

# Experimental Ethnography: The Marriage of Qualitative and Quantitative Research

By  
LAWRENCE W. SHERMAN  
and  
HEATHER STRANG

Experimental and ethnographic research methods are often described as mutually exclusive. This article suggests how they could be combined in the method of “experimental ethnography.” Building ethnographic methods into the separate branches of randomized controlled trials could substantially increase the range of conclusions that can be produced by experimental research designs, as well as by ethnographic methods. Experimental designs offer greater internal validity for learning *what* the effects of a social program are, and ethnographic methods offer greater insight into *why* the effects were produced. The prospects for such integration depend on the capacity of two different communities within social science to work together for the common goal of discovering truth.

*Keywords:* randomized controlled trials; restorative justice; sampling; experimental research designs; quantitative; qualitative

Everyone loves a good story. Yet most of us are suspicious about “anecdotal evidence.” We often say we prefer “solid data” for making big decisions, especially in spending trillions of dollars in taxes. But if the numbers do not make sense to us, we often reject them as flawed or biased.

Audiences may be even more likely to reject program evaluations when they come to surprising or counterintuitive conclusions. Many people are shocked and incredulous when they hear that an expensive mentoring and summer-camp program in the 1930s caused harmful effects on

*Lawrence W. Sherman, president of the American Academy of Political and Social Science, is the Albert M. Greenfield Professor of Human Relations and chair of the Department of Criminology at the University of Pennsylvania.*

*Heather Strang is the director of the Centre for Restorative Justice at the Research School of Social Sciences, Australian National University.*

NOTE: The authors would like to thank the Jerry Lee Foundation for the support it has provided for analyzing twelve randomized controlled trials on restorative justice in England and Australia.

DOI: 10.1177/0002716204267481

a large sample of teenagers for the rest of their lives (McCord 1978). They may find it hard to believe that arresting men for wife-beating can increase rather than reduce the frequency of violence by these men in the future (Sherman and Smith 1992; Pate and Hamilton 1992). They may be skeptical that young mothers visited in their homes by registered nurses will be much less likely to abuse their children and that those children will become much less likely to commit crimes when they grow older (Olds et al. 1998).

These conclusions are hard to accept when they are stated only in terms of numbers. Yet if the numbers are linked to stories about how these program effects actually happened with real human beings, an audience might be more likely to understand and accept the numerical conclusions. More important, social scientists themselves might be more likely to discover the truth about the program effects and to reach more understanding about why effects may vary across individuals exposed to a program.

In this article, we propose that social science unite the insights of stories and numbers. This merger should increase the contributions of social science to the reduction of human misery. The unification of stories and numbers can be achieved by introducing ethnographic methods into the best-known research design for producing unbiased conclusions about the average, numerical effects of almost anything on human beings, who vary widely in their reactions to almost everything (Cox 1958). Ethnography focused on finding—or falsifying—differences between randomly assigned treatment groups would not be “experimental” from the standpoint of ethnographic methods; standard methods of observation and interviews would suffice. From the standpoint of field experiments, however, the introduction of ethnography as a form of data collection would create an entirely new domain within work already dubbed “experimental,” such as “experimental physics” or “experimental criminology.” This work could quite properly be called “experimental ethnography.”

Experimental ethnography is a tool for answering questions about *why* programmatic attempts to solve human problems produce *what* effects, on average, in the context of the strong internal validity of large-sample, randomized, controlled field experiments (Campbell and Stanley 1963). The hypothesis that the two methods are “a thousand steps removed” from one another (Maruna 2001) can be falsified by the joint efforts of ethnographers and experimentalists. This strategy can achieve experiments that create both a strong “black box” test of cause and effect and a rich distillation of how those effects happened inside that black box, person by person, case by case, and story by story.

The article begins with a brief historical context for the proposal for experimental ethnography. It then tells a story about one case in a controlled experiment testing a radical new approach to criminal justice. Using that story as the example for experimental ethnography, the article describes the sampling strategies and qualitative research methods that could be used to enrich any field experiment. The article then turns to the question of systematic reviews of experiments and the crucial role experimental ethnography can make in learning “what works.” It con-

cludes with a brief look at the likely or possible future of the invention of experimental ethnography.

## The Rise and Fall of the Wall

For the past two centuries, the study of society has been divided by a visceral difference in taste among scholars. For several millennia, historians and biographers have recorded the *stories* of human society, emphasizing the unique facts of each event and its specific antecedents. Since the early nineteenth century, however, natural philosophers (later called “scientists”) have computed the *averages* of human society: the central tendencies and distributions of the human condition that encompass millions of stories in a few numbers. Ever since then, students of society have gathered in two camps, separated by a thick wall: those who prefer stories (e.g., Mayhew 1861-1862; Anderson 1978) and those who prefer numbers (Quetelet 1835; Stouffer et al. 1949). Perhaps the high point of this wall was captured in W. H. Auden’s 1946 poem “Under Which Lyre”:

Thou shalt not answer questionnaires  
or quizzes upon world affairs.  
Thou shalt not sit with statisticians  
nor commit a social science.

At the close of the twentieth century, the wall began to fall between the two camps. Scholars began to integrate qualitative and quantitative materials to generate and test hypotheses in history (e.g., Cannadine 1990), psychology (e.g., Sulloway 1996), sociology (e.g., Massey and Denton 1993), political science (e.g., Putnam 2000), economics (e.g., Reuter 1983), criminology (e.g., Sampson and Laub 1993), and psychiatry (e.g., Vaillant 2002). While by no means unprecedented, the integration of stories and numbers in these “natural history” analyses created a new climate in which the assumptions of two centuries of division could be reexamined.

