Political Science Research and Methods Vol 3, No. 1, 113–131 January 2015
© *The European Political Science Association, 2014* *doi:10.1017/psrm.2014.10*

Assessing the Correspondence between Experimental Results Obtained in the Lab and Field: A Review of Recent Social Science Research*

ALEXANDER COPPOCK AND DONALD P. GREEN

A small but growing social science literature examines the correspondence between experimental results obtained in lab and field settings. This article reviews this literature and reanalyzes a set of recent experiments carried out in parallel in both the lab and field. Using a standardized format that calls attention to both the experimental estimates and the statistical uncertainty surrounding them, the study analyzes the overall pattern of lab-field correspondence, which is found to be quite strong (Spearman's ρ = 0.73). Recognizing that this correlation may be distorted by the ad hoc manner in which lab-field comparisons are constructed (as well as the selective manner in which results are reported and published), the article concludes by suggesting directions for future research, stressing in particular the need for more systematic investigation of treatment effect heterogeneity.

Lab experiments and field experiments offer complementary approaches to the study of cause and effect in the social sciences. Both methods attempt to isolate the causal influence of one or more interventions by eliminating the systematic intrusion of confounding factors. Experiments carried out in lab or field settings typically allocate subjects randomly to treatment and control groups, ensuring that those assigned to each group have the same expected potential outcomes. Apparent differences in outcomes between treatment and control groups therefore reflect either the effect of the treatment or random sampling variability.

The interpretation of experimental results, however, depends on the setting in which the study is carried out. Although the line between lab and field is sometimes blurry (Gerber and Green 2012; Harrison and List 2004), lab and field studies typically differ in terms of who the subjects are, the context in which they receive the treatments, the sort of treatments that are administered and the manner in which outcomes are measured. Field experiments tend to assess the effects of a real-world intervention on those who would ordinarily encounter it. Although field experiments frequently use surveys to assess outcomes (cf. Glennerster and Takavarasha 2013), they otherwise tend not to alert subjects to the fact that they are being studied in connection with a particular intervention, and often measure outcomes unobtrusively after the intervention occurs. Situating an experiment in a field setting risks implementation problems, either because some subjects do not receive the treatment to

* Alexander Coppock is a PhD student, Department of Political Science, Columbia University, 420 West 118th Street, Suite 710, New York, NY 10025 (ac3242@columbia.edu). Donald P. Green is Professor of Political Science, Columbia University, 420 West 118th Street, Suite 710, New York, NY 10025 (dpg2110@columbia.edu). We are grateful to Oliver Armantier, Peter Aronow, Lindsay Dolan, Uri Gneezy, John List, Jennifer Jerit, Andrej Tusicisny and two anonymous reviewers for helpful comments, Glenn Harrison and Nicholas Valentino for graciously providing replication data, and to Johannes Abeler and Ido Erev for providing supplementary information. To view supplementary material for this article, please visit http://dx.doi.org/10.1017/psrm.2014.10

which they were randomly assigned or because subjects go missing before their outcomes can be measured. Laboratory studies, on the other hand, place a premium on creating a controlled environment in which treatments can be administered and their effects observed (Morton and Williams 2010, 42). Subjects, who are often recruited from the university community, are informed that they are participating in a research study, and although they are seldom told the (true) purpose of the experiment, they are aware that their behavior is being observed. Lab sessions tend to be relatively short (an hour or less), which in turn implies that outcomes are typically assessed in the immediate wake of an experimental stimulus. There are many intermediate gradations between tightly controlled lab studies and naturalistic field studies, such as those that take place under controlled, lab-like conditions in non-university settings (for example, Habyarimana et al. 2009) and field studies that administer treatments and measure outcomes in an obtrusive manner (for example, Paluck 2009).

Although lab and field studies often share some of the same ingredients, the contrast between the two is frequently quite vivid. Consider, for example, the contrast between lab and field experiments on vote choice. Großer and Schram (2010) studied voter turnout in elections by providing undergraduate subjects with a schedule of monetary payoffs that varied depending on both the electoral outcome and each subject's private voting costs; the experimental intervention was whether some subjects' turnout decisions are observed, and by whom. Approximately 90 seconds after the intervention, the researchers measured two outcomes: whether subjects voted and, if so, for which candidate. By contrast, Gerber (2004) presents a field experiment that assesses the effects of information on voter turnout and candidate choice. Randomly selected registered voters were sent mail from an actual candidate during state legislative elections; several days later, outcomes were assessed by post-election interviews with subjects in the treatment and control groups. The subjects who received direct mail (and may or may not have read it) were unaware that they were part of a research study, and respondents to the survey were asked about their candidate preference and turnout before any mention was made of the mailings.

These two studies illustrate some of the ways in which field and lab studies may differ. The lab study relies on a convenience sample of undergraduate subjects, whereas the field study draws its subjects from the voter rolls. The lab study consists of an abstract election campaign in which the only information available to subjects is controlled by the experimenter; the field study takes place in the context of an actual election, which means that the intervention must compete with subjects' background knowledge, other messages and life's distractions. Outcomes in the two studies are measured at different points in time; the lab study gauges responses immediately after presentation of the stimulus, and the field study assesses effects days later, but with some loss of subjects due to non-response. Finally, the lab study is obtrusive in the sense that subjects are aware of the fact that researchers are studying their voting behavior; the field study's post-election interview notifies subjects that research is being conducted, but the connection to the experimental stimulus remains opaque.

Each of these design features affects the interpretation of the experimental results. Experimental outcomes may be distorted if subjects know they are being watched, especially if they perceive a connection between treatments and outcomes. Failure to control or measure how subjects receive the treatment in field settings leads to uncertainty about the meaning of an apparent treatment effect. Outcomes measured immediately after the administration of a treatment may be a poor guide to long-term outcomes. Attrition of subjects from follow-up measurement may introduce bias. Even if these methodological concerns were inconsequential, there remains the substantive concern that a lab study addresses a different type of causal effect than a corresponding field study, which assesses the effects of information in a context of competing messages and demands for voter attention. When researchers argue about the relative merits of lab and field experimentation (Camerer forthcoming; Levitt and List 2007; Falk and Heckmann 2009; Gneezy and List 2006), these issues—obtrusiveness, treatment fidelity, outcome measurement, and context-dependent treatment effects—tend to occupy center stage.

One way to address this controversy is to turn the sensitivity of results into an empirical question. In recent years a literature has emerged that assesses whether results obtained in the lab are echoed in the field and vice versa. Although no one, to our knowledge, has attempted a comprehensive assessment of the literature, two of the most prominent articles that address this question have come to different conclusions about lab-field correspondence. Levitt and List (2007) express skepticism about the potential for laboratory findings to generalize to the field, citing the ways in which a typical laboratory experiment changes the decision environment, whereas Camerer (forthcoming) stresses evidence showing agreement across the two domains.

In their critique, Levitt and List describe several characteristics of laboratory experimentation that may distort behavior relative to the field: subject pools that disproportionately feature Western undergraduates, experimenter scrutiny, the absence of moral considerations from the abstract choices presented to subjects and the small stakes for which subjects play. For each of these features, Levitt and List review lab-field comparisons, noting instances in which lab results were sensitive. They emphasize that economic theory predicts different behaviors in the lab and field, and therefore warrants skepticism: "Theory is the tool that permits us to take results from one environment to predict in another, and generalizability of laboratory evidence should be no exception" (170).

