responses were not incentivized. The authors conducted a complementary experiment where participants' answers on the SOEP survey could be compared with choices on standardized experimental paradigms used in the literature to elicit preferences in an incentive compatible manner.

The results suggest that incentivized lottery experiments typically used to elicit risk attitudes lack predictive power over the unincentivized general survey questions in predicting relevant real-world behavior such as investment choices. Similar to Hanoch et al. (2006), Dohmen et al. (2012) find that domain-specific questions are best at predicting risky behavior in the respective domain. Additionally, the general risk question that consists of a scale representing how willing participants are to take on risk in general explains a substantial amount of variance across domains of risky behavior, outperforming the incentivized lottery task. By using the lab-in-the-field methodology, the authors were able to directly test the external validity of commonly used measures of risk preference, finding that the general and domain specific questions to be more representative of individuals' willingness to take risk in theoretically relevant contexts.

In a similar vein, Barr and Zeitlin (2010) investigate how well measures of social preferences elicited using the dictator game reflect actual prosocial behavior in real-life "specific, naturally occurring, policy-relevant decision-making." Participants were primary school teachers in Uganda who took part in dictator game with their students' parents serving as recipients. The chosen allocation game was compared to teachers' allocation of time to teaching, which served as a real-life proxy for prosocial behavior. The results

Table 4 External validity

Article	Population and setting	Study
Insurance versus savings for the poor: why one should offer either both or none	Rural villagers in the Philippines	Study residents in developing countries' decisions regarding insurance, saving and risk-sharing. Sample is more compatible with the idea of risk sharing at the village level and strengthens external validity of results.
Galizzi and Martinez (2015)	University students and alumni (London School of Economics and Political Science)	Compare results from lab experiments, field experiments and self-reports of the past behavior to assess the external validity of social preference games.
Ligon and Schechter (2012)	Villagers in Paraguay	Studies the motive for sharing in rural villages. Participants from rural Paraguay communities so that their sharing decisions are closer to real-world results. Examined money transfer data from both the experiment and real-world record to examine external validity of experiment.
Benz and Meier (2008)	Students at the University of Zurich	Conducted donation experiments in order to compare students' behavior in games with their behavior in an unconnected decision situation about donating to social funds. Studies the relationship between participants' behavior in experiment and decisions outside the laboratory.
Hanoch et al. (2006)	Decision-makers who are regularly subjected to risks	Studies the domain specificity of risk-taking behavior. Subjects drawn from different real-life risk-taking domains for external validity.

Table 4 External validity—cont'd

Article	Population and setting	Study
Dohmen et al. (2012)	Representative sample of German population	Compared results from national survey and lab-in-the-field experiment to examine how well people's responses to general risk questions (and therefore their risk attitudes) reflect people's actual decision when facing real risks in life.
Barr and Zeitlin (2010)	Primary school teachers in Uganda	Studies the external validity of dictator game by comparing school teachers' responses in the games with their actual prosocial behavior in real life (extra time allocated to teaching).
Berge et al. (2015b)	Small-scale entrepreneurs in Salaam, Tanzania	Studies the external validity of competitiveness measures by comparing entrepreneurs' competitive choices in lab-in-the-field measures to their employment choices in the field, finding a significant correlation.

showed a weak correlation between the two measures, suggesting that behavior in the dictator game may be capturing a preference orthogonal to decisions involving allocations of time to teaching (Table 4).

5. CONCLUSION

In this paper, we consider an important element in the toolbox of researchers who are interested in behavior in the "real world"—lab-in-the-field. Such experiments can be used to inform researchers and policy makers about individual preferences, mechanism design, and best practices. Good studies take into account that, as argued by Harrison et al. (2015), "Any data generated by an experiment needs to be interpreted jointly with considerations from theory, common sense, complementary data, econometric methods and expected applications."

Using carefully designed lab-in-the-field experiments can increase our understanding of the mechanisms behind behavior, inform the design of field experiments and increase the effectiveness of policies at a relatively low cost to the researcher. This method can

bridge the gap between traditional lab experiments and field experiments, as well as take the powerful tool of lab experiments to new, theoretically relevant subject pools and environments.