Ironically, the branch of social science that had first advocated the integration of stories and numbers has yet to achieve it. Program evaluation, the applied social science that blossomed with the War on Poverty of the 1960s, created textbook doctrine about the integration of *process* evaluations and *impact* evaluations (Weiss 1972). Process evaluations were designed to tell the story about how a program was (or was not) implemented as planned, while impact evaluations were designed to analyze numbers measuring the effects that the program caused to occur on the people exposed to the program.

In theory, both methods were to be used in evaluations of every type of social program, from Head Start to welfare reform. In practice, so many programs broke down during implementation that there was only a story to tell and no numbers to analyze. This repeated experience led to a rising distaste for impact evaluations

among the more story-centric process evaluators, who began to attack the wisdom of even attempting to generate numbers that could measure average effects of programs across large numbers of people (e.g., Schorr 1997). Rejecting homothetic views of program effects on the medical model, some evaluation researchers have cast the idiographic or contextually “situated” conclusion as the only kind of knowledge that is possible.

### *Medicine and individual differences*

Nowhere has the wall between numbers and stories in treatment evaluations been more rigid than in medicine. When the bloodletting of patients was first subjected to a numerical impact evaluation (Louis 1835), the study was widely attacked by physicians who objected to the basic idea of generalizing about treatment effects. (Note that this was the same year, in the same city, in which Quetelet first published the idea of using social statistics to describe the “average man.”) One critic of Dr. Louis’s “numerical method” of evaluating medical treatments, Dr. Benigno Juan Isidoro Risueno D’Amador, said that the method could lead to a substitution of “a uniform, blind and mechanical routine for the action of the spirit and individual genius of the [physician] artist” (quoted in Millenson 1997, 98-99).

Almost two centuries later, many would say that D’Amador was right: that medicine has become far too blind to individual differences and that the quest for average effects has produced a uniform approach that ignores the variability of responses to medical treatments. Few modern critics of the method would want to give up the many benefits of the numerical method of medical impact evaluations: a vaccine for polio (Smith 1990), a cure for tuberculosis (Streptomycin Tuberculosis Trials Committee 1948), and a life-extending treatment for AIDS are but a few examples. But even a pharmaceutical company executive has declared that most drugs “don’t work for most people”—in the sense that average effects may reflect successful treatment of only a minority of patients in a randomized clinical trial—and that individualized pharmacotherapy based on individual DNA differences will lead to much higher rates of successful treatment among all patients with the same disease (Connor 2003).

The era in which DNA differences might also be taken into account in evaluating the effects of social programs may not be far off, given the advances in the basic science of interactions between genes and social environments (e.g., Caspi et al. 2002). The precision contributed by such new measures, however, is likely to be more numerical than narrative. As long as people are highly influenced by situational-specific emotions in their reactions to attempts to influence their behavior, the need will be great for adding knowledge about a social individuation in program effects that may be equivalent to DNA differences in medical treatment effects. And as long as people generally feel more qualified to hold opinions about the causes and cures of human behavior than about the causes and cures of disease, it will be more important to address those opinions with individual-level stories as well as with numbers.

## A Story from an Experiment

One story that illustrates this point comes from an experiment in a series of tests of a new program called “restorative justice” (RJ) (Braithwaite 2002; Strang 2002). The story describes the way the program works and sets the stage for proposing the structure of an experimental ethnography research design to examine key questions about RJ—or any other randomized trial that requires consent from one or more parties to create eligibility.

RJ offers new values and processes for society’s response to crime. New values, endorsed by the United Nations and many religious organizations, stress healing over punishment, reconciliation over anger, and reintegration over rejection. New processes include participation by victims, offenders, and all persons affected by a crime in decisions about how offenders should repay their debts to society—and to their victims. The processes also provide a forum for expressing the emotions about a crime that are intentionally suppressed by the fact-finding procedures of conventional criminal justice.

The research question of ten separate tests involving more than 1,000 crime victims and offenders in England and Australia (Sherman and Strang 2004) is whether RJ can change people’s lives for the better. Can it cure the post-traumatic stress and improve the health of crime victims? Can it help offenders to stop committing crimes? Can it motivate offenders to accept drug treatment, get jobs, and turn their lives around? And can such effects persist over the life course, lasting decades rather than a few months or years? All of these are questions that can be studied both quantitatively and qualitatively, although research funding tends to be limited to quantitative measures.

In its best-known form, RJ consists of a two- to three-hour conference led by a police officer or other trained facilitator. The agreements reached at these conferences may either take the place of formal court sentencing as a diversion from court, inform court sentencing prior to judges’ deciding the sentence, or follow court sentencing as a supplement to prison or probation.

### *A conference in prison*

One such conference took place in a London prison for women in late 2002. The conference was about the robbery of a sixty-something woman nurse at the doorstep of the emergency room where she worked. A young woman—a crack-cocaine user—approached the nurse to grab her purse. When the nurse resisted, the offender hit her over the head with a glass Coca-Cola bottle. The nurse staggered into the nearby emergency room, where she was given more than sixty stitches to heal the cut on her scalp. The nurse returned home, where she remained a recluse for over a year. During that time, she refused to buy another handbag or to leave a room without a member of her family escorting her.