Writing in response to Levitt and List, Camerer (forthcoming) discusses six close comparisons of lab and field studies conducted in parallel settings.[1] The criteria by which Camerer judges correspondence are study specific: one pair of studies estimated effects of the same sign, another recovered similar coefficients and another displayed modest "prosociality" correlations across contexts. In addition, Camerer reviews lab-field correspondence in which the study designs are imperfectly matched or in which the subject populations are quite different and finds general agreement, concluding, "There is no replicated evidence that experimental economics lab data fail to generalize to central empirical features of field data" (Camerer forthcoming, 35)

Given these conflicting characterizations of the evidence for lab-field correspondence, we conduct a systematic assessment of the extant literature. Our aim is to investigate not only the apparent degree of correspondence, but also whether the literature as it now stands convincingly supports either side of this debate. The degree of lab-field correspondence is an important research question in its own right, as the stakes are high. All else being equal, if correspondence is strong, the marginal research dollar might be better spent in the lab than in the field. Field experimentation can be expensive, logistically challenging and ethically encumbered. If laboratory experiments can consistently predict field treatment effects, then the lab offers clear advantages, particularly the relative ease of conducting experiments and systematically extending them through replication.

[1] Two of these studies are included in our set of 12, but the remaining four failed to meet our inclusion criteria.

Further, theoretical control over incentives and behaviors might allow for the investigation of cause-and-effect relationships that lie beyond the feasible reach of field experiments. Yet when field experiments are feasible and have the potential to augment a lab-based literature in a methodologically convincing and substantively meaningful way, the extra complications of field experimentation seem worthwhile. That said, we recognize that idiosyncratic field experiments do not necessarily have a claim to greater generalizability to other (perhaps more interesting) field contexts than do well-executed lab experiments.

This article is structured as follows. We begin by describing and formalizing four key dimensions along which experiments may vary. Next, we discuss in detail the strategy by which we assembled the relevant literature, focusing in particular on the criteria by which pairs of lab and field studies were selected and analyzed. Although this collection of studies is too small to test specific theories about factors that contribute to lab-field correspondence, we can assess the overall level of lab-field agreement in the extant literature. Our statistical results reveal a surprisingly high degree of correspondence (Spearman's $\rho = 0.73$) that is robust to the inclusion or exclusion of particular studies. This correlation, however, must be interpreted with caution, given the ad hoc and selective manner in which lab-field comparisons are constructed and reported. We conclude by suggesting directions for future research, stressing in particular the need for more systematic investigation of treatment effect heterogeneity.

HYPOTHESES ABOUT CORRESPONDENCE

Imagine that two parallel experiments could be conducted with the same subjects under identical conditions. The results of the two experiments would differ only because random assignment may place a different collection of subjects into treatment and control groups. Apart from the fact that random sampling variability may cause the two parallel experiments to generate different estimates, the underlying degree of correspondence is perfect, and, in expectation, both experiments will generate precisely the same results.

As we depart from this hypothetical ideal, we confront the fact that any arbitrary pair of experiments—even when both are conducted in the lab or field—that employ different procedures may differ in countless ways. Theories of correspondence reduce this complexity to a manageable level by focusing attention on dimensions along which differences are likely to have a material effect on results. Although theories of correspondence in economics tend to be rooted in different microfoundations than comparable theories in social psychology (cf. Levitt and List 2007; Shadish, Cook and Campbell 2002), both literatures emphasize a similar set of relevant dimensions when assessing whether two or more experiments are likely to produce compatible results. Attention centers on "the extent to which the effect holds over variations in persons, settings, treatments, or outcomes" (Shadish, Cook and Campbell 2002, 22).

Subjects

When the units of observation in two distinct experiments are sampled or recruited in different ways, their measured and unmeasured traits may be quite different. These differences alone may produce divergent experimental results, and for decades a vigorous debate has raged over the question of whether lab findings would be materially affected if the subjects were drawn from less Western, less affluent and older segments of the population (Sears 1986; Henrich, Heine and Norenzayan 2010). This concern has led to an increasing number of studies that involve non-Western participants, who are frequently recruited from areas where poverty rates are high (Henrich et al. 2001; Habyarimana et al. 2009). It has been argued that lab studies involving standard economic games produce similar results across different societies (Oosterbeek, Sloof and Van de Kuilen 2004). Yet the recurrent finding that treatment effects vary depending on participants' background attributes such as education and income suggests that the subject pools used in lab and field research may be a source of systematic variation in experimental results. For example, the Asch experiment on conformity to social norms has been replicated dozens of times using many different subject pools—some studies report the same conformity finding, but most do not. Lalancette and Standing (1990) and Bond and Smith (1996) argue that subjects' characteristics account for discrepant findings across countries and eras.

One way to address concerns about systematic differences across subject pools is to randomly assign subjects with similar background attributes to parallel experiments. This design-based approach is rarely used, however. Only one of the studies discussed below (Jerit, Barabas and Clifford 2013) attempted to do so, and that study was only partially successful; subjects were initially recruited from the same population and invited to participate in lab and field experiments, but self-selection may have led different types of subjects to participate in each study. The remaining studies compare two different convenience samples. In principle, a researcher could re-weight the data from two distinct convenience samples so that both sets of participants have similar measured attributes (Hotz, Imbens and Mortimer 2005; Harder 2010). However, this method of making subject pools equivalent has important limitations. First, it does not address the problem of unmeasured differences that may persist after the data are re-weighted to achieve balance on all measured characteristics. Second, given the many measurable ways in which subject pools may differ, declaring two subject pools "similar" is a matter of judgment. It is unclear how one would convincingly re-weight the sample of Canadian university students who participated in the lab bribery study reported by Armantier and Boly (2013) to mimic their field study's subject pool of temporary workers in Burkina Faso.

Treatments

Experiments are designed to assess the effects of an intervention that some subjects receive but others do not. Depending on the experimenter's aims, an intervention may be administered in a highly controlled manner (for example, carefully worded instructions that explain to each group of players the payoffs associated with voting in a laboratory election) or more loosely (for example, a set of talking points that a door-to-door canvasser might convey while encouraging experimental subjects to vote in an upcoming municipal election). In the former case, precise control of the treatment allows the researcher to credibly claim that the laboratory study presents subjects with choices that are analogous in key respects to the choice of whether to vote in an election. In the latter case, the experimenter may be primarily concerned that the intervention falls clearly under the rubric of door-to-door canvassing; what canvassers say is less important than whether they convey the encouragement to vote in a natural, unscripted manner.

The way in which a treatment is defined and deployed has important implications for assessing the correspondence between the treatments used in two or more experiments. Sometimes experiments are analyzed together because they deploy the same intervention (for example, Arceneaux and Nickerson 2009). More often, experiments are compared because the interventions they deploy share some abstract property. For example, three recent experiments test whether enforcement of social norms increases the probability that a person will contribute to a collective good. In one field experiment, social norms about the obligation to vote were conveyed by means of a postcard reminding the recipient that voting is a matter of public record (Gerber, Green and Larimer 2008); in another field experiment, subjects were informed by door-hangers and handwritten notes about their energy consumption and the need to conserve energy (Shultz, Khazian and Zaleski 2008); and in a laboratory dictator game, subjects made allocation decisions in public or private (Charness and Schram 2013). Although the treatments are quite different, they arguably operate via the same causal mechanisms: namely, the subject's sense of obligation to uphold an injunctive norm, and concern that a failure to do so will be noticed by others.

