J Exp Criminol (2007) 3:299–321
DOI 10.1007/s11292-007-9044-y

The power few: experimental criminology and the reduction of harm

The 2006 Joan McCord Prize Lecture

Lawrence W. Sherman

Published online: 9 November 2007

Abstract The *promise* of experimental criminology is finding ways to reduce harm from crime and injustice. The *problem* of experimental criminology is that so few experiments produce evidence of big effects from the interventions they test. One solution to this problem may be concentrating scarce resources for experiments on the "power few:" the small percentage of places, victims, offenders, police officers or other units in any distribution of crime or injustice which produces the greatest amount of harm. By increasing the homogeneity and base rates of the samples enrolled in each experiment, the power few hypothesis predicts increased statistical power to detect program effects. With greater investment of resources, and possibly less variant responses to greater dosages of intervention—especially interventions of support, as distinct from punishment—we may also increase our chances of finding politically acceptable interventions that will work.

Keywords Crime forecasting · Crime prevention · Ethics · Experimental criminology · Hot spots · Power curves · Repeat offenders · Repeat victims

How can criminology produce more experiments with big effects in reducing harm?

This question is different from Joan McCord's lifelong concern with "cures that harm" (McCord 1978, 2003). Her discovery of the long-term harms caused by the well-intentioned and expensive Cambridge Somerville experiment still stands as a powerful exemplar for the principle of "First, do no harm—in crime prevention as in

L. W. Sherman
Jerry Lee Center of Criminology, University of Pennsylvania, Philadelphia, USA

L. W. Sherman (✉)
Institute of Criminology, Cambridge University, Cambridge, UK
e-mail: Lawrence.Sherman@crim.cam.ac.uk

medicine." The potential for any program to cause harm, no matter how sensible it may appear in theory, remains the primary ethical justification for experimental criminology (Federal Judicial Center 1981). Absent any other widely accepted means of creating unbiased estimates of treatment effects (Chalmers 2003), the only way we can be sure that treatments do not harm people is to subject those practices to randomized controlled experiments in field settings. One promise of such tests is that they will guide democratic societies in deciding what *not* to do, as one way to reduce human suffering.

The promise of experimental criminology, however, is far broader than detecting *unintended* harm. Its main objective is to develop programs to test their *intended* benefits of policies in reducing harm (Sherman 2006). No matter how much negative evidence we may produce as gatekeepers barring harmful programs, our best chance to reduce human misery is with positive evidence of programs that work well (Sherman 2000). By working well, I mean not only that they are modestly successful or cost-effective. I also mean that, in at least some cases, the results are *impressively* cost-effective. For it is by such major successes that our work can become more widely known, influential, and capable of helping to reduce the harms of crime and injustice.

Experimental medicine, for example, did not achieve its high level of financial support just from detecting cures that harm—although that has been impressive in itself. Great harm was prevented, for example, by keeping the drug thalidomide from being sold in the USA before it had undergone randomized clinical trials (Knightly et al. 1979). Yet medical experiments are more renowned for preventing human suffering through big effects at prevention and treatment of disease: preventing scurvy (Tröhler 2003) and cholera (Johnson 2006), discovering vaccines for smallpox and polio (Barry 2004; Smith 1990), or finding a cure for tuberculosis (Doll 1998; Hill 1990). Similarly, the yield of experimental criminology in reducing human misery could be far greater from *designing*—as well as testing—programs seeking major discoveries of ways to reduce crime and injustice, rather than limiting our role as experimental criminologists to the evaluation of programs designed by others (Sherman 2006).

In this tribute to Joan McCord, I would like to extend her concern with discovering treatments that *increase* harm to discovering treatments that *reduce* harm. My question is this: how can we do a better job? How can we produce more big discoveries in the limited time on earth that each experimental criminologist has to pursue this goal? How can we best address the reasons that most of our experiments fail to find the big effects that could make criminology in general, and experimental criminology in particular, more central to making policy for harm reduction?

My answer is this: let us focus our experiments on supportive treatments for the "power few" cases associated with most of the harm in any distribution of harmful events. This answer is a hypothesis, not a conclusion: the "*power few hypothesis*." That hypothesis suggests that experiments are most likely to discover big effects with test samples that are *unrepresentative* subsets of larger populations of places or people associated with any type of harm, even while those samples feature the majority of harm in their populations. Experiments focusing on such samples may also be more ethically and politically acceptable when they provide support, or even

regulation, than when they simply increase punishment. If that is as true of places as it is of people, such tests may help turn the discussion of criminal events away from a narrow focus on offenders.