When the robber was caught and pled guilty, a Scotland Yard police officer asked whether she would be willing to meet with the victim to discuss the crime

before her sentencing date in Crown Court. When she agreed, the officer asked the nurse whether she would also consent to a meeting. In both cases, the officer explained that there would be only a 50 percent chance of a meeting, even if both parties agreed because of the experimental nature of the meetings. The officer also said that agreement to meet in principle was just as valuable for the research as an actual meeting and that both victim and offender would make an important contribution to improving justice by their agreement to participate in the study. With this fully informed consent, first the offender and then the victim agreed to participate in the randomized controlled trial (RCT). Immediately after the victim gave her consent, the London officer called a research office in Philadelphia, where a staff member consulted a computerized formula that selected the case for a RJ conference.

---

*Ironically, the branch of social science that  
had first advocated the integration of stories  
and numbers [program evaluation]  
has yet to achieve it.*

---

After a series of telephone calls to family and friends of the victim and offender, all the parties arrived at a small room in the prison, where the offender was in custody awaiting sentencing. The police officer who would be the facilitator of the conference sat everyone down in chairs arranged in a circle, introduced all the participants, and outlined the discussion that would take place. He then asked the offender to tell everyone present what she did; she responded with a brief, shamefaced account of the crime. He then asked the victim to describe the harm caused by the crime to her and to others. While the victim said very little, her family members and friends spoke at length. Similar statements were then offered by the offender's family.

Supporters of both the offender and the victim expressed a great deal of anger, including the offender's grandmother. "How could you do this to somebody's else's granny?" she demanded. "Would you ever want someone to do this to me?"

The offender repeatedly asked the victim to forgive her, but the victim remained silent. When the harmful effects of the crime had been fully aired, the facilitator asked the entire group what the offender might do to repair the harm that the crime has caused. At this stage in the conference, victims often ask for the offenders' commitment to turning their lives around. Offenders often suggest ways in which they might do something for the victim, such as donating money to a charity

or performing a community service. Only rarely does the group agree on an actual restitution payment in cash to the victim. In this case, the group asked that the offender commit herself to getting off drugs and never going back to them.

Once the agreement was reached, the facilitator called a recess in the proceedings to write up the agreement. The participants stood up to get tea and biscuits from a sideboard. Supporters of the victim and the offender, as usual, talked informally over tea about common concerns, such as raising children. The normal release of tension was interrupted in this case, however, by the victim calling the offender over to her seat, where the victim had remained motionless throughout the two-hour conference. The offender complied.

The victim took the offender's hands, pressed them to her forehead, and began to pray—loudly and at length. She asked divine guidance for the offender and offered her own forgiveness as a means to help the young woman find the path to God's mercy and redemption.

The next day, the victim bought a handbag and later went back to work. The offender was sentenced to five years in prison.

#### *Counterfactuals: The conferences that did not happen*

So far, this story is consistent with the numerical evidence available. Robbery victims who volunteer for and participate in RJ conferences in this RCT are significantly less likely to suffer post-traumatic stress disorder and more likely to maintain (or resume) their normal employment patterns than are victims who volunteer for but are not randomly selected to participate in RJ conferences (Angel 2003; Strang 2004). It is only possible to draw that conclusion, however, by comparison of the stories of those victims who have the conferences and victims who do not.

Ethnographic studies of programs rarely make these comparisons, qualitatively or quantitatively. Ethnography naturally focuses on the people participating in a program and tells the story of how the program intersects with their lives. The logic of causation implicit in a chronological narrative is a before-and-after comparison. Interpreted through the language of a participant, the program observer can suggest plausible hypotheses about the causal links between program event *X* and human development *Y*. But these hypotheses cannot be tested rigorously within the cases participating in the program because there is no comparison to those who do not participate.

It is only by comparison to those with no program, or a different program, that we can reach valid conclusions about the effect of a program relative to some other option (Cox 1958; Campbell and Stanley 1963). Otherwise, causal analysis is vulnerable to the fallacy of *post hoc, ergo propter hoc*. Only a comparison to what happens in the absence of a program can rule out the possibility that a benefit would have occurred anyway. In this story, for example, the victim might have been just as likely to go back to work and buy a handbag even if she had not participated in an RJ conference. The only way to know if the effect of RJ is to increase the chances of victims' going back to work—on average—is to compare victims who had RJ to victims who would have, but did not, come to an RJ conference. While we can never

know what would have happened in the individual case if the treatment had not been available, the great strength of an RCT design is that it allows us to estimate the average effect across the entire range of cases in treatment and control groups.

Nothing in this logic requires an RCT design, however, to measure results in numbers. In principle, it is possible to imagine an RCT in which all events after random assignment were studied in purely qualitative ways. Numbers merely simplify and summarize the comparisons of program cases to counterfactual cases. For someone steeped in factual narrative, such as a historian, it is quite manageable to interpret the effects of a program on a hundred cases, with fifty in the treatment group and fifty in the control. Without ever committing a single number to the page, a qualitative analyst comparing the RJ and control group would still “commit a social science.” It is the logic of the RCT design, and not any particular method of data collection, that gives it internal validity—as long as most cases are treated as randomly assigned and as long as all cases are analyzed by intention to treat rather than by treatment actually received (Boruch 1997; Sherman and Strang 2004).

The use of qualitative materials in quantitative social science is often designed to “illustrate,” or to bore deeper into, patterns that are evident from quantitative summaries. Thus, for example, in a study of five hundred juvenile delinquents through age seventy, Laub and Sampson (2003, 9) drew a sample of fifty-three men for interviews, stratified on the basis of the three lifelong patterns of criminal offending (“desisters,” “persisters,” and “intermittents”). But as these interviews revealed, hypotheses emerge from such qualitative material that would be very interesting to test. In the case of retrospective interviews after quantitative data have already been collected and analyzed, the feasibility of testing new hypotheses is limited.