Whether treatments are considered sufficiently analogous is a matter of theoretical perspective. In the preceding example, it is assumed that all three interventions "enforce social norms." Absent this theoretical frame, the three treatments might seem to be disjointed attempts to get people to vote, lower their energy bills or earn extra cash by participating in a group exercise. Moreover, there may be more than one way to characterize the salient theoretical features of an intervention. For example, close inspection of the messages presented to subjects may reveal that some interventions stress descriptive norms (that is, what others tend to do) while others emphasize injunctive norms (that is, what one ought to do). Likewise, some treatments signal that failure to comply with norms will tarnish one's image in the community, while others make no reference to social punishment. As noted by Cook and Campbell (1979, 56–9), any intervention comprises a limitless number of ingredients, and theory and extensive experimentation are required to isolate the types of ingredients that matter most. A seldom-noted implication of this perspective on treatment effect heterogeneity is that "null" findings suggesting that two or more treatments work equally well can be enormously helpful. Null findings simplify the task of comparison and generalization by suggesting that certain theoretically meaningful differences among treatments are not empirically consequential.

Context

The setting in which an experiment takes place may affect the way in which subjects respond to the intervention. When experiments focus on the behavioral responses of individual subjects to information or other cues in the environment, the context may determine whether subjects are attentive to the treatment. Is a television commercial presented in a lab context, in which subjects are prevented from changing channels and discouraged from checking their email, or in a field context, where these and other distractions are readily available? Another concern is the subject's mindset when exposed to the treatment. When subjects are aware that they are participating in a research study, they may try to figure out the "trick" or "right" answer. Researchers sometimes go to great lengths to hide the true purpose of the experiment in order to minimize Hawthorne effects and socially desirable answers. A classic example is Milgram's (1963) obedience study, which placed the subject in the role of the experimenter's assistant, creating the impression that the purpose of the study was to teach a "learner" to memorize word pairs. However, even when the experimental aims are successfully concealed, subjects may behave in an especially attentive manner. Much of the impetus behind unobtrusive experimental designs is to observe subjects in naturalistic settings, where treatments are part of everyday life rather than an unusual experience connected to a research project.

Outcomes

The choice of outcome measure determines the mapping of latent quantities to observable behavior. Two experiments may be said to estimate the same average treatment effect to the extent that the observable behaviors map onto the same latent quantities, possibly with a different scaling factor. Take, for example, laboratory and field experiments that aim to understand the effects of electoral competition on voter turnout. In the field, outcomes are measured by voting behavior; in the lab, they are measured in costly vote tokens. The tokens are worth money; a subject can wager a token that her party will win, and by wagering the token she also increases the probability of victory. If that event occurs, the subject receives a payoff that is worth more than the token; if not, she has "wasted her vote." Now imagine an intervention that manipulates the extent to which subjects feel a civic obligation to vote, and that this treatment is effective in *both* the lab and field. In this case, the impulse to act in accordance with civic duty is the latent quantity that both treatments have perturbed. It is nevertheless possible that this impulse affects voting in actual legislative elections but has little effect on whether subjects expend vote tokens in a laboratory setting. The lack of lab-field correspondence in this case reflects the manner in which outcomes are measured.

A less extreme scenario is one in which the same latent quantity manifests itself in both outcome measures, but the scaling factors that translate the latent quantity into observable measures are somewhat different in the two settings. Just as two physics experiments would estimate treatment effects differently if outcomes were measured in miles instead of kilometers, two social science experiments might produce seemingly discrepant results if outcomes were measured using voter turnout in one study and intention to vote in another. When researchers compare experiments that use different outcome metrics, controversy often erupts over whether discrepancies are due merely to differences in scaling or more fundamentally to the fact that each measure taps into different latent dimensions (Morton and Williams 2010, chapter 10).

Formalizing Cross-Study Differences

How do different subjects, treatments, contexts and outcomes contribute to variation in experimental results between studies? In order to appreciate the identification challenges that arise when researchers attempt to trace interstudy differences back to some or all of these elements, it is useful to consider a set of hypothetical experiments and the results they would generate in expectation.

For simplicity, we will reduce the myriad ways in which treatments, subjects, contexts and outcome measures can vary into a series of binaries. Consider a research scenario in which there are two subject types (A and B), two variants of the treatment (T1 and T2), two outcome measures (money and effort) and two contexts (lab and field).[2] Varying each dimension while holding the others constant yields a set of 16 hypothetical experiments.

Table 1 illustrates a set of hypothetical experimental effects. Each entry in the "lab" and "field" columns refers to the average treatment effect recovered from each experimental scenario. Within both lab and field, subjects, treatments and outcome measures can be combined in eight unique arrangements. Each row of the table displays a lab-field pair that holds these other experimental features constant. The within-pair difference in average

[2] We will assume that there are no other relevant contextual features beyond the setting of the experiment in the lab or field.

TABLE 1 *Average Treatment Effects in 16 Hypothetical Experiments*

Pair	Lab ATE	Field ATE	Subjects	Treatments	Outcomes
1	2.0	0.5	A	T1	money
2	5.0	4.0	A	T1	effort
3	3.0	3.0	A	T2	money
4	0.0	3.0	A	T2	effort
5	2.0	−1.5	B	T1	money
6	0.5	4.0	B	T1	effort
7	−2.0	1.5	B	T2	money
8	−2.0	2.0	B	T2	effort

Note: entries in Columns 2 and 3 are hypothetical average treatment effects recovered under the experimental conditions described in each row.

TABLE 2 *Meta-analytic Regression Predicting ATEs*

	OLS
Subjects (A)	2.00
Treatments (T1)	1.00
Outcomes (Effort)	1.00
Contexts (Field)	1.00
Constant	–0.94

treatment effects is the effect of the experimental setting on the average treatment effect. When this difference is small, lab-field correspondence is high, and when this difference is large, correspondence is low. Looking across rows, we see pairs that are similar in sign and magnitude (for example, Pair 3) and other pairs that are quite different (for example, Pair 8). The correlation between treatment effects in the lab and field serves as a summary measure of correspondence. In this example, the correlation is weakly positive (0.14).

Suppose a researcher were to conduct the full array of lab and field studies represented in Table 1, for a total of 16 experiments: the independent effects of subjects, treatments, outcomes and contexts could be gauged through meta-analysis, which would provide a quantitative assessment of lab-field correspondence. In particular, as suggested by Camerer (forthcoming, 6), a coefficient close to zero for the estimated slope of the lab-field dummy variable would indicate close lab-field correspondence. The results of such a meta-analytic regression are presented in Table 2. On average, treatment effects are two units higher when A-type subjects are used, one unit higher when treatment T1 is employed, one unit higher when effort is the outcome variable and one unit higher in the field as opposed to the lab. The non-zero coefficient on the field dummy is further evidence of weak lab-field correspondence in this example.

In our review of attempts by researchers to conduct parallel experiments in the lab and field, we found no examples of scholars systematically varying one feature of the design while keeping all other relevant characteristics constant. Instead, we found instances of researchers presenting a single lab-field pair that varied on multiple dimensions. The authors then argue with theory and qualitative evidence that the differences in subjects, treatments, outcomes and contexts are relatively minor. In many cases, the treatment effect estimates from the lab and field appear to agree, which nevertheless leaves open the question of whether the experiments are measuring the same underlying causal effect.