This discussion proceeds in three parts. It begins with the *power curve problem* and its relationship with small effect sizes in experimental criminology. Next, I suggest a *power few solution*, by which experiments could capitalize on the non-linear nature of the distributions of harm to demonstrate much bigger benefits. Third, I consider the *ethical and political context* for developing and testing power curve responses to crimes and injustices. In all three parts, I address Gladwell's (2006) statement of the moral dilemma about the power few, concluding that power few solutions are no more problematic than any other response to crime. The fact that they may be the most effective responses may even make them morally preferable, especially when they take the form of support or regulation rather than punishment.

The power curve problem

Gladwell's choice

New Yorker Magazine writer Malcolm Gladwell (2006) recently described innovations in Denver, Reno and Seattle for dealing with mentally disturbed homeless people. The key illustration was a man named "Million Dollar Murray," who was costing the city of Reno (NV, USA) over one million dollars in just the costs of emergency health care services he consumed. These costs did not include police, social services, jails, prosecution, legal aid and probation expenses. Gladwell described Murray as an example of the "power few" homeless people who were consuming a highly disproportionate share of the total resources available to help the homeless. The data he summarized challenged the prevailing paradigm of a normal distribution, in which the problem of homelessness was thought to incur roughly equal costs per homeless person.

This result was no surprise to students of what economists call the "Pareto curve," or the non-linear "hockey stick" skewed distribution J-curve, in which a small fraction of the units in a population account for a large part of the total volume of any characteristic of that population (Zipf 1949; Simon 1955; Eck et al. 2007). This phenomenon has been popularized as the "80–20" rule (Kock 1999), in which 80% of the volume of outcomes in a group are generated by 20% of its members. Such curves are actually far more variable in their degree of concentration (skewness) than a simple "80–20" rule implies, with some as skewed as 100% of a problem found in 3% of units at risk, and others with no apparent skewness at all (Sherman et al. 1989).

Many cases of nonlinear skewness have been found in criminology. In the distributions of frequency of crime across units associated with criminal events, in fact, it is hard to find any normal distributions at all. In what we might call "the criminology of this and that," a small proportion of "this" unit produces a high proportion of "that" kind of crime in repeated settings:

- A few juvenile delinquents produce the majority of all arrests in a birth cohort (Wolfgang et al. 1972).

- A few street addresses experience the majority of most crimes in a city (Sherman et al. 1989).
- A small proportion of places of a certain kind, such as taverns or apartment buildings, experiences a large proportion of offenses of a certain kind (Eck et al. 2007). For example, about 10% of the taverns in Milwaukee in the late 1980s produced over half of all the violence occurring in taverns.
- A few victims account for a large portion of all victimizations (Farrell 1995).
- Some 15% of all police officers produce over half of all arrests resulting in criminal convictions of criminals (Forst et al. 1978).
- A small percentage of Los Angeles police officers generate a large portion of all complaints against police and generate the majority of reports stating that they have used force against citizens (Independent Commission on the Los Angeles Police Department 1991).
- Only 15% of the street blocks in Seattle accounted for almost all of the city's crime drop in the 1990s (Weisburd et al. 2004).
- Only 35 out of 9,589 street blocks in Brooklyn consume over one million dollars a year in the cost of incarcerating their residents (Wagner 2005).

Most criminologists would not be surprised that a handful of homeless people consumed the majority of the costs of serving the homeless. What would surprise most criminologists is what the City of Denver did in response to its power curve analysis. After calculating the cost of providing *equal* services to all homeless people on demand (or upon causing trouble to other people), Denver decided to provide an *unequal* investment of resources in the *most expensive* people. For these "power few," the policy Denver adopted was to pay for each one of them to have a private apartment, a full-time caretaker, and a program of interventions that would provide daily support for their staying out of trouble. This support included employment that the caretaker helped them to get and keep–the latter being harder and more important. The program included caretaker management of their clients swallowing mental health medications on a regular basis, yielding far greater compliance than when the clients were homeless and contemptuous of shelters. This supervision meant that "forgetting to take their meds" would become a less frequent cause of a psychotic episode that scared tourists or generated expensive calls for police to take the clients into custody, treat them in an emergency room, and send them back out to the streets unsupervised until the next episode.