In the case of prospective ethnography, however, it is possible to both develop and test grounded theory (Glaser and Strauss 1967) as the research progresses. The hypotheses that are generated from interviews or observations of one case can immediately be tested against new data on the same hypotheses collected on other cases. Even if these hypotheses and their tests are later reduced to quantitative form, the fact that they would not have emerged without ethnographic work provides a strong justification for the added cost and effort of experimental ethnography. In our story, for example, the severity of the victim’s reaction to the crime may suggest this hypothesis: the magnitude of potential benefit of RJ on the victim’s mental health is directly proportionate to the magnitude of the harm the victim suffered from the crime. Because only limited quantitative data were being gathered on the severity of victim harm, the introduction of qualitative evidence offered both a way to “discover” this grounded theory and a way to test it—on average, in constant comparisons between treatment and control groups.

This kind of testing is best done when all the cases in an experiment can be included in an ethnography—an expensive but not unimaginable strategy for at least smaller samples of a hundred or so. Even the use of a much smaller sample, however, has the potential to generate hypotheses that could be tested against more comprehensive but less direct measures. If only ten cases—five treatment and five controls—could be studied ethnographically out of a sample of a hundred,

for example, the selection of those cases (preferably at random from within each treatment stratum) for detailed “boring down” could yield hypotheses that could be tested in future questionnaire interviews sought from the entire sample of a hundred cases. Just as Laub and Sampson (2003) used a 10 percent sample to illustrate quantitative patterns retrospectively, experimental ethnography could use 10 percent—or even less—samples to generate hypotheses for testing prospectively.

To the extent that one ethnographer can compare cases from both treatment groups, it may be easier to detect hypotheses that emerge only from comparisons. With five program cases and five control-group cases, major differences in treatment effects may well emerge through qualitative work that might never appear in

---

*In the case of restorative justice, the fact that  
both offenders and victims must give consent  
multiplies the possible comparisons.*

---

preconceived quantitative measures. In some RJ conferences, for example, victims have said that they were moving away from the city because of the crime. Victims in the control group may have done the same. It may be only from the frequent contact with victims that ethnographic work entails that a comparison of rates of moving away because of the crime (at least according to victims’ accounts) would become possible. This would be harder to estimate in a 10 percent sample than in a full sample, but if it were a very large effect, it would at least be noticed and be measurable against other kinds of data. With a good relationship to research subjects, ethnographers may also get a call from a victim who moves back into town after a brief period away. Constant comparisons by each ethnographer between treatment and control cases may lead to discovery of many such hypotheses.

#### *Comparisons that did not happen*

What RCTs usually fail to do is to make enough comparisons. Beyond the volunteers who are treated as assigned (control or treatment groups), there are many other people whose lives are affected by a RCT. This critique has several components. One is comparisons of people who volunteer to those who do not. This comparison is extremely important for external validity, in knowing how far to generalize the results of a program evaluation. That issue is especially important for programs likely to be imposed on the kinds of people who did not volunteer for the test phase. The results obtained from volunteers and nonvolunteers could be completely different. But neither medical nor social program evaluations invest much,

if any, research effort in studying those who do not volunteer and in comparing them to those who do.

In the case of our story, there would be three possible comparisons. One is between all volunteers and all refusers, on average. A second would be between those volunteers who attended RJ conferences and those who did not volunteer. The third would be between volunteers assigned to the control group and refusers. All three comparisons would be biased by self-selection and would lack the internal validity of a randomized design. But both comparisons would yield insights from qualitative exploration that could be important in understanding the limits, or even the potential, of the program being tested.

Imagine, at the case level, the kind of data that could emerge from ethnographic comparisons of volunteers and refusers. What is the story of at least one refuser? How many days or years did that victim who refused participation suffer from the robbery—missing work, getting divorced, moving to another country to escape the associations of the crime? How was her child-raising affected? Her religious faith or respect for the law? Did she herself commit any crimes in the future? Or did the ones who volunteered get over the crime much more quickly and suffer less long-term effects than those who participated? And with similar stories from other refusers, how typical do any of the stories seem to be? If cardiac epidemiologists study the risk of death in men who differ in how angry they get (a self-selection or at least not a randomly assigned difference), then why not study the correlates and stories of self-selection out of a potentially beneficial program?

Similar comparisons can be made in asking questions about how volunteers assigned to the control group differed from refusers. Were the controls worse off than those who refused consent? Did they get their hopes up about what could happen in an RJ conference, only to have them dashed by their assignment to the control group? Or were they relieved that they did not actually have to go to prison to sit down with someone who hurt them? What happened in their life stories from the time they were assigned to the control group until some years later? Whether by numbers or narrative, this comparison is relevant to understanding the effect of general adoption of the program, especially if the program may not always be possible after a victim agrees.

Another comparison too often omitted in RCTs is between those who drop out and those who do not—except that such comparisons alone are sometimes used, in error, to estimate the true treatment effects, which can create disastrous selection bias and terribly mislead the research audience (Gorman 2002). Victims who consent and are assigned to the program but then change their minds may suffer the most. They may have more remorse about having refused a sure thing, never knowing why the criminal chose to rob them and never hearing the apology that they thought was their due (Strang 2002).