Stated differently, the reader must decide if a lab-field pair looks more like Row 3 of Table 1 (3, 3) or more like a combination of the lab estimate from Row 1 and the field estimate from Row 8 (2, 2). Both pairs obtain estimates that "agree," but the first pair holds constant all other relevant experimental factors, while the second pair allows the subjects, treatments and outcome measures to vary. The case for lab-field correspondence is stronger in the first pair than in the second, because we cannot be certain that the agreement is not a happy accident due to the multiplicity of factors contributing to treatment effect estimates.

The investigation of lab-field correspondence is complicated by the fact that existing studies do not systematically vary the lab and field context while holding other experimental conditions constant. Some combinations of context, subjects, treatments and outcomes are never explored, or are explored but never reported. Suppose that a researcher conducted all eight pairs of experiments but failed to report Pairs 1 and 5. The overall correlation between lab and field, originally weakly positive (0.14), would appear to be much stronger (0.76). As we consider the existing literature on lab-field correspondence, we must bear in mind that we do not observe the full set of relevant lab-field comparisons.

METHOD

In order to construct a systematic review of lab-field correspondence, we sought to (1) identify a set of lab-field comparisons, (2) define our measure of correspondence and (3) standardize analytic procedures to facilitate cross-study comparison.

Study Selection

We gathered a comprehensive set of recent studies in the social sciences that mention or reference the comparison of results across lab and field. We expanded on the many comparisons noted by Camerer and Levitt and List by following chains of citations and conducting searches on the terms "lab experiment," "field experiment," "lab-field correspondence" and "generalize from lab to field." This initial search netted approximately 80 journal articles and unpublished manuscripts, which are listed in the appendix. The investigations are from many domains: experimental economics, sociology and political science. We whittled this sample of 80 articles down to 12 in three steps:

1. Explicit pairing of studies: we kept only those studies whose authors set out to make an explicit lab-field comparison. Either the authors conducted parallel lab and field experiments themselves or they named a specific field (lab) experiment their lab (field) study was addressing. We excluded several field experiments that sought to test in the field some well-established laboratory result (loss aversion, for example) but did not name a particular comparison study.[3] In a small number of cases, we excluded studies that reported lab and field experiments that were not, in our judgment, sufficiently parallel (for example, King and Ahmad 2010).

[3] We could have chosen a representative lab study against which to assess lab-field correspondence, but the correspondence might be strong or weak depending on the comparison chosen. The main advantage of this approach is that it limits our own discretion. An alternative approach is to focus on a specific substantive domain (e.g., voter turnout) and a specific intervention (e.g., providing voters with financial rewards if they vote), reviewing all of the lab and field evidence. We regard this approach as a fruitful next step.

2. Definitions of lab and field: as noted above, the distinction between lab and field experimentation is not always clear. They may differ along a number of dimensions, including treatments, subjects, contexts and outcome measures. The "explicit pair" filter eliminated almost all borderline cases, which obviated the need for strict definitions of lab and field. When defining field experiments, we excluded "lab in the field" studies in which subjects played a laboratory game outside a university context (for example, Benz and Meier 2008).
3. Estimation of a treatment effect: the set of lab and field studies was further restricted to randomized experiments that estimated treatment effects. In other words, admissible experiments had to assess the effect of a randomly assigned manipulation. Measurement studies, by contrast, estimate the level of an outcome variable rather than a change in that variable. An example of pure measurement from the laboratory is a dictator game in which an average level of pro-sociality is estimated. In principle, a comparison of laboratory measurements and field measurements is possible and potentially informative (Benz and Meier 2008). However, similar baseline measurements would be no guarantee of similar treatment responses across lab and field. The investigation of treatment effects generally requires a fully randomized design, but we relaxed the definition of randomization to include pseudo-randomizations such as assignment according to the day on which subjects arrived in the lab.[4]

The purpose of this investigation is to examine the existing evidence on lab-field correspondence. Readers may find it useful to consult Online Appendix Table 2, which includes a detailed description of our 12 study pairs, including the subject pools, treatments, outcome measures and context of each experiment. With one exception (List 2006), the lab experiments take place in a university laboratory; the "fieldness" of each field study is open to debate. The subject pools differ substantially within each lab-field pair, though subjects are more similar in some studies (Jerit, Barabas and Clifford 2013; List 2006b) than others. Treatments are often analogous across lab and field, and are in some cases identical (Armantier and Boly 2013; Valentino, Traugott and Hutchings 2002). Outcome measures are frequently dissimilar: for example, real versus hypothetical donations (Shang and Croson 2008), food and beverage consumption versus lab production units (Gneezy, Haruvy and Yafe 2004; Abeler and Marklein 2013), or oranges picked versus contributions from a lab endowment (Erev, Bornstein and Galili 1993; Bornstein, Erev and Rosen 1990).

Data Collection and Reanalysis

The experimental results used in our analyses were gathered from publicly available datasets, correspondence with the authors, or from tables and charts in the original papers. In some cases, we corrected minor mistakes in the original analyses. Further, we standardized the presentation format so that we could assess correspondence within particular studies and across the entire set. For each of the studies, we collected mean outcomes (without covariate adjustment) in each treatment group, standard errors and group sizes. From these data, we calculated treatment effects and 95 percent confidence

[4] Several studies would have been excluded either because they did not use a fully randomized design or they did not report their randomization procedures. See Green and Tusicisny (2013) for a critique of faulty randomization procedures and inadequate reporting. There is the further issue of accounting for clustered random assignment; we use the authors' reported standard errors but recognize that these estimates probably understate the true sampling variability.

intervals using normal approximations. Most of these studies have multiple treatment groups and measure a single outcome, but one (Jerit, Barabas and Clifford 2013) has only two treatment groups and measures a large number of outcomes.

Assessing Correspondence

One issue that arises in existing debates about lab-field correspondence is the question of how to assess correspondence statistically. Intuition suggests that one should simply compare treatment effects. However, for many comparisons, outcomes are not measured using the same scale. For example, in one of the studies discussed below, the lab outcome is the number of computer mazes solved, and the field outcome is playground footrace times. There is no universally accepted technique for determining the appropriate theoretical maze-to-footspeed conversion ratio. It might be argued that treatment effects could be put in percentage terms: a 10 percent increase in mazes completed compared with a 10 percent increase in velocity. This too may have problems if small percentage changes are substantively quite large in some domains but not others.

Another technique applied by some researchers to assess correspondence is to compare the sign and statistical significance of effects: if the treatment effects are positive and significant in both lab and field, the results are said to indicate strong correspondence. This approach has two weaknesses. First, it is not clear that lab and field studies would be considered in strong agreement if they both recovered insignificant effects. Second, this approach may conflate the magnitude of the effect size with the power of the study. For example, even if the estimated effects were the same, a large field experiment might generate a significant *p*-value, whereas a small lab experiment may not.

In order to sidestep the issue of incomparable scaling and sample-size-dependent conclusions, we assess correspondence using rank-order correlations. We collect mean outcomes in each experimental group in both the lab and field studies. Spearman's ρ assesses the degree to which the ordering of means in the lab corresponds to the ordering in the field. This approach has a number of advantages. First, it is robust to the scaling problem described above. Second, it keeps the question of effect size separate from the question of statistical uncertainty. Third, it accommodates what some researchers describe as "general" or "qualitative" (Kessler and Vesterlund, forthcoming) correspondence, insofar as larger effects in the lab are associated with larger effects in the field. One weakness of our approach is that correlations tend to be exaggerated in absolute value when *N* is small (Student 1908). For this reason, our overall conclusions are based on the full set of lab-field comparisons, which we standardize for purposes of meta-analysis.