What surprised Gladwell about this innovation is not that it was expensive, but that it was so unequal. Many people would have welcomed such care, but they were not offered it because they did not cause enough trouble. The implicit reward for bad conduct, Gladwell suggested, is both politically and morally offensive. This is true, even though this "Cadillac service" for the power few actually reduced their total cost. Lavish spending on *prevention* was far cheaper than minimal spending on *response* to one client's outbursts. By Gladwell's analysis, Denver's difference in costs per "power few" case was $15,000 per year in prevention vs $45,000 per year in response. This did not even count the costs of scaring people away from city center commerce, which hurt city sales taxes and convention bookings. Nor did it count injuries to police officers and others trying to subdue the clients when they were living on the street without regular medication.

Gladwell (2006, p. 104) found all of this perfectly rational and morally reprehensible. Framing the choice in terms of the benefits to the clients, rather than to their victims or to the entire community, he presented a standard Kantian argument against using people as means to an end, rather than as an end in themselves. If each person has equal worth, the allocation of resources to them should be equal—even if it causes more harm to the community. Using these principles, he generalized his conclusion about the Denver program to its possible application for any other kind of power curve:

> Power-curve problems leave us with an unpleasant choice. We can be true to our principles or we can fix the problem. We cannot do both.

That conclusion, of course, depends on just which moral principles we adopt, as I argue below. It also depends on the premise that power curve solutions will always work, which is far from clear without controlled experiments. Denver's power curve solution to a problem of law and order has substantial implications for the ways in which our experiments could be designed. Before we consider Gladwell's moral conclusion, then, let us consider the evidence for his empirical conclusion that such power curve solutions can indeed "solve the problem."

Bell curve experiments, small effects

The problem, for us, is that experimental criminology so often finds that policies and programs have weak, or no, benefits. That conclusion can be drawn from Farrington's reviews of 35 randomized experiments in crime and justice reported in 1957–1981 (Farrington 1983), of 83 more reported in 1982–2004 (Farrington and Welsh 2005), and in his 2005 McCord Lecture on a systematic review of the effects of community-based prevention experiments (Farrington 2006). It is also the conclusion reached by Weisburd et al. (2001) in their analysis of 308 program evaluations summarized in Sherman et al. (1997) and in 68 evaluations summarized in MacKenzie and Hickman (1998). While there are some important exceptions, the typical effect size of the results of experiments meeting the eligibility criteria for these reviews is small or near zero.

Farrington and Welsh (2005, pp. 28–30) used meta-analyses to go beyond significance tests in each study. While this approach yielded more optimistic conclusions, they nonetheless found the count of positive results to be "depressing:"

> .....only 16 out of 83 experiments produced significantly desirable results, with two nearly significant and four producing significantly undesirable results. While 16 out of 83 is much greater than the chance expectation of 4, these results do not seem impressive.

Weisburd and his colleagues place the findings for randomized experiments in the larger context of other research designs. In general, they show that the greater the internal validity of the research design, the smaller the average (positive) effect size. While they do not claim that this proves randomized experiments necessarily offer more *accurate* assessments of program effects, they do provide evidence for a "kill-the-messenger" conclusion: that program advocates might wish to avoid randomized experiments if at all possible. On average, we random assigners seem likely to be

bearers of bad news. And that, in turn, is bad news for experimental criminology, reducing our chances of even doing experiments, let alone reducing harm.

The limitation of such assessments is that they lack a control group. What we do not know is the odds of an experiment succeeding in trying to reduce any other kind of harm, from cigarette smoking to illiteracy to automobile accidents. It is even conceivable that criminology is doing better than other fields in its ratio of successes to failures, low as that ratio may be. If we are unimpressed with our average results in an absolute sense, it may be due to evidence-free expectations that we should "win" all or even most of the time. Experiments are not like team sports, in which one side must win in every game. That metaphor implies that we might win in half of all experiments, just by chance. The history of experimentation shows otherwise, with many examples of success rates below 1% of all tests. Edison's apocryphal comment on this point was that his many "failed" experiments in a long-lasting filament for a lightbulb were all "successes" because they showed what materials would *not* work. Josiah Wedgwood's carefully documented notebooks show over 5,000 tests of different ways to apply a glazing to china plates and crockery (Uglow 2003).