In the case of restorative justice, the fact that both offenders and victims must give consent multiplies the possible comparisons. All of these comparisons between people who attend conferences and those who do not are tripled, depending on the reason that they did not attend conferences. Offenders as well as victims who volunteer are placed by random assignment into the experimental and control

groups. In addition, the “would-have-if-could-have” offenders who consented but whose victims did not join the victims who would have consented but their offenders did not. Since offenders are asked first, and victims are not approached unless offenders say they would consent, some victims never know that they lost the opportunity for a conference. Police tell the offenders who consented that they will not have a conference, although they do not say why. Placing the story in the full context of counterfactual stories would require examination of fourteen groups and 196 pairwise comparisons (14 squared) logically possible across the groups:

*Offender Groups:*

- O-1. Offenders who do not consent and whose victims would have consented had they been asked.
- O-2. Offenders who do not consent but whose victims would have refused had they been asked.
- O-3. Offenders who did consent but whose victims refused when asked.
- O-4. Offenders whose victims said yes but whose cases were assigned to the control group.
- O-5. Offenders whose victims said yes and whose cases were assigned to have an RJ conference, which was held with a victim present.
- O-6. Offenders whose victims said yes, whose cases were assigned to have an RJ conference, but who changed their minds and refused to meet with the victim.
- O-7. Offenders whose victims said yes, whose cases were assigned to have an RJ conference, but whose victims changed their minds and refused to meet with the offender.

*Victim Groups:*

- V-1. Victims who would have consented but were never asked because offenders refused.
- V-2. Victims who were never asked but would not have consented even if they had they been asked.
- V-3. Victims whose offenders consented but who refused their own consent.
- V-4. Victims whose offenders consented and who gave their own consent but whose cases were randomly assigned to the control group.
- V-5. Victims whose offenders consented, who gave their own consent, whose cases were randomly assigned to the RJ group, and who participated in an RJ conference.
- V-6. Victims whose offenders consented, who gave their own consent, who were randomly assigned to an RJ conference, but whose offenders changed their minds and refused to meet with the victim for RJ.
- V-7. Victims whose offenders consented, who gave their own consent, who were randomly assigned to an RJ conference, but who changed their minds and refused to meet with the offender for RJ.

These categories are depicted in Table 1.

Most of the 196 possible pairwise comparisons in the fourteen victim and offender groups would not be of much theoretical or policy interest. Comparisons between victims and offenders, for example, might not be of interest to an analysis focusing on victim effects or to one focusing on whether RJ reduces repeat offending. The main comparisons of interest would be between groups 4 and 5, for both victims and offenders, respectively, with the other groups compared with both 4 and 5. Smaller investments could also be made in studying the groups other than 4 and 5. Just how large or small an investment in ethnography would be appropriate,

TABLE 1  
TREATMENT AND COMPARISON GROUPS IN A  
CONTROLLED TRIAL OF RESTORATIVE JUSTICE

Victim and Offender Group No.	Offender Consent	Victim Consent	Random Assignment	Offender Change	Victim Change
1	No	Yes			
2	No	No			
3	Yes	No			
4	Yes	Yes	Control		
5	Yes	Yes	Program	No	No
6	Yes	Yes	Program	Yes	No
7	Yes	Yes	Program	No	Yes

or possible, in any of the groups merits a separate discussion of the sampling issues inherent in the logic of an RCT design.

### Sampling for Experimental Ethnography

The question of “how many cases” bedevils qualitative social science. The direct tradeoff between depth and breadth is unavoidable, given a fixed amount of resources. Whether it is a choice of the lone scholar, with only so many hours to work on a research project, or of a research team with many members, the tradeoff is the same: very detailed insights on a small and perhaps atypical sample of people versus less detailed insight on a larger and perhaps more typical cross-section of people relevant to the research question.

In the case of experimental ethnography, the question of how many cases can become more sharply focused. Rather than seeking to represent the experience of a huge class of people, such as all poor black men in big cities or all women who work in suburban offices, experimental ethnography is challenged to say something only about the specific categories of people in the “pipeline” of the experiment’s universe of eligible cases (Boruch 1997, 14). The sampling frame and the universe are identical for experimental ethnography, a truly rare opportunity for ethnography to be based on systematic sampling methods. The challenge of creating probability samples remains immense, of course, even for relatively small experiments with only fifty cases per treatment group. The ratio of potential cases to available resources may result in a sample that is too small to create a stable estimate of population parameters.

One solution to this problem is to focus entirely on the cases in the randomly assigned groups. With ten ethnographers and a one-year time frame following up an experiment in criminal sanctioning, for example, each ethnographer could be

assigned ten offenders to interview or observe each week. By design, five of the offenders in each ethnographer's sample could be taken, by random selection, from the randomly assigned treatment group, while the other five in the sample could be taken from the control group. This would allow each ethnographer to be making "grounded theory" comparisons (Glaser and Strauss 1967) as they gathered their data, sharpening the things they look for in each treatment group by real-time comparisons to the other. While the cost of ten ethnographers for one year is not trivial, it is not inconceivable. The cost-benefit ratio would be especially high for evaluating a treatment, like RJ, in which the accumulation of findings shows highly diverse and unpredictable responses of different kinds of people (or offense types) to the same treatment (Bottoms, Gelsthorpe, and Rex 2001, 229).