In Figure 1 below, laboratory results are plotted on the x-axis and field results are plotted on the y-axis. Perfect correspondence, by our metric, would be represented by means falling along any strictly increasing line. The numeric data used to generate these charts may be found in Online Appendix Table 1.

The 12 pairs we selected are displayed in Table 3. They cover a range of economic and political domains such as motivation and effort, social dilemmas and political attitudes. Fuller descriptions of each study can be found in Online Appendix Table 2.

RESULTS

We present the results of these 12 close comparisons in Figure 1. With the exception of the Jerit, Barabas and Clifford study, these graphs present treatment group means and 95 percent confidence intervals. Lab results are presented on the horizontal axis, and field

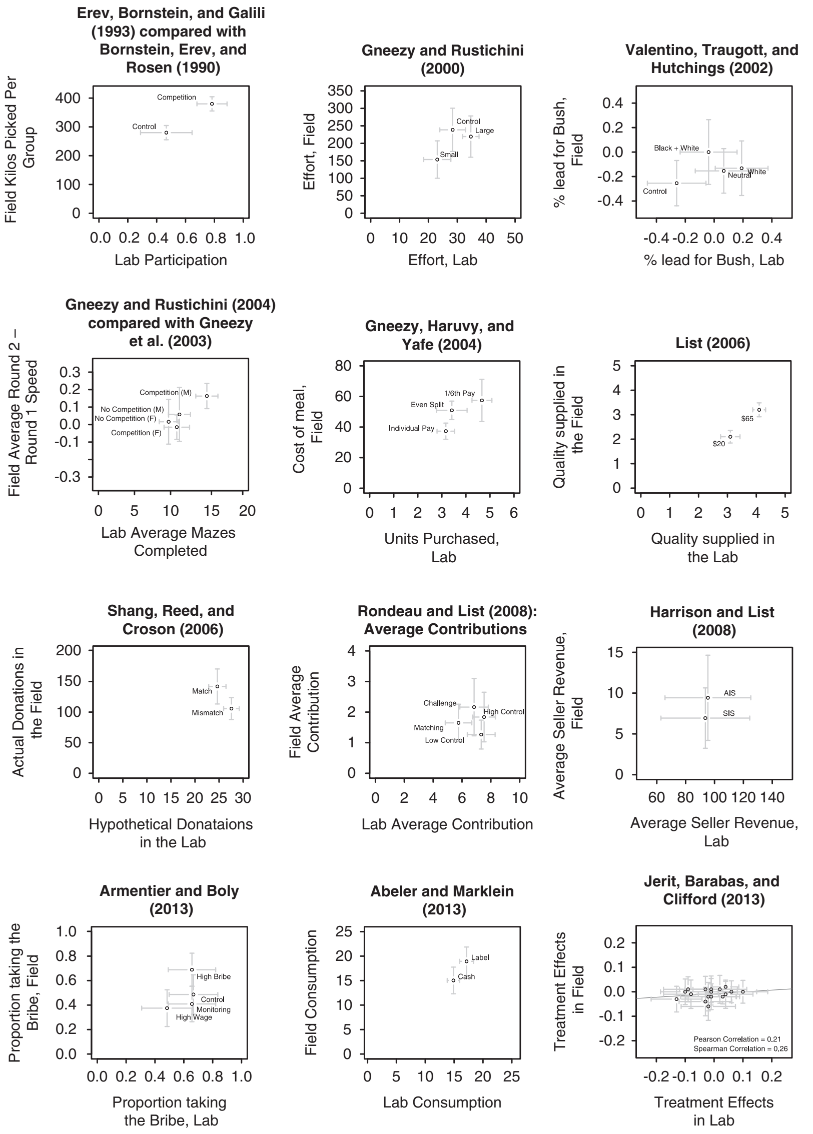

Fig. 1. 12 Lab-field comparisons
Note: each panel presents unstandardized means by experimental group, except for the last panel, which presents estimated treatment effects.

results are presented on the vertical axis. For the Jerit, Barabas and Clifford study, we present treatment effects, not group means. Taken together, these 12 studies demonstrate a reasonably strong correlation between group means in the lab and group means in the field.

In order to provide some measure of this correlation, we combined the 12 studies to form a single dataset.[5] In order to facilitate cross-study comparison, we recorded each

[5] When constructing an overall assessment of lab-field correspondence, we included only a single treatment effect pair from the Jerit, Barabas and Clifford study. Which pair we included caused slight changes in the overall correlation, ranging from 0.699 to 0.762. Figure 2 is generated using the treatment effect pair that presents the "median case" for lab-field correspondence. We believe that including all

TABLE 3 *12 Parallel Lab-Field Study Pairs*

Lab-Field Study Pair	Study Purpose
Erev, Bornstein and Galili (1993) & Bornstein, Erev and Rosen (1990)	Effect of competition on effort
Gneezy and Rustichini (2000)	Effect of small incentives on intrinsic motivation
Valentino, Traugott and Hutchings (2002)	Effect of racialized political advertising on support for a candidate
Gneezy and Rustichini (2004) & Gneezy et al. (2003)	Effect of competition on effort, by gender
Gneezy, Haruvy and Yafe (2004)	Evidence of the "Diners' Dilemma"
List (2006b)	Evidence of "Gift Exchange"
Shang and Croson (2008)	Effect of "Identity congruence" on donations
Rondeau and List (2008)	Relative effectiveness of matching versus challenge grants
Harrison and List (2008)	Effect of information on the "Winner's Curse"
Armantier and Boly (2013)	Effect of wages and monitoring on corruption
Abeler and Marklein (2013)	Effect of restricted vouchers on consumption
Jerit, Barabas and Clifford (2013)	Effect of newspapers on political knowledge and attitudes

study's estimates of the average treatment effects in the lab and field (rather than each study's treatment and control group averages). We standardized these treatment effects using the Cohen's *d* procedure as shown in Equations 1 and 2. This process allows us to compare effect sizes in standard units.

$$\text{Standardized Average Treatment Effect} = \frac{\mu_{Treat} - \mu_{Control}}{\sigma_{Control}} \tag{1}$$

$$\text{Standardized ATE Standard Error} = \frac{\sqrt{\frac{\sigma^2_{Treat}}{N_{Treat}} + \frac{\sigma^2_{Control}}{N_{Control}}}}{\sigma_{Control}} \tag{2}$$

Figure 2 shows our results. Treatment effects in the lab and treatment effects in the field show a upward-sloping relationship, with a rank-order correlation of 0.73. In order to be sure that the correlation was not driven by any single pair or a particular set of pairs of treatment effects, we carried out the following robustness check. First, we calculated the set of rank-order correlations that would occur if we dropped any single pair of treatment effects. Next, we calculated the set of correlations that would occur if we dropped any two pairs, and so on, out to any ten pairs.[6] The results of this procedure can be seen in Figure 3, moving leftward from the center of the graph. As we move rightward from the center of the graph, we take our 21 observed pairs and first duplicate any single pair, then any two pairs, and so on out to any ten pairs. With some abuse of notation, we label the x-axis as ranging from $\binom{21}{11}$ to $\binom{21}{31}$—strictly speaking, beyond $\binom{21}{21}$, we are appending $\binom{21}{X}$ combinations to our 21 observed pairs.

