

Memo on Experimental Approach Used in the Field

Macartan Humphreys and Jeremy Weinstein
3 February 2008

We used experimental methods “in the field” (not field experiments) to try and answer a specific context-dependent question: *how does ethnic diversity matter for public goods provision in Kampala, Uganda?* To answer this question, we placed a series of theoretical predictions in competition with one another. The research thus speaks clearly to the larger question “how does ethnicity matter?” but our ambition in the project was only to try to answer this question in a given context.

In this memo, we describe some of the challenges we faced in using experimental methods to answer this research question. We think that many of these challenges are quite general to experimental approaches in political science and are worth discussing at the conference. In confronting some of them, we felt that our methods represent reasonable responses to the challenges; in other cases, however, we are much less satisfied. In some of these cases, we are doubtful that an experiments-in-the-field approach can resolve the problems and are more supportive of a more thorough-going field experiment approach. For some problems, however, we are unsure that even a field experiment approach can overcome the challenge.

Experimental Methods and Theory Testing

We describe a series of challenges we think that experimental methods are faced with. How important these different challenges are depends of course on *what the purposes of an experiment are* in the first place. At a superficial level, experiments are designed to measure a quantity, typically the effect of x on y . Beyond this general goal, however, we think it is useful to distinguish two families of distinct motivations for using experiments (there are many more):

- (i) **To “demonstrate,” “check” or, in some accounts, “test” the logic of a formal model.** Say a model supports the *analytical* proposition that under a set of conditions Z , x has a positive effect on y , then one might use an experiment to show that indeed in some actual setting like Z^* , some actual variable x^* does indeed have a positive impact on some actual variable y^* . This approach uses experiments to move from analytic to empirical claims but suffers from the problem that if the predictions are not supported, this says nothing about the analytical proposition (it says more about the ability of the researcher to put conditions like Z in place).
- (ii) **To test an empirical proposition** (which may perhaps be informed by a formal model) of the form: “In setting Z^* in some family of actual settings Z^{**} , some actual variable x^* does indeed have a positive impact on some actual variable y^* .” Unlike case (i) the proposition that is tested here is not *analytically* true, rather, it is an empirical proposition. If a body of evidence fails to find support for the empirical proposition, then this suggests that the empirical proposition is false but not that any analytic proposition that generated it is false (although the evidence against the empirical proposition would suggest that the analytic proposition might not be very helpful for understanding the world!).

Our work uses experiments to address a project like that described in (ii). For (ii), a lot depends on what the family of settings Z^{**} is. In some cases, the claims examined in experiments are universal claims about humans: eg. “humans use Bayes’ rule to update” or “humans use backward induction.” In other cases, the claims are much more likely to be context dependent, for example “humans pay more attention to newspapers than TV” or “humans are racist.” In our case (and we believe for many studies of interest to

political science), the claims we are interested in are undeniably context-dependent and this fact shapes our research design in important ways, as we describe below.

Challenges in Experimental Research Design

1. **The manipulability of the independent variables of interest.** A common criticism of field experiments is that some manipulations – such as sending a UN mission into a country – are not subject to experimental control. We believe in fact that a very wide range of variables of interest to political science can be manipulated for experimental purposes. However a key limitation of *this* project is the fact that the central independent variable of interest could *not* be manipulated. We cannot randomly assign an ethnic identity to someone, keeping all other factors of the person constant. In fact, more subtly, we cannot randomly determine the ethnic composition of a group, since doing so involves changing other features of its membership. For this reason, we cannot lay claim to the greatest advantage of experimentation: exogeneity. We cannot guarantee that the co-ethnicity of a pairing is uncorrelated, even *ex ante*, with features such as education or wealth. We believe that this is a general problem in the use of experiments for the study of constituent features of humans, such as ethnicity or gender; this problem, we believe is not resolved by a move to more field based approaches (we note of course that this inferential problem is a feature of the subject matter, not of the experimental method; many observational studies will face just the same problem).¹

“Assigning” identities to individuals, as is done in minimal group experiments, does not get around the problem since it answers a different question (What is the effect of placing someone in an artificial group? Not: What is the effect of membership in a particular social group?). The cost of not assigning identities in this way is that the identities that our subjects brought with them to the lab may be correlated with other things outside of our control. The advantage, however, is that we are able to observe actual attitudes and behavior by members of particular groups with respect to other groups, not simply behavior and attitudes generated in the lab. In that sense we tradeoff a desire to study actual ethnic identities with the ability to identify truly causal effects of the identities that we construct.

Frustrating as this is, there are still benefits to the random assignment of individuals to pairs. In our experiments, whether an individual *interacts* with a co-ethnic or not is uncorrelated with all features other than those that are constituent of the individuals themselves; notably it does not depend on features such as the past play of the individuals in our games. In particular, by using randomization to assign players to partners we overcome selection effects (for example, absent randomization we might observe individuals interacting with non-co-ethnics only in cases in which both players expect positive returns).