---

*The sampling frame and the universe are identical for experimental ethnography, a truly rare opportunity for ethnography to be based on systematic sampling methods.*

---

Even a 10 percent sample of cases, or smaller, as discussed above, would serve the role of generating hypotheses and of illustrating the patterns of reaction. Sampling for ethnographic cases on the basis of initial reactions to random assignment would sharpen the differences to be explored and about which more grounded theory could be developed.

The main argument in favor of focusing only on the randomly assigned experimental sample is that it would increase the internal validity of the test. This is true at the exterior of the black box, in which the experimental design rules out alternative rival hypotheses at a quantitative level. It is also true within the black box, where the qualitative level could more precisely describe the causal mechanisms that allow the treatment to succeed or to fail in changing behavior. The combination of the two would be virtually unprecedented at the level of cases as randomly assigned.<sup>1</sup>

Another argument in favor of focusing on treatment and control groups is that it offers a clear basis for sampling on theoretical grounds based on early differences in experience. In RJ conferences, for example, most offenders and victims come away fairly well satisfied (Strang 2002), but a small minority come away unhappy. Similarly, among control-group cases some are angrier about the conventional process than others. If only ten victims and ten offenders can be included in an ethnography, the use of initial posttreatment interviews with the full sample could allow

the selection of an ethnographic sample on the basis of treatment evaluation. In each group of ten, six “satisfied” (three treatment and three control) and four “unsatisfied” (two treatment and two controls) could be selected for ethnography. While these samples are so small that they could easily generate atypical results, there would still be more basis at the outset to expect differences that would affect the comparison of treatment and control groups.

The argument against focusing only on the experimental sample is that it would limit the external validity that could be gained from experimental ethnography. The unasked questions about the groups excluded from the sample (by self-selection or someone else’s decision) could be best explored through qualitative methods. While a quantitative analysis could examine differences in demographic characteristics, health or crime outcomes, and other officially recorded variables, it would not be possible to use such records to tap into emotions and life experiences. Only ethnography—prospective and ongoing, by definition—would be likely to learn what questions should even be asked about why and with what consequences people do or do not consent, for example, to participate in an experimental test of a new idea. When the process of consent is conditional upon the consent of others, the various pipeline categories become even more difficult to understand without the open-ended scope of ethnography.

The external-validity argument is simultaneously strengthened and challenged in situations of high refusal rates. The argument is strengthened by the greater need to understand why people refuse to participate. The better the benefit that may result from an experimental test, the greater the need to understand why people may refuse to accept the benefit. This argument for experimental ethnography is challenged, however, by the problem of the large numbers to be sampled. Whenever the number of refusals exceeds the numbers of consents, the use of ongoing ethnography with refusal cases becomes more problematic.

In the 2002-2004 Thames Valley postsentencing RCTs of restorative justice, for example, the numbers of eligible cases outnumbered the randomly assigned cases by almost five to one. Of the 706 eligible offenders, 699 could be contacted to request their consent. Of the 699 contacted offenders, 443 (73 percent) consented. Of those 443, victims could be contacted in only 367 cases (83 percent) and consented in only 147 cases (33 percent of all 443 offender consents and 40 percent of victims contacted). With only some 100 offenders and victims lost to contact, almost 450 people refused to participate. Ethnography on all of them would be impossible (unless the authors won the lottery). Yet even sampling 10 people each from categories 1, 2, and 3, for offenders and victims (see Table 1) would require six ethnographers, a high price for such a small sample of each large group. One ethnographer, on the other hand, could handle two cases from each category.

One justification for ethnography in such a context would be the generative function of identifying research questions. Neither the ethnography nor the experimental design could identify causal relationships between the offenders’ (or victims’) decisions to refuse consent and some later developments in their lives. What the ethnography could do, however, is to explore how the kinds of people who refused, their circumstances, and the circumstances of their offenses differed from

the kinds of people and circumstances where consent was forthcoming. This exploratory analysis could lead to future analyses drawing on quantifiable measures that could be gathered at low cost per case.

One option for adding qualitative information to the universe of cases in the experimental pipeline would be to ask each operational person seeking consent to prepare a short memoir of each case. This memoir could be required to follow a standard set of questions, such as how the offenders felt about the crimes, the victims, their lives, and their futures, and how the idea of participating in the experimental treatment looked to them in light of all those factors. The items could also cover the social context of the decision: who in their family or friendship network did they consult in making the decision? Who was in favor, who was against, and why? These are all questions that ethnographers could ask and interpret with greater objectivity. But for direct recall of the emotions and discussions at the time, in relation to cost of measurement, enlisting operating people after the fact may be a compromise approach to collection of qualitative materials at manageable cost from people who did observe the research participants prospectively and directly.

## Experimental Ethnography and Systematic Reviews

A further justification for focusing ethnography on the experimental sample is the increasing use of systematic reviews of randomized trials (Chalmers 2003). With each experiment seen as merely one tree in a forest of evidence, the importance of knowing how each tree compares to all others becomes even greater (Sherman and Strang 2004). Rather than writing off the average effect of a series of RCTs as negligible or negative, it may be more useful to isolate the one or two most successful RCTs and determine how they differed from the majority. It is possible or even likely that the difference in question was merely due to chance. But there may also be substantive differences in the sample or the way the treatment was delivered that could help explain the difference in outcomes. Those substantive differences may even point to refinement of the treatment for future RCTs, treating the accumulated knowledge as a trial-and-error process of invention rather than as a verdict based on the average result.

At present, the methods of systematic reviews do not accommodate the kinds of data that ethnography could generate. Yet there is no reason such methods could not be developed. Sensitivity analyses can already be done by quantitative sorting of the RCTs on various criteria, such as the nature of the control group comparison (see Fonagy et al. 2002; Assendelft et al. 2003). Similar methods could be used for qualitative data.