It is instinct rather than evidence, however, which plagues experimental criminology. The instinctive reaction is that we should try to "win" the struggle to reduce harm in each and every experiment. Rather than denying that instinct, we could make the most of it by raising the bar of experimental design. That bar could be based on the evidence of what may explain the generally small effects of randomized experiments on crime and justice. That evidence suggests that, while there are three hypotheses about what causes *small* effects, there is one constant factor among experiments with *big* effects: they have all focused on "power few" samples, at least to some degree. Although some power few experiments have also failed to show big effects, the rate of "big" success may be higher with such samples than with less selective samples. Raising the rate of big successes could do much to increase the volume and yield of experiments in criminology.

Three hypotheses

If we look only at the reasons why experiments fail to show big effects, there are at least three plausible hypotheses. It is only when we look at the few cases of big effects, however, that we make better sense of these hypotheses and can integrate them into a coherent proposal. Stated systematically, these hypotheses are as follows:

1. Randomized experiments are simply more likely than other evaluation designs to detect *weak programs*, when weaknesses are due to poor theory, poor integrity of implementing a theory, or insufficient dosage of a theoretically powerful intervention.
2. Our experiments have been focused on samples that are *too heterogeneous*, increasing the variance in subject response to intervention and reducing statistical power to detect truly powerful effects on some subgroups but not others.
3. Our experiments have been focused primarily on *low-harm units*, when most of the crime problem is concentrated in the "power few" of high-harm units.

These hypotheses can be summarized as (1) weak programs, (2) diverse samples, or (3) low-harm samples.

Weak programs? Farrington's (1983) initial discovery of the generally weak effects of randomized experiments led him to suggest that the results may be due to weak programs. As he noted, programs tested with random assignment can be weak in a variety of ways. One is that the political difficulty of randomly assigning the "strong" interventions in crime and justice, such as arrest and incarceration, may have limited experiments to testing relatively marginal interventions in the lives of offenders or potential offenders. That limitation would be an example of a *weak theory*. For example, recidivism differences between delinquents sent to a boot camp and those undergoing standard confinement (Farrington and Welsh 2005, p. 20) may be near zero, because boot camps are really no more "severe" than standard confinement.

Policies can also be based on strong theory, but experiments can implement the theories with less than complete *integrity*. In the Minneapolis hot spots patrol experiment (Sherman and Weisburd 1995), for example, the strong difference in patrol dosage between experimental and control hot spots in the first 7 months broke down during summer conditions of more crime and fewer police. This breakdown yielded much stronger effects for the period with complete integrity when compared to the full 1-year experiment (Sherman and Weisburd 1995). When the dosage difference was as strong as the theory required, the crime prevention effects were large. When the dosage difference between the groups disappeared, so did the difference in crime prevention.

Finally, weak programs may be based on a strong theory that is well implemented, but *insufficient dosage* may be available for providing a valid test of the theory. If counseling for domestic batterers, for example, fails to produce less repeat offending (Feder and Dugan 2002), the cause may be either that the theory was wrong (weak theory) or that offenders were not offered (or did not attend) enough counseling sessions (insufficient dosage). In another example, Landenberger and Lipsey (2005) noted that the effects of cognitive behavioral therapy (CBT) on recidivism are weaker when fewer CBT sessions are completed than when the full course of treatment is carried out.

Diverse samples? A second explanation for weak or no effects in experimental criminology was discovered by Weisburd et al. (1993) in the first systematic review of findings from randomized experiments in criminal sanctions. Weisburd and his colleagues coded the sample size, standard deviation, effect sizes and significance levels for their universe of eligible experiments. In what may be called the "Weisburd paradox," the likelihood of finding a significant difference between treatment and control groups in this sample of experiments was contrary to the conventional wisdom about sample size. The larger the sample size, the *less* likely the experiment was to report a statistically significant difference. Weisburd elegantly demonstrated that the larger samples had higher standard deviations than the smaller samples. That, in turn, reduced the statistical power of the test. There was less statistical power to detect "significant" results in larger samples because power depends on variance, not just on sample size or base rates of outcomes. Thus, if effect size was held constant, the chances of a significant result became lower as the standard deviation and sample size increased (see Cohen 1988).

