Lab in the Field: Measuring Preferences in the Wild

Uri Gneezy
Rady School of Management
University of California, San Diego

Alex Imas
Social and Decision Sciences
Carnegie Mellon University

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 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.

Although we have learned quite a lot from carefully designed experiments that impose a strict structure on decision-making, it is important to explore how individuals' endogenous 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 to study social preferences may yield more insightful results than the same manipulation in an anonymous giving game.[2]

[1] See Fehr, Kirchler, Weichbold and Gachter (1998) and Gneezy and List (2006) for the qualitative difference effort and reciprocity depending on methodology used.

[2] See Andreoni and Miller (2003) and Karlan and List (2007) for qualitative differences in price sensitivities in giving depending on the methodology used.

Field experiments are conducted in naturalistic environments and typically use a non-student 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 theoretically 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, and this makes it difficult to 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 paradigm. Targeting the theoretically relevant population and setting increases the applicability of the results. Employing a standardized paradigm permits the experimenter to maintain tight control over 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. Non-standard 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. 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 non-professional 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 loses 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, Haigh, and List (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 the lab-in-the-field 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, Krause and Vesterlund (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 older individuals 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, Krause and Berry (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 preciously 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 7 already display a high degree of rationality, with a significant proportion demonstrating choices consistent with GARP. By age 11 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 such as Fehr and Schmidt (1999) and Rabin (1993) aim to capture the systematic violations of the purely selfish, money-maximizing actor typically assumed in neoclassical economics. People have been observed to share money with strangers (Forsythe, Horowitz, Savin and Sefton, 1994), sacrifice money by rejecting unfair offers in ultimatum bargaining games (Guth, Schmittberger and Schwarze, 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, Bernhard and Rockenbach (2008) sought to answer this question by examining the allocation decisions of young children. The authors recruited groups of children aged 3-4, 5-6 and 7-8 at local pre and elementary schools. 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 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.

Dohmen, et al. (2012) jointly elicit preferences of both children and their parents, examining 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, Brown and Wall, 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.

The authors 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 with respect to models of rational choice and social preferences 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.

Another reason to conduct experiments with special population is to understand how background characteristics influence real world behavior. Burks et al. (2009) used a sample of 1000 trainee truckers at a company operated training facility to study how cognitive skills (CS) affect economic preferences and behavior. They 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. Those with higher CS scores also displayed greater strategic sophistication in sequential Prisoner's Dilemma games. 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 suggesting that interventions aimed at fostering cognitive skills may have a significant positive impact on human capital accumulation and employment outcomes.

In development contexts, in order to design effective policy and interventions it is important to understand how the environment and prior experiences of the target population have shaped their preferences. 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, 2001; see Charness, Gneezy and Imas, 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, El-Gamal and Wilson (2009) document an analogous relationship between natural hazards and risk attitudes for individuals who experienced a natural disaster versus those who did not. Particularly, the authors 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-Grossman (Eckel and Grossman, 2002) method which offered individuals a choice between 6 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. This method has been used to demonstrate gender differences in risk attitudes in a standard student sample (Eckel and

Grossman, 2008), as well as to elicit the preferences of French farmers (Reynaud and Couture, 2010). Eckel, El-Gamal and Wilson (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. The result suggest that, similar to exposure to natural disasters, exposure to violence also leads individuals to make riskier choices.