RCTs that enjoy the benefit of experimental ethnography could be coded in very rich and surprising ways, such as a strong tendency of the control group to resent the fact that they were not randomly assigned to the treatment group. While such resentment is occasionally suggested (e.g., Killias, Aebi, and Ribeaud 2000), it

seems unheard of to study it systematically. How this hypothesized resentment might play out in the lives of the control group, or disappear after an initial complaint or two, is exactly the kind of question that ethnography could raise. Comparisons to samples of similar offenders who were never eligible for the randomly assigned experimental treatment might also help inform such analysis. Since the issue of control-group reactivity is one of the most fundamental threats to the external validity of RCTs, this task should be reason alone for funding experimental ethnography.

---

*The rising power of systematic reviews to dominate the policy conclusions about innovative ideas makes it all the more important to get them right.*

---

The rising power of systematic reviews to dominate the policy conclusions about innovative ideas makes it all the more important to get them right. Black-box conclusions may be much more likely to mislead, especially if the contents of the black box actually vary. While some quantitative measures can examine the consistency of black-box causal mechanisms, few would doubt the increased insight that could be gained from ethnographic materials.

Some may argue about how large a sample would be needed to generate reliable insights about the causal processes operating with the randomly assigned groups. The possibility of placing an ethnographer with every research participant—victim or offender in our example—should cut short such an argument. Even with a 50 percent sample in RCTs of fifty cases per treatment group, many would place great credence in consistent conclusions found across a research team of five or ten ethnographers each studying both experimental and control cases.

The more detailed the descriptive material gathered and published about each RCT, the greater the possibility for systematic reviewers to comb the details looking for ground theory (Glaser and Strauss 1967) about why some RCTs produced better outcomes in the treatment group than other RCTs did. This “tertiary” analysis of the data from multiple RCTs would strengthen the use of RCTs to discover things that do work and not just to reject the treatments that do not work.

Given the greater likelihood that more rigorous methods (compared with less rigorous designs) will find that promising ideas are ineffective (Weisburd, Lum, and Petrosino 2001; Glazerman, Levy, and Myers 2003), the risk of prematurely rejecting promising treatments may rise along with evaluation rigor. One protec-

tion against that risk could be experimental ethnography. As former Attorney General Janet Reno often said about negative program evaluations, “Please don’t just tell me that the crime prevention program doesn’t work. Please tell me *why* it doesn’t work, and how it *might* be made to work if it were improved in some way.”

It is hard for quantitative analyses of RCTs to respond to such a request from a public official who seeks out the guidance of social science. No one could be better placed than an experimental ethnographer to answer the call. After we know that in one RCT, or in a series of them, a program does not work, there still may be substantial public pressure to continue a program. Generating ideas for revising the program is just what experimental ethnography could do, before handing a revised approach back to the general experimental evaluators (also known as inventors) for further RCTs.

## The Prospects for Experimental Ethnography

This proposal requires two necessary and sufficient conditions to become a reality. One is that ethnographers and experimentalists be willing to work together. In our case, at least, we have invited several distinguished ethnographers to join us, and in principle, they have agreed to do so. Which brings us to the second necessary condition: funding.

The campaign to encourage evaluation funding to allow RCTs has been difficult wherever it has been mounted. Adding ethnography to the cost may just break the bank. Then again, it may just sweeten the package. In social science cultures that have strongly opposed RCT designs (e.g., Pawson and Tilley 1997), the addition of experimental ethnography may be just the compromise needed to bring greater value to the high costs of nonexperimental evaluations. Those who attack RCTs say that while they may tell what worked in one sample, they cannot tell why, and therefore, the RCTs’ value for developing externally valid policy conclusions is limited. Those who attack nonexperimental evaluations say that because they cannot tell what worked, their insights as to why a program functions the way it does will have little value. Rather than choosing between more “whys” than “whats” or more “whats” than “whys,” experimental ethnography could provide the win-win answer that works. What a story that could make!

### Note

1. Multisite randomized trials, with many cases randomized at each site, sometimes include qualitative observations about the site. But conclusions from such analyses do not directly address variability across individual cases within randomly assigned treatment groups.

### References

Anderson, Elijah. 1978. *A place on the corner*. Chicago: University of Chicago Press.