The larger point here is not sample size but diversity: crime prevention interventions seem to have very different effects on different kinds of people. When

more differences that could affect outcomes are introduced into a sample, the differences drive up the standard deviation. These differences could include base rates, by which small differences in effect sizes across low and high base-rate offenders could yield large differences in the overall count of the offenses in each treatment group. Or they could include effects of opposite direction.

How do we know that such diversity is greater with larger sample sizes than with smaller samples? Weisburd not only demonstrated this quantitatively by correlating standard deviation with sample size. He also found qualitative evidence that when many criminological experiments failed to meet their target sample sizes within the research funding period, their leaders broadened the eligibility criteria to enroll more diverse kinds of cases.

The conventional expectation is that samples might become larger without becoming more diverse. Sherman et al. (2000), for example, took that risk knowingly when they broadened the offender age criterion for their restorative justice diversion experiment with violent crimes from a maximum age of 18 years to 29 years. After several years of not receiving enough referrals from police to reach a projected minimum sample size of 100, they decided that broadening the criteria seemed better than abandoning the experiment. That gamble succeeded in achieving a large and statistically significant effect, but not because the expanded sample was normally distributed. The treatment only worked because the older offenders were more responsive to the treatment than the younger ones were. That finding can be taken as evidence of the main point that larger samples may generally increase diversity, even when one subgroup swamps another so that large effects can be found overall.

Even without changing the criteria for eligibility over the course of an experiment, targets for minimum sample size may press experimenters towards designs with more diverse sample composition rather than less. In theory, eligibility criteria could be based upon a single type of offense (like bank robbery) or offender characteristics (like prior record, gender or age). In practice, however, experimental designs may be driven by the number of available cases in a jurisdiction within a given funding period. That, in turn, may force the experiment to include a far broader range of offenses, victims and offenders than a theory might suggest. Gladwell's empirical argument suggests that the quest for large samples is unlikely to solve the problem, if only because it dilutes the available dosage of a program across too many people who cause too little harm. This claim converges with the Weisburd paradox that larger samples produce smaller effects.

Both Gladwell and Weisburd find that the quest for large samples is based on a false premise. The premise is that the distribution of harm across units follows a bell curve. The data show otherwise. Weisburd's evidence of standard deviations increasing with sample size suggests that the distribution of harm is not normal but highly skewed. The more cases we draw, the greater the range of responses to the intervention becomes.

Low harm samples? A skewed distribution, by definition, means that in the majority of cases in the sample, each case causes little harm. Even if the treatment being tested works for those cases, it may be difficult to detect that effect if

(a) most of the harm is caused by a small fraction of the cases, and
(b) most of the result depends on how those few cases respond to treatment

The assumption that consistent application of a treatment across all cases is the best test of its effect may therefore be misguided. While it is true of random assignment in statistical theory, in practice the inferential force of random assignment may be wasted on skewed distributions. The accumulating evidence of differential responses within samples could explain the average finding of weak to no effect. When two or more subgroups react in completely opposite ways to the same consistent treatment, the resulting effect may be near zero. That does not mean, however, that the treatment cannot work for anyone. Instead, it throws the spotlight on the question of "what works for whom."

The salience of this question is growing. More findings of "no effect" in experiments have been found to mask the large, but contradictory, effects in subgroups that were identified prior to random assignment. Arrests for misdemeanor domestic assault, for example, have opposite effects on employed and unemployed suspects (Pate and Hamilton 1992; Sherman and Smith 1992; Berk et al. 1992). Pre-school participation from age 2 years reduced violence by age 12 among youngsters who had no medical conditions at birth, but it increased violence among those who did have medical conditions at birth (Pagani et al. 1998). White offenders in Canberra were largely unaffected by meeting with their victims of property crime and had much lower rates of recidivism than controls after restorative justice for violent offences, but Aboriginal offenders assigned to restorative justice increased their offending rates by over 200% compared to that of their controls (Sherman and Strang 2007).