Table 1: Non-Standard Populations

Article	**Population and Setting**	**Study**
Burks et al. (2009)	1000 trainee truckers at company operated trainee facility.	Effect of cognitive skills on three tests of preferences, strategic behavior and perseverance in the job.
Harbaugh, Krause and Vesterlund (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, Krause and Berry (2001)	7 and 11 year old children and college undergraduates in classrooms	Testing whether age affects rationality and consistency of preferences in line with GARP
Alevy, Haigh and List (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, Kong and Liu (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, Roosen and Blanchemanche (2011)	201 households – with women between 25 and 35 years old, with at least one child under 15, who eat fish at least 2x 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 (2015)	2,597 Ugandan farmers in rural communities	Tested the whether group attachment and relative position in social networks affects prosocial behavior towards in-group
Gilligan, Pasquale and Samii (2011)	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
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, Kinnan and Larreguy (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, Barr, Cardenas, Genicot and Meghir (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 pro-social 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
Voors, Nillesen, Verwimp, Bulte, Lensink, van Soest (2012)	Villagers from Burundi	Studies how exposure to violence affects risk preferences.
Bchir and Willinger (2013)	Communities in Arequipa, Peru	Studies how differing exposure to background risk in the form of mudslides affects risk preferences
Eckel, El-Gamal, and Wilson (2009)	Individuals who were evacuated after Hurricane Katrina	Studies how exposure to natural disasters affects risk attitudes.

2. Comparing between contexts

A benefit of the lab in the field methodology 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 punish 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 (as the standard student population) and non-industrialized populations.

A total of 1,762 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 3 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 decided on 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. Lastly, 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 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 rates 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 6 distinct cultures, Gachter, Herrmann and Thoni (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 response 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 on the 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, Leonard and List (2009) examined whether culture influences the gender gap in willingness to compete or if the gap was due to innate differences in preferences. Gneezy, Niederle and Rustichini (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, Leonard and List (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.

Results revealed that similar to 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, Gneezy and List (2011) similarly examine the effect of culture on the gender gap in spacial ability. Voyer, Voyer and Bryden (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, Gneezy and List (2011) tested whether the gender gap was due to nature versus nurture by having two genetically similar participants' pools (the Khasi and the Karbi) complete a puzzle task involving special abilities. Importantly, 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 2: 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, Herrmann and Thoni (2010)	120 participants across 6 different cultures	Willingness to contribute and cooperate in public goods games with and without punishment
Herrman, Thoni and Gachter (2008)	120 participants across 6 different cultures	Willingness to engage in antisocial punishment in public goods games
Gneezy, Leonard and List (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, Gneezy and List (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, Au and Fock (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

3. 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, Kirchsteiger and Riedel (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.

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, Johnson and Wilke (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.

The authors elicited risk perception and likelihood of engaging in risky activity across a wide array of 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 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 socio-economic 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 particiapnts' 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, Johnson and Wilke (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, e.g. "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 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 3: 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 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, Johnson and Wilke (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.

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).

References

Akerlof (1982). Labor Contracts as Partial Gift Exchange. *The Quarterly Journal of Economics, 97*(4), 543-569.

Alevy, Haigh, and List (2007). Information Cascades: Evidence from a Field Experiment with Financial Market Professionals. *Journal of Finance, 62*(1), 151-180.

Alexander and Christia (2011). Context Modularity of Human Altruism. *Science*, *334*(6061), 1392-1395.

Anderson and Holt (1997). Information Cascades in the Laboratory. *The American Economic Review, 87*(5), 847-862.

Andreoni (1989). Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence. *Journal of Political Economy, 97*(6), 1447-1458.

Andreoni and Miller (2002). Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism. *Econometrica, 70*(2), 737-753.

Attanasio, Barr, Cardenas, Genicot, and Meghir (2012). Risk Pooling, Risk Preferences, and Social Networks. *American Economic Journal: Applied Economics, 4*(2), 134-167.

Barr and Zeitlin (2010). Dictator Games in the Lab and in Nature: External Validity Tested and Investigated in Ugandan Primary Schools. *mimeo*

Bchir and Willinger (2013). Does the exposure to natural hazards affect risk and time preferences? Some insights from a field experiment in Peru. *mimeo.*

Benartzi and Thaler (1995). Myopic Loss Aversion and the Equity Premium Puzzle. *The Journal of Economics, 110*(1), 73-92.