2. **External Validity I: Site Selection.** There are a number of deep external validity concerns associated with our approach. Surprisingly, we are most commonly confronted with a naïve one: *It may be true in Kampala but is it true generally?* The answer is quite simple: we don’t know. To answer that question, we would need to do similar experiments in many more places. In fact – and here is the difficulty – to answer that question convincingly for a population of cases *P*, we would want to run the experiments not just in many sites, but in a random sample of sites in *P*. Ultimately we would like to be able to

¹ We note that *in principle* if ethnicities can be changed then they can be manipulated; in practice however this is beyond our control.

account for variation in results across sites. This problem of external validity, though commonly leveled against experimental approaches, isn't in fact all that different from the one confronted by any study in a single place that tries to speak to a broader population of cases. The deeper external validity question (inherent to lab experiments) with which we are concerned is: Is this true in the Kampala beyond the lab, and are these results valid for different social dilemmas than the ones we study?

3. **External Validity II: Subject Selection.** For the questions of interest in our research we cannot reasonably maintain the claim that is implicit in much lab experimentation that the quantities of interest—such as attitudes of co-ethnics or co-ethnic specific behavior—are universal to humans. Instead we expect that ethnic behavior and attitudes are *likely to be acutely context dependent*. We expect for example that the extent of co-ethnic favoritism or the ability to “locate” a co-ethnic are not universal features of human beings, but specific to a population in a particular location at a particular point in time. In this sense, our experiments play a role not unlike any other measurement strategy (ie. the use of surveys, interviews, etc.) in political science research. Although our experiments mirror those used in behavioral economics, this concern with settings forces us to pay far greater attention to the issue of the population for which reliable inferences can be drawn. *We have found that this is a concern that seems not to resonate with many in the wider community of experimentalists.*

This concern has important implications for the selection of our site and subjects. First, we took the laboratory to the “field,” electing to carry out our experiments in Uganda, rather than with college students in the US. Cognizant of the evolution of work in behavioral economics that convincingly demonstrates differences in the social preferences people exhibit in different societies, it was clear to us that to speak to issues of ethnic diversity and public goods provision in settings that our research focuses on, we needed to carry out our experiments in a real-world context of substantive interest. Following Alesina, Baqir, and Easterly, we might have looked at how ethnic identity conditions behavior in U.S. counties by playing experimental games with voters; instead, given our interest in the relationship between diversity and collective action in Africa, we took our games to Kampala. Within Kampala we set up the laboratory in a community that exhibited high levels of ethnic diversity and low levels of public goods provision (based on data collected in a pre-survey). In doing so we focused on a single point in a scatterplot that describes a more general relation between diversity and low public goods provision; with more macro evidence to motivate the claim that diversity and low public goods provision are related in the area our question then was why, in this area, is this the case.

Finally, in contrast to most of the work in behavioral economics, we drew a random sample of subjects from the community. Sampling has not been a big focus in behavioral economics, especially in the early work, as many experiments were seen as useful for demonstrating *any deviations* from the standard assumptions underlying neo-classical economics. Work that focuses on showing the *possibility* of certain types of behavior is also often less concerned about identifying the population for which inferences are valid. But work in behavioral economics is moving now more in the direction of characterizing how features of different societies shape individual behavior, putting a premium on appropriate sampling to ensure that representative populations are employed in measuring the behaviors of interest. Yet, aside from Rick Wilson's work in Russia, we know of almost no studies in behavioral economics that use random samples of the population of interest. Of course, in survey experiments (with the rise of Knowledge Networks), we are increasingly seeing random samples used to test hypotheses about racial priming (Hutchings et al), the impact of international law on voter perceptions of foreign policy (Tomz), etc.