In sum, the power curve problem is that it makes it difficult for one to interpret the central tendency of any difference between experimental and control groups. In a normal distribution the interpretation is more straightforward, with the mean values falling close to the median and the mode. In a power curve distribution, the measures of central tendency are much farther apart. The mean response can be driven by a few cases, while the median values of experimental and control groups may be much closer. The paradox can be what pharmaceutical executive Allen Roses described by saying that most drugs (that his company sold) "don't work for most patients" (Connor 2003) Only a minority of patients derive benefit from the not-so-miraculous drugs used to treat Alzheimer's disease, hepatitis or cancer, even though the average effect of each drug is beneficial. One manifestation of such differential response across subgroups can be, on average, weak overall effects in experimental results.

The three hypotheses of what causes weak effects are not independent. A weak theory that fails to address the heterogeneity of responses to an intervention may provide insufficient dosage for the cases that do the most harm. The larger the sample, the less dosage may be available for implementing the theory with integrity. Larger samples may also be more heterogeneous, with greater chances of missing the strong effects on some subgroup. Yet designers of experiments may fear that smaller samples would weaken statistical power or provide "not enough" cases for a "valid" test. The solution to all these potential causes of weak effects may be found, again, in the few randomized experiments that have succeeded in finding large benefits—all of them using a selective, power few sample to some degree.

The "power few" solution

Gladwell's hypothesis implies that the "power few" solution is to invest large *quantities* of resources in the most harmful tail of the distribution of cases, regardless of the *qualities* of that investment. What stops us from adopting this solution, he implies, is that we are wasting too much money on low-risk cases that contribute less to total harm than the cost of serving them justifies. If we concentrate our scarce resources on the big targets, the overall harm level of the problem will shrink more quickly than if we spread them too thinly. In this respect, Gladwell is confusing necessary and sufficient causes, as if big spending alone will bring about a big result.

What Gladwell skips over in this analysis is the role of experimental science, and the high failure rate of experiments. It is entirely possible to concentrate resources on high-harm cases and still produce no effect. A more cautious view of the "power few" is that it is just a better place to start experimenting, not a solution in itself. Concentrated resources may make eventual success more likely, but many failed experiments may be needed before a program is discovered that works. That much is true of any kind of experimental sample. It may be even more likely with the power few. Absent an excellent theory of how best to spend any extra money, power few samples may be even *more likely to fail* than less selective samples. The very reasons that success with the power few would yield big effects are the same reasons that all treatments with this group may be more resistant to change.

Power few targets are anything but the proverbial "low-hanging fruit" that is easiest to harvest. The power few may, in fact, be the hardest nuts to crack: the cases that are most difficult to solve because they have so many simultaneous or "co-morbid" problems. This is true not only of people, but also of places, neighborhoods, schools, prisons, police units, or any other population with harmful characteristics of crime or justice. Hot spots of crime may have co-morbid concentrations of people with risk factors such as alcoholism, mental illness, post-traumatic stress disorder or maternal rejection at birth. They may also (but not always) have multiple kinds of crimes (Weisburd et al. 1988). The same may be true for prisons suffering weak or brutal leadership or with concentrations of inmates who are skillful at generating conflict among their fellow inmates. Whatever the type of unit, the fraction of the population of such units with most of the harm will likely pose complex challenges of co-morbidity.

Some people counsel against working with such hard cases at all, let alone accepting my recommendation to make them the priority. The classic battlefield system of triage did not try to help those most likely to die even with extensive treatment; it gave medical priority the medium risk cases over high- and low-risk as the best way to save lives. Battlefield injuries, however, had much higher mortality rates than most crime and justice problems. The "power few" crime hot spots, or brutal prison guards, or violent schools do not die if they are left untreated. They go on and on, causing the majority of harm in their category. Leave them unchecked, and most of the harm in the population remains unchecked. Power few solutions must, therefore, provide a very different logic from that of classic triage. In the battlefield metaphor, the power few would not be the patients most likely to die but the enemy in pillboxes shooting the most soldiers—which are also the hardest to attack. Many soldiers have died in the cause of such high gain, high risk, attacks.

The power few strategy for crime and justice is therefore a paradox: *the best chance of a successful result may be found with those cases most likely to fail*. That paradox depends not on the responsiveness of the cases to treatment, which may be lower than for all other cases. Rather, the paradox requires that programs seeking big effects work on big problems: high base rates, high seriousness, and high proportions of the total harm in a much larger population.