Running lab-in-the-field experiments brings with it new challenges not present in traditional laboratory environments. In this section, we outline several lessons that we have learned from conducting lab-in-the-field studies.

1. *Simpler is better*

   Nonstandard populations, especially in developing countries, can exhibit large variation in their level of literacy and mathematical ability. With that in mind, it is important to develop methodology that is easy for participants to understand and mirrors real-world decision making. As an example, consider the elicitation of risk preferences from a population in rural Senegal using a series of validated methods, with some more complex than the others. Charness and Villeval (2009) found that presenting participants with a complicated list of choices between lotteries yielded inconsistent, noisy data more than 50% of the time; in contrast, a simpler measure analogous to a choice of how much resources one is willing to expose to risk yielded more precise estimates that correlated with real-world behavior.

2. *Use standardized, validated methods*

   When designing a study to take to the field, it is important to use standardized, validated methods to ensure that results are comparable and replicable across contexts. As outlined in Section 3, one of the main advantages of the lab-in-the-field methodology is the ability to compare results across contexts and cultures. Using new methods to elicit preferences in a standard student population is less of a concern because one can readily have access to a similar context and population again in order to validate the novel methodology and compare it to prior work either directly in the same experiment or across studies. In the field, however, using novel methods that have not been validated may handicap the researcher in arguing that the elicited measure corresponds to the construct of interest.

3. *Be aware of cultural factors when developing methodology*

   It is critical to be aware of cultural differences when developing the methodology for a study. A design that can be carried out easily in one context, e.g., the United States, may be impossible to implement in another, e.g., Tanzania. In turn, it may be useful to include an individual deeply familiar with the culture and environment for a given study as a part of the research process. For example, the four countries experiment of Roth et al. (1991) compared the market and bargaining behavior amongst participants in the United States, Yugoslavia, Israel, and Japan. The authors found that while market behavior was consistent between countries, bargaining behavior diverged. The research team came up with a number of design modifications in order to be able to argue that culture rather than flaws in the experimental design drove the results. For example, in order to ensure that language in the instructions was consistent between countries, the

English instructions were first translated to the respective language, e.g., Hebrew. The Hebrew instructions were then translated back to English to ensure that meaning was not lost in the translation.

For a lab-in-the-field study, one of us (Gneezy) wanted to draw blood from villagers near Shillong in India, in order to correlate biological markers with measures of competitiveness. When arriving to the research site, it turns out that the participants refused to have their blood drawn because of a belief that "people from the mountains will come over, take their blood, and feed it to snakes."

Religious factors can play a role as well. For example, for a study in Bosnia on the effects of post-war partitioning on investment behavior, Imas et al. planned to use a validated risk measure to proxy for individuals' willingness to invest that involved a financial gamble. However, financial gambling was against custom in the predominantly Muslim population so another method needed to be devised. At the end, the authors settled on a hypothetical scenario involving willingness to invest in either a risky business venture with both a high upside and downside or a safer venture involving less variance.

4. *Plan out logistics ahead of time*

When going into the field, it is vital to plan out the logistics well ahead of time and factor delays into your timeline. Obtaining the right permissions and getting the relevant people on board early is critical for successfully running a lab-in-the-field experiment.

For example, in many contexts, it will be difficult to follow-up with a population if your study involves a delayed payment. In such situations, it may be more effective to design a study where a random subset of the subjects is selected to be paid rather than taking on the impossible task of tracking everyone down. Similarly, busy people such as financial professionals may not care about the low stakes involved in typical incentivized lab experiments. Since incentivizing an experiment for the very rich would make the stakes consequential is outside the budget of most experimentalists, it may be more effective to run a hypothetical study with large stakes.