Bendick, Brown, and Wall (1999). No Foot in the Door: An Experimental Study of Employment Discrimination Against Older Workers. *Journal of Aging & Social Policy, 10*(4), 5-23.

Benz and Meier (2008). Do People Behave in Experiments as in the Field? - Evidence from Donations. *Experimental Economics, 11*(3), 268-281.

Bikhchandani and Sharma (2000). Herd Behavior in Financial Markets. *IMF Staff Papers, 47*(3), 279-310.

Binzel and Fehr (2013). Giving and sorting among friends: Evidence from a lab-in-the-field experiment. *Economic Letters, 121*(2), 214-217.

Burks, Carpenter, Gotte, and Rustichini (2008). Cognitive Skills Explain Economic Preferences, Strategic Behavior, and Job Attachment. *PNAS, 106*(19), 7745-7750.

Chandrasekhar, Kinnan, and Larreguy (2014). Social Networks as Contract Enforcement: Evidence from a Lab Experiment in the Field. *NBER Working Papers, 20259.*

Charness and Villeval (2009). Cooperation and Competition in Intergenerational Experiments in the Field and Laboratory. *American Economic Review, 99*(3), 956-78.

Charness, Gneezy, and Imas (2013). Experimental Methods: Eliciting Risk Preferences. *Journal of Economic Behavior and Organization, 87,* 43-51.

Croson and Gneezy (2009). Gender Differences in Preferences. *Journal of Economic Literature,* 47(2), 448-474.

Dohmen, Falk, Huffman, and Uwe Sunde (2012). The Intergenerational Transmission of Risk and Trust Attitudes. *Review of Economic Studies, 79*(2), 645-677.

Eckel and Grossman (2002). Sex Differences and Statistical Stereotyping in Attitudes Toward Financial Risk. *Evolution and Human Behavior, 23*(4), 281-295.

Eckel and Grossman (2008). Men, Women, and Risk Aversion: Experimental Evidence. *Handbook of Experimental Economic Results, 1,* 1061-1073.

Eckel, El-Gamal, and Wilson (2009). Risk loving after the storm: A Bayesian-Network study of Hurricane Katrina evacuees. *Journal of Economic Behavior & Organization, 69*(2), 110-124.

Fehr, Bernhard, and Rockenbach (2008). Egalitarianism in Young Children. *Nature, 454*, 1079-1083.

Fehr, Kirchsteiger, and Riedl (1993). Does Fairness Prevent Market Clearing? An Experimental Investigation. *Quarterly Journal of Economics, 108*(2), 437-459.

Fehr, Kirchler, Weichbold, and Gachter (1998). When Social Norms Overpower Competition: Gift Exchange in Experimental Labor Markets. *Journal of Labor Economics, 16*(2), 324-351.

Fehr and Schmidt (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics, 114*(3), 817-868.

Frijters, Kong, and Liu (2015). Who Is Coming to the Artefactual Field Experiment? Participation Bias among Chinese Rural Migrants. *NBER Working Papers, 20953.*

Forsythe, Horowitz, Savin, and Sefton (1994). Fairness in Simple Bargaining Experiments. *Games and Economic Behavior, 6*(3), 347-369.

Gachter and Fehr (2000). Cooperation and Punishment in Public Goods Experiments. *The American Economic Review, 90*(4), 980-994.

Gachter, Herrmann, and Thoni (2010). Culture and Cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1553), 2651-2661.

Galizzi and 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, and Samii (2014). Civil War and Social Cohesion: Lab-in-the-Field Evidence from Nepal. *American Journal of Political Science, 58*(3), 604-619.

Gneezy and List (2006). Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments. *Econometrica, 74*(5), 1365-1384.

Gneezy and Potters (1997). An Experiment on Risk Taking and Evaluation Periods. *The Quarterly Journal of Economics, 112*(2), 631-645.

Gneezy, Leonard, and List (2009). Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society. *Econometrica, 77*(5), 1637-1664.