The evidence in experimental criminology appears consistent with the basic Gladwell hypothesis. While a systematic review would be needed to confirm this conclusion, my own scanning of the findings of randomized experiments finds the strongest effects in the most homogeneous, high-risk samples. Four examples illustrate these effects. Three other cases, however, show that experiments with such samples can and do fail. I conclude that selecting a high-risk sample is a necessary, but not sufficient, condition for finding big effects in experimental criminology.

What worked: effect sizes and eligibility criteria

Four well-known examples illustrate big effects with power-few samples: pre-school education, nurse home visits, repeat offender surveillance and hot spots policing.

Perry pre-School project In one of the longest-lasting effects of a crime prevention program, the Ypsilanti, Michigan trial of the Perry pre-school project found substantial crime reduction benefits through age 40. The power few sample consisted of 123 African–Americans who were risk-assessed at age 2 as highly likely to fail in school. The fact that both experimental and control groups had high prevalence of arrests by age 40 is one indication that the sample selected for random assignment was homogeneously high-risk yet still responsive to this low-cost program early in life. People who were randomly selected for the program were one-third less likely to have been arrested five or more times by age 40 than those whose parents volunteered to put them in the program but who apparently stayed home during pre-school years (Schweinhart 2005). For males, this difference was even greater. Overall, the analysis estimated almost $13 in cost savings for every dollar invested in the program, 88% of which was due to lower costs of crime and justice. Of that total return, 93% of it was associated with the males—the higher risk subcategory—and only 7% was associated with females, suggesting even bigger effects among the power few (male) subset of the power few. These effects have received widespread attention worldwide, arguably because the effects are so large.

Nurse-family partnerships The original randomized trial of this program of almost 50 home visits to new mothers by Registered Nurses was conducted in Elmira, New York, USA, with a range of risk levels (Olds et al. 1998). The comparison of effect sizes across subgroups of higher and lower risk levels is consistent with the power few solution: the higher the risk, the bigger the effect. The original analysis found a 40% reduction in child abuse among the highest risk group. The smaller but positive impact among the lower risk group was not statistically significant. Olds concluded that, as a matter of policy, scarce resources would yield the greatest return if they were limited to the power few. Estimates of the return on investment from this program ranged up to $7 in cost savings through age 18 years for every dollar

invested before age 2. While more recent sensitivity analysis has shown that even low-risk families may have statistically significant effects, the magnitude of benefits remained greatest among the highest-risk families.

Repeat offender surveillance Two independent randomized experiments in police surveillance of high-risk, previously convicted, recently released, offenders both found very large effects (Martin and Sherman 1986; Abrahamse et al. 1991). In the Washington, DC case, 214 offenders, serious and reportedly active ex-prison inmates, assigned by a coin-flip to covert surveillance, were about 400% more likely to be arrested than if they were not assigned to that treatment. Their prior crimes included rape, robbery and attempted murder. While there is no direct measure of the number of crimes prevented by their arrest, these "power few" offenders were far more likely to go back to prison after assignment to the Repeat Offender Project (ROP) than if they were not assigned to the ROP.

Hot spots police patrols In a series of randomized experiments and quasi-experiments, a variety of police efforts to increase patrol visibility in the power few locations called crime "hot spots" has yielded consistently large benefits (Braga 2001). Sherman and Weisburd's (1995) randomized experiment in Minneapolis at 110 hot spots averaged 3 hours of observed uniformed patrol per night, compared with the observed control group average of 1 hour. In the first 7 months of the experiment when treatment integrity was maintained, the effect of extra patrol was substantial: experimental sites had two-thirds less of an increase in crime (at a time of city-wide increase) than was found in the control group locations. As the universe of the highest crime frequency (and spatially independent) locations in the city, the cases assigned to different conditions were complex, co-morbid environments with many diverse problems. They were homogeneous, however, in their high volume of crime in a small geographic area (usually centered on a street corner).

In all four of these examples, a sample size of no more than several hundred high-risk cases provided adequate power for finding large effect sizes from a highly intensive concentration of resources in support of a clear and evidence-based theory. The fact that these characteristics are not *sufficient* to guarantee large effects is demonstrated by the three failures that follow below. Yet these characteristics may still be necessary for big effects. What follows suggests why other reasons may have led the solutions to fail.