- Angel, Caroline M. 2003. Effects of restorative justice on post-traumatic stress symptoms of burglary and robbery victims in London: A preliminary report. Jerry Lee Center of Criminology, University of Pennsylvania.
- Assendelft, Willem J. J., Sally C. Morton, Emily I. Yu, Marika J. Suttorp, and Paul G. Shekelle. 2003. Spinal manipulative therapy for low back pain: A meta-analysis of effectiveness relative to other therapies. *Annals of Internal Medicine* 138:871-81.
- Boruch, Robert. 1997. *Randomized experiments for planning and evaluation*. Thousand Oaks, CA: Sage.
- Bottoms, Anthony, Loraine Gelsthorpe, and Sue Rex. 2001. *Community penalties: Change and challenges*. Devon, UK: Willan.
- Braithwaite, John 2002. *Restorative justice and responsive regulation*. Cambridge, UK: Cambridge University Press.
- Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.
- Cannadine, David. 1990. *Decline and fall of the British aristocracy*. New Haven, CT: Yale University Press.
- Caspi, Avshalom, Joseph McClay, Terrie E. Moffitt, Jonathan Mill, Judy Martin, Ian W. Craig, Alan Taylor, and Richie Poulton. 2002. Role of genotype in the cycle of violence in maltreated children. *Science* 297 (5582): 851-54.
- Chalmers, Iain. 2003. Trying to do more good than harm in policy and practice: The role of rigorous, transparent, up-to-date evaluations. *Annals of the American Academy of Political and Social Science* 589: 22-40.
- Connor, Steve. 2003. Glaxo chief: Our drugs do not work on most patients. *The Independent*, December 8, p. 1.
- Cox, D. R. 1958. *The planning of experiments*. London: Wiley.
- Fonagy, P., M. Target, D. Cottrell, J. Phillips, and Z. Kurtz, eds. 2002. *What works for whom? A critical review of treatments for children and adolescents*. New York: Guilford.
- Glaser, Barney, and Anselm Strauss. 1967. *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.
- Glazerman, S., D. M. Levy, and D. Myers. 2003. Nonexperimental versus experimental estimates of earnings impacts. *Annals of the American Academy of Political and Social Science* 589:63-93.
- Gorman, D. M. 2002. The "science" of drug and alcohol prevention: The case of the randomized trial of the life skills training program. *International Journal of Drug Policy* 13:21-26.
- Killias, Martin, Marcelo F. Aebi, and Dennis Ribeaud. 2000. Learning through controlled experiments: Community service and heroin prescription in Switzerland. *Crime & Delinquency* 46:233-51.
- Laub, John and Robert Sampson. 2003. *Shared beginnings, divergent lives: Delinquent boys to age 70*. Cambridge, MA: Harvard University Press.
- Louis, Pierre Charles Alexandre. 1835. *Recherche sur les effets de la saignée* [Research on the effects of bloodletting]. Paris: De Mignaret.
- Maruna, Shadd. 2001. *Making good: How ex-convicts reform and rebuild their lives*. Washington, DC: American Psychological Association.
- Massey, Douglas, and Patricia Denton. 1993. *American apartheid: Segregation and the making of the underclass*. Cambridge, MA: Harvard University Press.
- Mayhew, Henry. 1861-1862. *London labour and the London poor: A cyclopaedia of the condition and earnings of those that will work, those that cannot work, and those that will not work*. London: Griffin, Bohn.
- McCord, Joan. 1978. A thirty-year follow-up of treatment effects. *American Psychologist* 33 (3): 284-89.
- Millenson, Michael. 1997. *Demanding medical excellence*. Chicago: University of Chicago Press.
- Olds, D. L., C. R. Henderson, R. Cole, J. Eckenrode, H. Kitzman, D. Luckey, L. Pettitt, K. Sidora, P. Morris, and J. Powers. 1998. Long-term effects of nurse home visitation on children's criminal and anti-social behavior: 15-year followup of a randomized controlled trial. *Journal of the American Medical Association* 280:1238-44.
- Pate, A. M., and E. H. Hamilton. 1992. Formal and informal deterrents to domestic violence: The Dade County Spouse Assault Experiment. *American Sociological Review* 57:691-97.
- Pawson, Ray and Nick Tilley. 1997. *Realistic Evaluation*. Thousand Oaks, CA: Sage.
- Putnam, Robert. 2000. *Bowling alone*. New York: Simon and Schuster.
- Quetelet, Adolphe. 1835. *A treatise on man, and the development of his faculties*. Paris: Bachelier
- Reuter, Peter. 1983. *Disorganized Crime*, Cambridge, MA: MIT Press.

- Sampson, Robert, and John Laub. 1993. *Crime in the making*. Cambridge, MA: Harvard University Press.
- Schorr, Lisbeth. 1997. *Common purpose: Strengthening families and neighborhoods to rebuild America*. New York: Anchor.
- Sherman, L. W., and D. A. Smith. 1992. Crime, punishment and stake in conformity: Legal and informal control of domestic violence. *American Sociological Review* 57:680-90.
- Sherman, L. W., and H. Strang. 2004. Verdicts or inventions? Interpreting results from randomized controlled experiments in criminology. *American Behavioral Scientist* 47:575-607.
- Smith, Jane S. 1990. *Patenting the sun: Polio and the Salk vaccine*. New York: William Morrow.
- Strang, Heather. 2002. *Repair or revenge: Victims and restorative justice*. Oxford: Oxford University Press.
- . 2004. Victims and restorative justice. Presentation to Second Winchester International Restorative Justice Conference, *Restorative Justice Approaches: From Inspiration to Results*.
- Stouffer, Samuel A., Edward A. Suchman, Leland C. DeVinney, Shirley A. Star, and Robin M. Williams Jr. 1949. *The American soldier: Adjustment during army life*. Vol. 1. Princeton, NJ: Princeton University Press.
- Streptomycin Tuberculosis Trials Committee. 1948. Streptomycin treatment of pulmonary tuberculosis: A Medical Research Council investigation. *British Medical Journal* 20:769-82.
- Sulloway, Frank J. 1996. *Born to rebel: Birth order, family dynamics and creative lives*. New York: Pantheon.
- Valliant, George. 2002. *Aging well*. Boston: Little, Brown.
- Weisburd, D., C. Lum, and A. Petrosino. 2001. Does research design affect study outcomes in criminal justice? *Annals of the American Academy of Social and Political Science* 578:50-70.
- Weiss, Carol H. 1972. *Evaluation research: Methods of assessing program effectiveness*. Englewood Cliffs, NJ: Prentice Hall.