Gneezy, Niederle, and Rustichini (2003). Performance in Competiitive Environments: Gender Differences. *The Quarterly Journal of Economics, 118*(3), 1049-1074.

Grossman and 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, and Schwarze (1982). An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior & Organization, 3*(4), 367-388.

Hui, Au, and Fock (2004). Empowerment Effects Across Cultures. *Journal of International Business Studies, 35*(46-60), 46-60.

Haigh and List (2005). Do Professional Traders Exhibit Myopic Loss Aversion? An Experimental Analysis. *The Journal of Finance, 60*(1), 523-534.

Hanoch, Johnson, and Wilke (2006). Domain specificity in experimental measures and participant recruitment. *Psychological Science, 17*(4), 300-304.

Harbaugh, Krause, and Berry (2001). GARP for Kids: On the Development of Rational Choice Behavior. *The American Economic Review, 91*(5), 1539-1545.

Harbaugh, Krause, and Vesterlund (2001). Are Adults Better Behaved than Children? Age, experience, and the endowment effect. *Economic Letters, 70*(2), 175-181.

Henrich et al. (2006). Costly Punishment Across Human Societies. *Science, 312*(5781), 1767-1770.

Herrmann, Thoni, Gachter (2008). Antisocial Punishment Across Societies. *Science, 319*(5868), 1362-1367.

Hoffman, Gneezy, and List (2011). Nurture Affects Gender Differences in Spatial Abilities. *PNAS, 108*(306), 14786-14788.

Holt and Laury (2002). Risk Aversion and Incentive Effects. *The American Economic Review, 92*(5), 1644-1655.

Karlan and List (2007). Does Price Matter in Charitable Giving? Evidence From a Large-Scale Natural Field Experiment. *The American Economic Review, 97*(5), 1774-1793.

Knetsch (1989). The Endowment Effect and Evidence of Nonreversible Indifference Curves. *The American Economic Review, 79*, 1277-1284.

Kovalchick et al. (2005). Aging and Decision-making: A Comparison between Neurologically Healthy Elderly and Young Individuals. *Journal of Economic Behavior and Organization, 58*(1), 79-94.

Landmann, Vollan, Frolich (2012). Insurance versus Savings for the Poor: Why One Should Offer Either Both or None. *Discussion Paper Series/IZA.*

Ligon and Schechter (2012). Motives for Sharing in Social Networks. *Journal of Development Economics, 99*(1), 13-26.

Locke and Mann (2005). Professional Trader Discipline and Trade Disposition. *Journal of Financial Economics, 76*(2), 401-444.

Marette, Roosen, and Blanchemanche (2011). The Combination of Lab and Field Experiments for Benefit-Cost Analysis. *Journal of Benefit-Cost Analysis, 2*(3), 1-34.

Niederle and Vesterlund (2007). Do Women Stay Away from Competition? Do Men Compete too Much? *Quarterly Journal of Economics, 122*(3), 1067-1101.

Poole and Stanley (1972). A Factorial and Predictive Study of Spatial Abilities. *Australian Journal of Psychology, 24*(3), 317-320.

Rabin (1993). Incorporating Fairness into Game Theory and Economics. *The American Economic Review, 83*(5), 1281-1302.

Reynaud and Couture (2010). Stability of Risk Preference Measures: Results from a Field Experiment on French Farmers. *Theory and Decision, 73*(2), 203-221.

Spears (2010). Economic Decision-Making in Poverty Depletes Behavioral Control. *CEPS Working Paper*, 213.

Spelke and Pinker (2005). The science of Gender in Science: A Debate. *Edge Foundation*.

Voors, Nillesen, Verwimp, Bulte, Lensink, and van Soest (2012). Violent Conflict and Behavior: A Field Experiment in Burundi. *The American Economic Review, 102*(2), 941-964.

Voyer, Voyer, and Bryden (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117,* 250-270.