On the y-axis of Figure 3, we plot the five-number summary of the set of rank-order correlations. Unsurprisingly, the median correlation stays constant across all permutations. As we remove or duplicate more observations, the maximum and

(F'note continued)

17 pairs gives the Jerit, Barabas and Clifford study too much weight, but doing so reduces the rank-order correlation to 0.56.

[6] When dropping more than any six pairs, we randomly sampled 50,000 of the possible combinations.

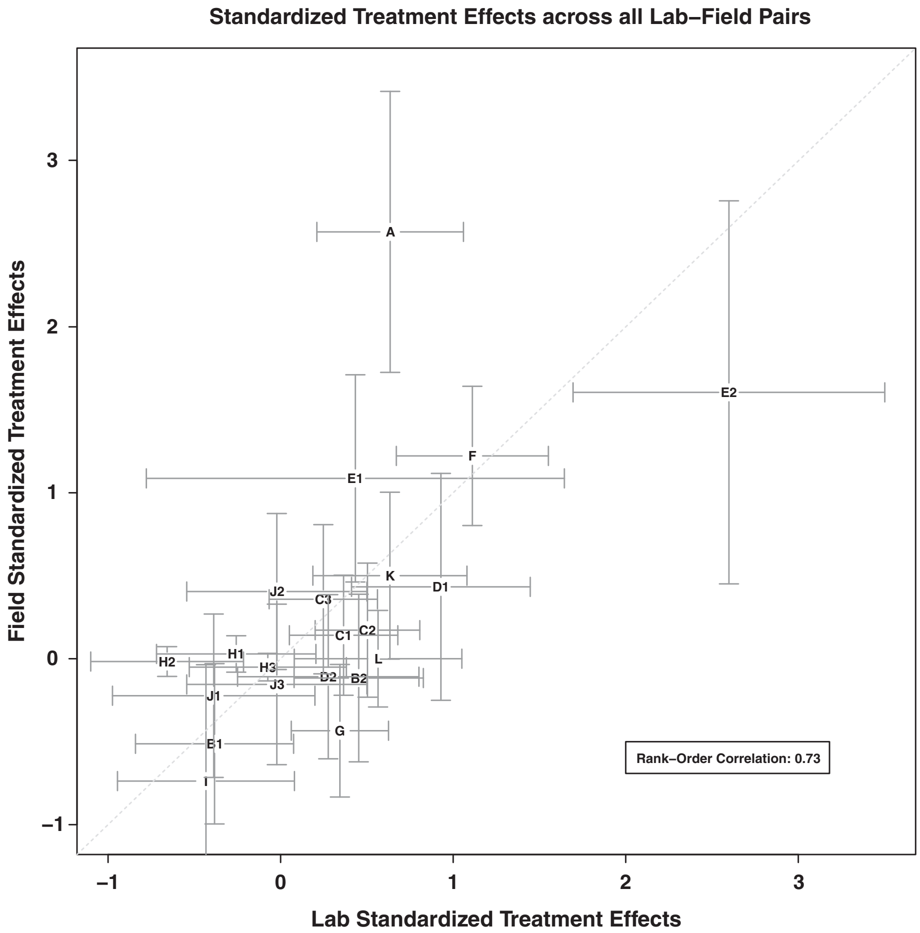

Fig. 2. Standardized treatment effects across all lab-field pairs
Note: each point represents a standardized lab-field treatment effect pair. A = Erev, Bornstein and Galili (1993) and Bornstein, Erev and Rosen (1990); B = Gneezy and Rustichini (2000); C = Valentino, Traugott and Hutchings (2002); D = Gneezy and Rustichini (2004) and Gneezy *et al.* (2003); E = Gneezy, Haruvy and Yafe (2004); F = List (2006b); G = Shang, Reed and Croson (2008); H = Rondeau and List (2008); I = Harrison and List (2008); J = Armantier and Boly (2013); K = Abeler and Marklein (2013); L = Jerit, Barabas and Clifford (2013).

minimum correlations spread out—even crossing zero and reaching one in the case of dropping any ten pairs. A striking feature of this graph, however, is that the interquartile range stays within the narrow band of about 0.6 to 0.8.

In summary, our review of parallel lab and field experiments indicates a strong overall pattern of lab-field correspondence. Although many of the studies we reviewed lacked sufficient power to discern the level of correspondence by themselves, the pattern of lab-field correspondence becomes clear when all of the studies are pooled. As Figure 2 indicates, stronger effects in the lab are associated with stronger effects in the field.

DISCUSSION

The overall pattern of agreement is surprising, given the many ways in which lab and field studies differ. In the collection of 12 studies we examined, seven of the lab studies

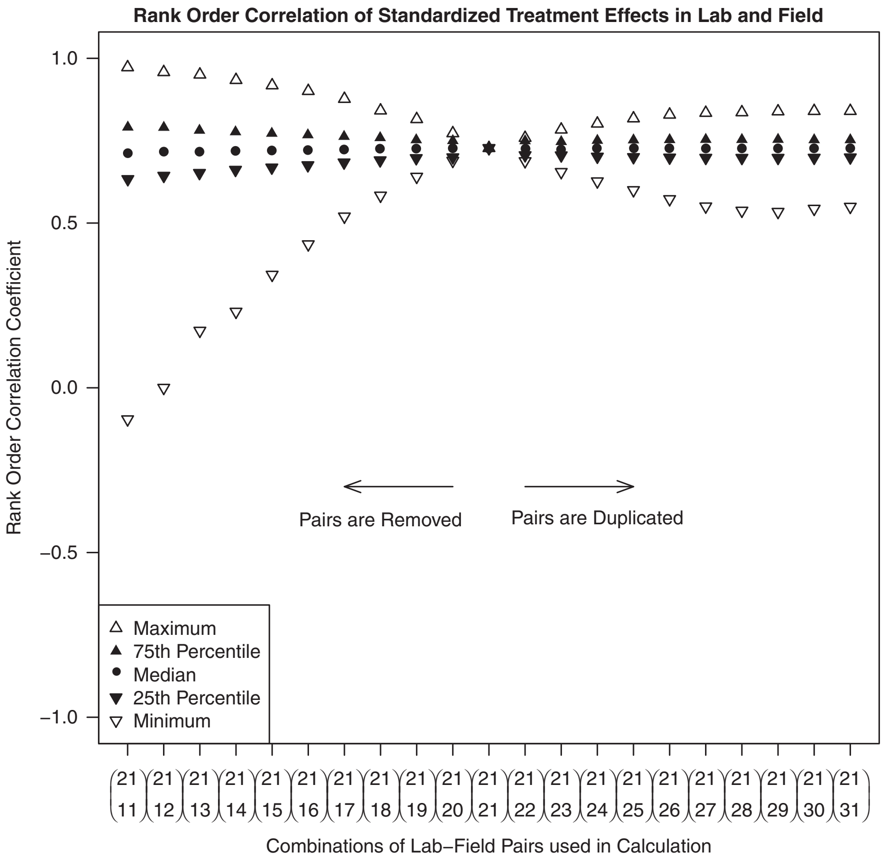

Fig. 3. Rank-order correlation of standardized treatment effects in lab and field
Note: we calculate the rank-order correlations of the 21 treatment effect pairs included in the meta-analysis, under all possible combinations duplicating or removing any 1, 2, 3… or 10 of them.

involved undergraduate subjects. In six of the lab studies, subjects encountered an abstract rendering of a real-life situation, such as the exchange of tokens meant to simulate the division of a restaurant bill. In all 12 of the lab studies and nine of the field studies, outcomes were measured immediately after subjects encountered the treatment. Our collection of studies is too sparse to permit a more fine-grained analysis of how changes in lab-field similarity in terms of subjects, treatments, contexts and outcomes affect correspondence. Nevertheless, correspondence is remarkably high given that our collection of lab and field studies often diverges on more than one of these dimensions. It is not hard to think of sound theoretical arguments for expecting that small incentives should improve performance in the lab (Gneezy and Rustichini 2000), though they did not, or that the label condition should not influence allocations (Abeler and Marklein 2013), though it did. Indeed, against a backdrop of other meta-analyses that detect meaningful variation in results *within* experimental literatures that are entirely situated either in the lab (Engel 2011; Johnson and Mislin 2011) or in the field (Baird et al. 2013; Stewart et al. 2012), it is remarkable to find such a high degree of lab-field correspondence when so many features of the lab and field differ.