REFERENCES

Akerlof, 1982. Labor contracts as partial gift exchange. Q. J. Econ. 97 (4), 543–569.
Alevy, Haigh, List, 2007. Information cascades: evidence from a field experiment with financial market professionals. J. Finance 62 (1), 151–180.
Alexander, Christia, 2011. Context modularity of human altruism. Science 334 (6061), 1392–1395.
Almås, I., Cappelen, A.W., Salvanes, K.G., Sørensen, E.Ø., Tungodden, B., 2010. Fairness and the development of inequality acceptance. Science 328, 1176–1178. Management Science.
Almås, I., Cappelen, A.W., Salvanes, K.G., Sørensen, E.Ø., Tungodden, B., 2016. Willingness to compete: family matters. Manag. Sci. (forthcoming).
Anderson, Holt, 1997. Information cascades in the laboratory. Am. Econ. Rev. 87 (5), 847–862.
Andreoni, 1989. Giving with impure altruism: applications to charity and Ricardian equivalence. J. Polit. Econ. 97 (6), 1447–1458.

Andreoni, Miller, March, 2002. Giving according to GARP: an experimental test of the consistency of preferences for altruism. Econometrica 70 (2), 737–753.
Ashraf, N., Karlan, D., Yin, W., May 2006. Tying Odysseus to the Mast: evidence from a commitment savings product in the Philippines. Q. J. Econ. 121 (2), 635–672.
Attanasio, Barr, Cardenas, Genicot, Meghir, 2012. Risk pooling, risk preferences, and social networks. Am. Econ. J. Appl. Econ. 4 (2), 134–167.
Barr, Zeitlin, 2010. Dictator Games in the Lab and in Nature: External Validity Tested and Investigated in Ugandan Primary Schools. Mimeo.
Bchir, Willinger, 2013. Does the Exposure to Natural Hazards Affect Risk and Time Preferences? Some Insights From a Field Experiment in Peru. Mimeo.
Benartzi, Thaler, 1995. Myopic loss aversion and the equity premium puzzle. J. Econ. 110 (1), 73–92.
Bendick, Brown, Wall, 1999. No foot in the door: an experimental study of employment discrimination against older workers. J. Aging Soc. Policy 10 (4), 5–23.
Benz, Meier, 2008. Do people behave in experiments as in the Field? – Evidence from donations. Exp. Econ. 11 (3), 268–281.
Berge, L.I.O., Pires, A., Bjorvatn, K., Tungodden, B., 2015a. Competitiveness in the lab, successful in the field. J. Econ. Behav. Organ. 118, 303–317.
Berge, L.I.O., Bjorvatn, K., Tungodden, B., 2015b. The role of human and financial capital in microenterprise development: experimental evidence from Tanzania. Manag. Sci. 61 (4), 707–722.
Bikhchandani, Sharma, 2000. Herd behavior in financial markets. IMF Staff Pap. 47 (3), 279–310.
Binzel, Fehr, 2013. Giving and sorting among friends: evidence from a lab-in-the-field experiment. Econ. Lett. 121 (2), 214–217.
Burks, Carpenter, Gotte, Rustichini, 2008. Cognitive skills explain economic preferences, strategic behavior, and job attachment. PNAS 106 (19), 7745–7750.
Cappelen, A.W., Nygaard, K., Sørensen, E.Ø., Tungodden, B., 2015. Social preferences in the lab: a comparison of students and a representative population. Scand. J. Econ. 117 (4), 1306–1326.
Chandrasekhar, Kinnan, Larreguy, 2014. Social Networks as Contract Enforcement: Evidence From a Lab Experiment in the Field. NBER Working Papers, 20259.
Charness, G., Kuhn, P., February, 2011. Lab labor: what can labor economists learn in the lab? Handb. Labor Econ. 4a, 229–330.
Charness, Villeval, 2009. Cooperation and competition in intergenerational experiments in the field and laboratory. Am. Econ. Rev. 99 (3), 956–978.
Charness, Gneezy, Imas, 2013. Experimental methods: eliciting risk preferences. J. Econ. Behav. Organ. 87, 43–51.
Croson, Gneezy, 2009. Gender differences in preferences. J. Econ. Lit. 47 (2), 448–474.
Dohmen, Falk, Huffman, Sunde, U., 2012. The intergenerational transmission of risk and trust attitudes. Rev. Econ. Stud. 79 (2), 645–677.
Eckel, Grossman, 2002. Sex differences and statistical stereotyping in attitudes toward financial risk. Evol. Hum. Behav. 23 (4), 281–295.
Eckel, El-Gamal, Wilson, 2009. Risk loving after the storm: a Bayesian-Network study of Hurricane Katrina evacuees. J. Econ. Behav. Organ. 69 (2), 110–124.
Fehr, Bernhard, Rockenbach, 2008. Egalitarianism in young children. Nature 454, 1079–1083.
Fehr, Kirchsteiger, Riedl, 1993. Does fairness prevent market clearing? An experimental investigation. Q. J. Econ. 108 (2), 437–459.
Fehr, Kirchler, Weichbold, Gachter, 1998. When social norms overpower competition: gift exchange in experimental labor markets. J. Labor Econ. 16 (2), 324–351.
Frijters, Kong, Liu, 2015. Who Is Coming to the Artefactual Field Experiment? Participation Bias Among Chinese Rural Migrants. NBER Working Papers, 20953.
Forsythe, Horowitz, Savin, Sefton, 1994. Fairness in simple bargaining experiments. Games Econ. Behav. 6 (3), 347–369.
Gachter, Fehr, 2000. Cooperation and punishment in public goods experiments. Am. Econ. Rev. 90 (4), 980–994.