4. **External Validity III: Estimated Average Treatment Effects Depend on the Levels at Which Controls are Set.** Consider a data generating process in which $y = x + xz + z$. The treatment of interest is x . Let's say that z can take values in $\{0, -1, -2\}$. The marginal effect of x on y is of course $(1+z)$ which can take on values, depending on z , of -1 , 0 or 1 . The *average treatment effect* can thus be negative, positive or zero depending on the distribution of z in the sample. What is the quantity of research interest? In the best of all worlds, it is to get the marginal effect of x for all possible values of z . But that may require knowing more about the data generating process (or having more power) than we typically do. Indeed, the researcher may not know about z , or z may be unobservable. The great advantage of a controlled environment is that it can allow for a better measure of any particular parameter, However a problem with a "controlled environment" is that z may unwittingly be part of what might be controlled. If so, then the estimated average effect is only good for the particular value of z that has been set (possibly unknowingly) by the researcher. We believe that we have avoided some such problems by relying on random samples of subjects (the combination of a random sample of subjects with random matchings means not just that our subjects are representative of a population but that an individual's matchings reflect the types of individual encounters she would expect to have). However other features that we controlled likely do affect the external validity of our experiments with respect to the population of interactions for which we wish to make inferences. For example, we "set" ethnic salience to 0 in the sense that ethnicity was not explicitly referred to as part of the games. If, in the universe of interest, ethnic salience is higher and this has an effect on how people behave, our estimated average treatment effects are not valid for this wider class of settings. This problem of "contamination from control" is, we believe, less salient for many field experiment designs.
5. **Metrics and Aggregation.** Since we are ultimately interested in making a claim about a particular community, and not simply about our experimental subjects, we are faced with a set of "translation" problems. One of these is a problem of identifying "metrics." How does a particular effect size recorded in the lab translate into an effect size in an actual social interaction? Using our present methods, we do not know. Unless we can answer this question however it is not clear that our strategy has any merit. The metric problem is especially germane when we turn to comparing the results from different games or from different mechanisms. What inferences would we have made if we had gotten positive results on multiple mechanisms? how could we have placed them on a common metric and determined which was more important, or which would likely dominate in a particular social setting? A second problem relates to aggregation. Having identified patterns of play in the laboratory, we face the challenge of using dyadic behavior to attempt to understand outcomes that result from more complex processes in the real world. To what extent can behavior observed in dyadic interaction be used to generate predictions that might apply to the outcomes of more complex processes? In Chapter 6 of the book, we undertake an exercise that uses a matrix of patterns of play across ethnic dyads to predict public goods production for different ethnic demographics. We find some encouraging results, but are conscious that the exercise relies on a very strong assumption about aggregation. These problems we feel are common to lab experimentation which often examines individual behavior on a scale that is quite different to that which is of importance to the researcher. This is one problem however which we feel is greatly aided with a field experimentation approach in which the experiment can be on the same scale as (and be in the same class as) the family of outcomes of interest.
6. **Equilibrium play is not learned quickly.** In many instances, experiments are used to work out what equilibrium, if any, players will play in a given game. Indeed, in many cases the game played is new to the subjects. In such cases, a problem arises that knowing how to play in equilibrium may take some learning. The problem is especially severe if the game has *multiple* equilibria since players must not only be able to play in equilibrium but also coordinate on a single one. Absent a hypothesis on

equilibrium selection, in such cases it is hard to know whether out-of-equilibrium play tells us anything about play in settings in which players do have time to learn and coordinate on equilibria. Our interest was not in finding out how players would play a given game, but more generally what strategies people *in fact* use when playing particular classes of games (with particular co-players). Our approach involved a very strong supposition: that individuals play some class of similar games similarly and will recognize our game as a class of games that they play in the real world. In addition to this, this goal shaped a number of design elements:

1. Our games had to be a member of some class of (or be sufficiently similar to) actual social problems (a fact that we checked through extensive debriefings with subjects);
2. We had to *prevent* learning, since our interest is not in whether people can learn to play together in a particular way but whether they do play together in a particular way (a fact we tested through a variant of the trust game after the experiments were completed);
3. Players had to interact with people who were “known quantities”; that is, with players with which they might typically engage in such interactions and of whom they might expect a particular type of equilibrium play. For this, random sampling played an important role. With a random sample (and with it being common knowledge that the set of subjects was random) we could ensure that subjects were playing with (and knew they were playing with) a set of people much like those that they would engage with in anonymous social settings.

7. **Stable Parameters and Simple Games.** To link our empirical results to rival theoretical accounts of the effects of ethnicity we assume that a set of individual level features – such as the extent of co-ethnic favoritism – are stable and measurable. Thus we assume that if players exhibit other regarding preferences towards each other in one game that that speaks of a general propensity to treat each other well that travels across games. This assumption is common to non-experimental approaches, including qualitative accounts (such as accounts that explain ethnic violence in terms of ethnic hatreds). It is however a very strong presupposition and (beyond the observed consistencies across games) we do not have much evidence to support it. Absent such an assumption however it is difficult to see how one can extricate rival explanations for a complex phenomenon.

To illustrate the point consider the problem of cooperation in a collective action dilemma. Such cooperation can arise for a host of reasons. One is that people in fact care for each other’s wellbeing. Another is that they expect retribution for non-cooperating. Our aim is to disentangle these effects. Our approach is to try to measure a set of underlying features (or “primitives”) of a population, by playing a set of very simple games. For example, anonymous dictator games were used to measure altruism. And the difference between play in anonymous and non-anonymous dictator games was used to measure reciprocity. These games are far simpler than the set of standard games now used in the experimental economics literature (games such as the trust game, the ultimatum game, or multi-player public goods games). Moreover, they are simpler than the very complex games designed by some social scientists to imitate real political situations (Barr). The advantage of the “primitives approach” is that it can provide a handle on the mechanism question. (In contrast studies of play in the trust game are now plagued with debates about whether trusting behavior reflects altruism, risk-seeking behavior, or the operation of reciprocity norms (Ashraf et al; Karlan)). However the approach requires an inferential leap (does the same feature that leads people to share a dollar also lead them to spend an afternoon working together?). A key design question then is: are we better off asking people to undertake exercises that mirror their real world challenges in all of their complexity? Or should we continue to try to identify primitives? If primitives exist and can be measured, do we know enough to be able to aggregate up?