The pattern is all the more striking given the small samples used in several of the studies. In the 12 pairs of lab-field comparisons, the median sample size per treatment condition was 40 for the lab and 37.5 for the field. If we suppose that these samples were drawn at random from a superpopulation such that the usual formulas for sampling variability apply, the correlation we observe masks an even stronger underlying correlation that is attenuated somewhat by sampling error. The raw Pearson correlation among the 21 treatment effect pairs is 0.64. The reliabilities for the lab and field treatment effects are 0.78 and 0.82, respectively.[7] The disattenuated correlation is found to be $0.64/\sqrt{0.78^{*}0.82} = 0.80$, further attesting to the strength of lab-field correspondence in published work.

That said, we recognize that caution is warranted when drawing inferences about lab-field correspondence based on the current state of the literature. As noted earlier, studies of lab-field correspondence have emerged in an ad hoc fashion, without any attempt to systematically investigate variation in subjects, treatments, contexts and outcomes. The lack of systematic procedures raises two concerns. One is the so-called file-drawer problem (Rosenthal 1979). If authors or journal editors have a preference for noteworthy findings—either demonstrations of very high or very low correspondence—the distribution of correspondence reported in academic work might be unrepresentative of the broader set of studies that were conducted. We might see an exaggerated level of lab-field correspondence because high-correspondence findings are especially likely to find their way into published articles, conference papers or manuscripts posted online. As this literature matures, it will be interesting to see whether future research findings diverge from the strong correspondence apparent in currently available work. The threat of publication bias also underscores the need for preregistration of experiments and institutional arrangements facilitating the reporting of results even in the absence of publication (Humphreys, de la Sierra and van der Windt 2013).

A second concern relates to the way in which lab-field comparisons are chosen. In most of the studies considered here, a field experiment was conducted to confirm, validate or challenge a laboratory result. If the lab is to function as a cost-effective substitute for field research, it makes sense to take the opposite approach: start with a field experiment, and look for parallel tests in the lab. Similarly, if the aim is to calibrate lab designs so that their results agree with field research findings (Camerer forthcoming, 47), field experimentation is a natural starting point.

How might one go about looking for field research that lends itself to lab-field comparisons? One approach is to start with well-developed field literatures that examine the effects of treatments that can be delivered in lab or field contexts. For example, field experiments on the accountability of public officials have assessed how voters respond to revelations of politicians' behavior (Chong et al. 2010; Humphreys and Weinstein 2010; Banerjee et al. 2010). Field experiments on tax compliance (Fellner, Sausgruber and Traxler 2013; Castro and Scartascini 2013) have distinguished between the effectiveness of threats versus appeals to morals. In tests of the psychology of sunk cost, field experimental results have shown that charging for health products does not increase use relative to free provision (Cohen and Dupas 2010; Ashraf, Berry and Shapiro 2010). The challenge in each of these instances is to devise lab experiments that address the same causal or policy question, which in some cases involves outcomes that are expressed long after the administration of a treatment.

[7] The reliabilities were estimated using a simulation technique in which hypothetical treatment effects are drawn from the normal distributions implied by the estimated treatment effects and their standard errors. The estimated reliability is the square of the correlation between two treatment effect draws. We use the average of 10,000 estimated reliabilities for both lab and field to disattenuate the raw correlation.

The debate over the relative merits of laboratory and field experiments is often framed in terms of the questions that each is better equipped to address: a schematic version of this debate is that field experiments can answer real-world questions while lab experiments can isolate causal mechanisms. We have attempted to advance this debate by offering tentative evidence that when studies in the lab and field attempt to answer similar questions, they arrive at similar answers. The task going forward is to investigate the question of lab-field correspondence in a more systematic fashion, designing research specifically to assess the conditions under which correspondence is maximized.

REFERENCES

Abeler, J., and F. Marklein. 2013. 'Fungibility, Labels, and Consumption'. Unpublished manuscript.

Arceneaux, K., and D.W. Nickerson. 2009. 'Who Is Mobilized to Vote? A ReAnalysis of 11 Field Experiments'. *American Journal of Political Science* 53(1):1–16.

Armantier, O., and A. Boly. 2013. 'Comparing Corruption in the Laboratory and in the Field in Burkina Faso and in Canada'. *The Economic Journal* 123(573):1168–87.

Ashraf, N., J. Berry, and J.M. Shapiro. 2010. 'Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia'. *American Economic Review* 100(December):2383–413.

Baird, S., F. Ferreira, B. Özler, and M. Woolcock. 2013. 'Relative Effectiveness of Conditional and Unconditional Cash Transfers for Schooling Outcomes in Developing Countries: A Systematic Review'. Technical report, Campbell Systematic Reviews.

Banerjee, A., S. Kumar, R. Pande, and F. Su. 2010. 'Do Informed Voters Make Better Choices? Experimental Evidence from Urban India'. Unpublished manuscript.

Benz, M., and S. Meier. 2008. Do People Behave in Experiments as in the Field? Evidence from Donations. *Experimental Economics* 11(3):268–81.

Bond, R., and P.B. Smith. 1996. 'Culture and Conformity: A Meta-analysis of Studies Using Asch's (1952b, 1956) Line Judgment Task'. *Psychological Bulletin* 119(1):111–37.

Bornstein, G., I. Erev, and O. Rosen. 1990. 'Intergroup Competition as a Structural Solution to Social Dilemmas'. *Social Behaviour* 5(4):247–60.

Camerer, C.F. forthcoming. 'The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List'. In G. Frechette and A. Schotter (eds), *Methods of Modern Experimental Economics*. Oxford: Oxford University Press.

Castro, L., and C. Scartascini. 2013. 'Tax Compliance and Enforcement in the Pampas: Evidence from a Field Experiment'. Washington, DC: Inter-American Development Bank, 2013.

Charness, G., and A. Schram. 2013. 'Social and Moral Norms in Allocation Choices in the Laboratory'. Unpublished manuscript.

Chong, A., A.L. De La O, D.S. Karlan, and L. Wantchekon. 2010. 'Information Dissemination and Local Governments Electoral Returns, Evidence from a Field Experiment in Mexico'. Unpublished manuscript.

Cohen, J., and P. Dupas. 2010. 'Free Distribution or Cost-sharing? Evidence from a Randomized Malaria Prevention Experiment'. *Quarterly Journal of Economics* 125(1):1–45.