Gachter, Herrmann, Thoni, 2010. Culture and cooperation. Philos. Trans. R. Soc. B Biol. Sci. 365 (1553), 2651–2661.
Galizzi, Martinez, 2015. On the External Validity of Social-Preference Games: A Systematic Lab-Field Study. Working Papers 802. Barcelona Graduate School of Economics.
Gilligan, Pasquali, Samii, 2014. Civil war and social cohesion: lab-in-the-field evidence from Nepal. Am. J. Polit. Sci. 58 (3), 604–619.
Gneezy, List, 2006. Putting behavioral economics to work: testing for gift exchange in labor markets using field experiments. Econometrica 74 (5), 1365–1384.
Gneezy, Potters, 1997. An experiment on risk taking and evaluation periods. Q. J. Econ. 112 (2), 631–645.
Gneezy, Leonard, List, 2009. Gender differences in competition: evidence from a matrilineal and a patriarchal society. Econometrica 77 (5), 1637–1664.
Gneezy, Niederle, Rustichini, 2003. Performance in competiitive environments: gender differences. Q. J. Econ. 118 (3), 1049–1074.
Grossman, Baldassarri, 2013. The effect of group attachment and social position on prosocial behavior – evidence from lab-in-the-field experiments. PLoS One 8 (3), e58750.
Guth, Schmittberger, Schwarze, 1982. An experimental analysis of ultimatum bargaining. J. Econ. Behav. Organ. 3 (4), 367–388.
Henrich, Heine, Norenzayan, 2010. The weirdest people in the world? Behav. Brain Sci. 33, 61–135.
Hui, Au, Fock, 2004. Empowerment effects across cultures. J. Int. Bus. Stud. 35 (46–60), 46–60.
Haigh, List, 2005. Do professional traders exhibit myopic loss aversion? An experimental analysis. J. Finance 60 (1), 523–534.
Hanoch, Johnson, Wilke, 2006. Domain specificity in experimental measures and participant recruitment. Psychol. Sci. 17 (4), 300–304.
Harbaugh, Krause, Berry, 2001a. GARP for kids: on the development of rational choice behavior. Am. Econ. Rev. 91 (5), 1539–1545.
Harbaugh, Krause, Vesterlund, 2001b. Are adults better behaved than children? Age, experience, and the endowment effect. Econ. Lett. 70 (2), 175–181.
Harrison, G., Lau, M., Williams, 2002. Estimating individual discount rates for Denmark: a field experiment. Am. Econ. Rev. 925, 1606–1617.
Henrich, et al., 2006. Costly punishment across human societies. Science 312 (5781), 1767–1770.
Harrison, G.W., List, J.A., 2004. Field experiments. J. Econ. Lit. 42 (4), 1009–1055.
Harrison, G.W., Lau, M.I., Rutström, E.E., 2015. Theory, experimental design and econometrics are complementary. In: Frechette, G., Schotter, A. (Eds.), Methods of Modern Experimental Economics. Oxford University Press, Oxford, UK.
Herrmann, Thoni, Gachter, 2008. Antisocial punishment across societies. Science 319 (5868), 1362–1367.
Hoffman, Gneezy, List, 2011. Nurture affects gender differences in spatial abilities. PNAS 108 (306), 14786–14788.
Holt, Laury, 2002. Risk aversion and incentive effects. Am. Econ. Rev. 92 (5), 1644–1655.
Jakiela, P., Miguel, E., te Velde, V.L., 2015. Youve earned it: estimating the impact of human capital on social preferences. Exp. Econ. 18 (3), 385–407.
Karlan, December 2005. Using experimental economics to measure social capital and predict real financial decisions. Am. Econ. Rev. 95 (5), 1688–1699.
Karlan, List, 2007. Does price matter in charitable giving? Evidence from a large-scale natural field experiment. Am. Econ. Rev. 97 (5), 1774–1793.
Knetsch, 1989. The endowment effect and evidence of nonreversible indifference curves. Am. Econ. Rev. 79, 1277–1284.
Kovalchick, et al., 2005. Aging and decision-making: a comparison between neurologically healthy elderly and young individuals. J. Econ. Behav. Organ. 58 (1), 79–94.
Kremer, M., Miguel, E., Thornton, R., 2009. Incentives to learn. Rev. Econ. Stat. 91 (3), 437–456.
Lahno, A., Serra-Garcia, M., D'Exelle, B., Verschoor, A., 2015. Conflicting risk attitudes. J. Econ. Behav. Organ. 118, 136–149.
Ligon, Schechter, 2012. Motives for sharing in social networks. J. Dev. Econ. 99 (1), 13–26.

Liu, 2013. Time to change what to sow: risk preferences and technology adoption decisions of cotton farmers in China. Rev. Econ. Stat. 95 (4), 1386–1403.
Locke, Mann, 2005. Professional trader discipline and trade disposition. J. Financ. Econ. 76 (2), 401–444.
Marette, Roosen, Blanchemanche, 2011. The combination of lab and field experiments for benefit-cost analysis. J. Benefit-Cost Anal. 2 (3), 1–34.
Niederle, Vesterlund, 2007. Do women stay away from competition? Do men compete too much? Q. J. Econ. 122 (3), 1067–1101.
Poole, Stanley, 1972. A factorial and predictive study of spatial abilities. Aust. J. Psychol. 24 (3), 317–320.
Roth, A.E., Prasnikar, V., Okuno-Fujiwara, M., Zamir, S., December 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. Am. Econ. Rev. 81, 1068–1095.
Spears, 2010. Economic Decision-Making in Poverty Depletes Behavioral Control. CEPS Working Paper, 213.
Spelke, Pinker, 2005. The Science of Gender in Science: A Debate. Edge Foundation.
Viceisza, A., 2016. Creating a lab in the field: economics experiments for policymaking. J. Econ. Surv. (forthcoming).
Voors, Nillesen, Verwimp, Bulte, Lensink, van Soest, 2012. Violent conflict and behavior: a field experiment in Burundi. Am. Econ. Rev. 102 (2), 941–964.
Voyer, Voyer, Bryden, 1995. Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. Psychol. Bull. 117, 250–270.
Ward, P.S., Singh, V., 2015. Using field experiments to elicit risk and ambiguity preferences: behavioural factors and the adoption of new agricultural technologies in rural India. J. Dev. Stud. 51 (6), 707–724.