Cook, T.D., and D.T. Campbell. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin.

Engel, C. 2011. 'Dictator Games: a Meta Study'. *Experimental Economics* 14(4):583–610.

Erev, I., G. Bornstein, and R. Galili. 1993. 'Constructive Intergroup Competition as a Solution to the Free Rider Problem: A Field Experiment'. *Journal of Experimental Social Psychology* 29:463–78.

Falk, A., and J.J. Heckman. 2009. 'Lab Experiments are a Major Source of Knowledge in the Social Sciences'. *Science* 326(5952):535–8.

Fellner, G., R. Sausgruber, and C. Traxler. 2013. 'Testing Enforcement Strategies in the Field: Threat, Moral Appeal and Social Information'. *Journal of the European Economic Association* 11(3):634–60.

Gerber, A.S. 2004. 'Does Campaign Spending Work?: Field Experiments Provide Evidence and Suggest New Theory'. *American Behavioral Scientist* 47(5):541–74.

Gerber, A.S., and D.P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.

Gerber, A.S., D.P. Green, and C.W. Larimer. 2008. 'Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment'. *American Political Science Review* 102(1):33–48.

Glennerster, R., and K. Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton, NJ: Princeton University Press.

Gneezy, U., E. Haruvy, and H. Yafe. 2004. 'The Inefficiency of Splitting the Bill'. *The Economic Journal* 114:265–80.

Gneezy, U., and J.A. List. 2006. 'Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments'. *Econometrica* 74(5):1365–84.

Gneezy, U., M. Niederle, and A. Rustichini. 2003. 'Performance in Competitive Environments: Gender Differences'. *The Quarterly Journal of Economics* 118(3):1049–74.

Gneezy, U., and A. Rustichini. 2000. 'Pay Enough or Don't Pay at All'. *The Quarterly Journal of Economics* 115(3):791–810.

——. 2004. 'Gender and Competition at a Young Age'. *American Economic Review* 94(2):377–81.

Green, D.P., and A. Tusicisny. 2013. 'Statistical Analysis of Results from Laboratory Studies in Experimental Economics: A Critique of Current Practice'. Unpublished manuscript.

Großer, J., and A. Schram 2010. 'Public Opinion Polls, Voter Turnout, and Welfare: An Experimental Study'. *American Journal of Political Science* 54(3):700–17.

Habyarimana, J., M. Humphreys, D.N. Posner, and J.M. Weinstein. 2009. *Coethnicity: Diversity and the Dilemmas of Collective Action*. New York: Russell Sage Foundation.

Harder, V.S., E.A. Stuart, and J.C. Anthony. 2010. 'Propensity Score Techniques and the Assessment of Measured Covariate Balance to Test Causal Associations in Psychological Research'. *Psychological Methods* 15(3):234–49.

Harrison, G.W., and J.A. List. 2004. 'Field Experiments'. *Journal of Economic Literature* 42(4):1009–55.

——. 2008. 'Naturally Occurring Markets and Exogenous Laboratory Experiments: A Case Study of the Winners Curse'. *The Economic Journal* 118(528):822–43.

Henrich, J., R. Boyd, S. Bowles, C.F. Camerer, E. Fehr, H. Gintis, and R. McElreath. 2001. 'In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies'. *American Economic Review* 91(2):73–8.

Henrich, J., S.J. Heine, and A. Norenzayan. 2010. 'The Weirdest People in the World?' *Behavioral and Brain Sciences* 33:61–83.

Hotz, V.J., G.W. Imbens, and J.H. Mortimer. 2005. 'Predicting the Efficacy of Future Training Programs using Past Experiences at other Locations'. *Journal of Econometrics* 125(1–2): 241–270.

Humphreys, M., R. Sanchez de la Sierra, and P. van der Windt. 2013. 'Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration'. *Political Analysis* 21(1):1–20.

Humphreys, M., and J. M. Weinstein. 2010. 'Policing Politicians: Citizen Empowerment and Political Accountability in Uganda'. Unpublished manuscript.

Jerit, J., J. Barabas, and S. Clifford. 2013. 'Comparing Contemporaneous Laboratory and Field Experiments on Media Effects'. *Public Opinion Quarterly* 77(1):256–82.

Johnson, N.D., and A.A. Mislin. 2011. 'Trust Games: A Meta-analysis'. *Journal of Economic Psychology* 32(5):865–89.

Kessler, J., and L. Vesterlund. forthcoming. 'The External Validity of Laboratory Experiments: Qualitative rather than Quantitative Effects'. In G. Frechette and A. Schotter (eds), *Methods of Modern Experimental Economics*. Oxford: Oxford University Press.

King, E.B., and A.S. Ahmad. 2010. An Experimental Field Study Of Interpersonal Discrimination Toward Muslim Job Applicants. *Personnel Psychology* 63(4):881–906.

Lalancette, M.-F., and L. Standing. 1990. Asch Fails Again. *Social Behavior and Personality* 18(1): 7–12.

Levitt, S.D., and J.A. List 2007. 'Viewpoint: On the Generalizability of Lab Behaviour to the Field'. *Canadian Journal of Economics* 40(2):347–70.

List, J.A. 2006. 'The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions'. *Journal of Political Economy* 114(1):1–37.

Milgram, S. 1963. 'Behavioral Study of Obedience'. *The Journal of Abnormal and Social Psychology* 67(4):371–8.

Morton, R.B., and K.C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. New York: Cambridge University Press.

Oosterbeek, H., R. Sloof, and G. van de Kuilen. 2004. 'Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis'. *Experimental Economics* 7(2):171–88.

Paluck, Elizabeth Levy. 2009. 'Reducing Intergroup Prejudice and Conflict Using the Media: A Field Experiment in Rwanda'. *Journal of Personality and Social Psychology* 96(3):574–87.

Rondeau, D., and J.A. List. 2008. 'Matching and Challenge Gifts to Charity: Evidence from Laboratory and Natural Field Experiments'. *Experimental Economics* 11(3):253–67.

Rosenthal, R. 1979. 'The File Drawer Problem and Tolerance for Null Results'. *Psychological Bulletin* 86(3):638–41.

Schultz, W.P., A.M. Khazian, and A.C. Zaleski. 2008. 'Using Normative Social Influence to Promote Conservation Among Hotel Guests'. *Social Influence* 3(1):4–23.

Sears, D.O. 1986. 'College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature'. *Journal of Personality and Social Psychology* 51(3):515–30.

Shadish, W.R., T.D. Cook, and D.T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton, Mifflin and Company.

Shang, J., A. Reed, and R. Croson. 2008. 'Identity Congruency Effects on Donations'. *Journal of Marketing Research* 45:351–61.

Stewart, R., C.V. Rooyen, M. Korth, A. Chereni, N. Rebelo Da Silva, and T. de Wet. 2012. *Do Micro-credit, Micro-savings and Micro-leasing Serve as Effective Financial Inclusion Interventions Enabling Poor People, and Especially Women, to Engage in Meaningful Economic Opportunities in Low- and Middle-income Countries*. London: EPPI-Centre, University of London.

Student. 1908. 'Probable Error of a Correlation Coefficient'. *Biometrika* 6(2):302–10.

Valentino, N.A., M.W. Traugott, and V.L. Hutchings. 2002. 'Group Cues and Ideological Constraint: A Replication of Political Advertising Effects Studies in the Lab and in the Field'. *Political Communication* 19:29–48.