What didn't work and why

At least three examples of randomized experiments testing the "power few solution" show that it does not always work. Different reasons are suggested for each failure, illustrating the range of additional factors we must get right in order to reduce harm successfully.

Intensive probation In a series of multi-site randomized experiments, intensive supervision of high-risk probationers and parolees was tested in the 1980s. In general, these tests failed to find any clear reduction in new crimes by probationers

assigned to intensive supervision (MacKenzie 2006). Many factors could explain this result. One most often suggested is that the intense supervision provided *not enough contact*, and that more contact with probation or parole officers would be needed to undertake effective rehabilitation of very troubled people. Joan McCord may have suggested the opposite: that even weekly visits to a probation or parole office brought the offenders into *too much contact* with other high-risk people. This contact may have fostered a "deviant peer contagion" effect of increasing the likelihood of crime, counteracting any benefits from contact with the probation officer. A third alternative is that the nature of the relationship between offenders and their supervisors was too antagonistic for offenders to identify with the officers and try to achieve mutually agreed goals. All of these suggestions fall under the general heading of weak theory, by which there is no clear reason to believe that spending more time with probation officers should reduce criminal offending. This is confounded by weak measurement, in which it is always unclear whether increased supervision leads to more detection or more deterrence. Nor is it clear from the program descriptions whether the offenders saw their supervisors as providing support, the threat of punishment, or both, and in what balance.

Repeat call address policing The Minneapolis Police Department undertook the largest randomized experiment in problem-oriented policing (POP) in 1986, assigning to the treatment group 250 of the top 500 addresses in the city ranked by number of police cars dispatched to them (Sherman 1992). The failure of this experiment to show reductions in repeat calls has been blamed by POP supporters on inadequate dosage (Eck and Weisburd, personal communication). Yet, there is no independent evidence of how much dosage is "enough" for good results in problem-oriented policing, or any evidence that dosage levels matter more than the content of the strategy. It is not even clear that previously reported and successful case studies in POP (e.g., Eck and Spelman 1987) used any more dosage than repeat call address policing (RECAP): about 40 officer-hours per address in which to develop effective interventions in crime or disorder. What is clear, from the perspective of the experiment's designer (Sherman, personal communication), is that the program suffered from the same weak theory as the intensive probation experiments. Simply having police officers spend more time with the managers or residents of a wide range of properties is not much of a theory, unless there is a clear causal model of actions that police could take to reduce crime. The variance within the sample was also high, despite their common membership on the top 500 addresses (out of 115,000) for frequency of police calls; the locations varied by almost a factor of 10. These and other issues meant that a power few sample was not a sufficient cause for success.

High-risk peer counseling Two randomized experiments in peer counseling by "power few" people in trouble both failed to produce benefits. Both produced harms. One was a project of the Kansas City (MO, USA) Police Department to help aggressive young police officers to use less force in apprehending suspects (Pate et al. 1976). The other was a school-based program for at-risk teens to help each other by meeting to discuss their problems with their parents, teachers and life (Gottfredson 1987). Both programs allowed the randomly assigned subjects to meet in private with each other, without supervision by manifestly pro-social supervisors.

What was said in these meetings is not entirely clear, but it may have reinforced rather than counteracted the anti-social values and attitudes. The result was more use of force by treatment group police than by controls, and more delinquency by treatment group teens than controls. In both cases the theory was not weak, but strong—and wrong. The phenomenon of deviant peer contagion that Joan McCord described in the Cambridge Somerville experiment may have applied just as powerfully in these cases as well.

The tragedy is that no further experiments were conducted, testing other ways to deal with high-harm police officers or students. The basic lesson of all these failures is to expect failure and to plan for variations in approach to the same high-harm populations.

Seriousness vs frequency of harm: defining "high risk"

Whatever the theories of treatment may be, the key issue in defining a "power few" solution remains the definition of the "high risk" sample. This section closes our discussion of power few solutions with proposals for recasting our outcome measures around the idea of seriousness in measuring the total harm prevented. It then illustrates the proposals with the homicide prevention research at Penn's Jerry Lee Center of Criminology.

The concept of "high risk," when applied to neighborhoods, offenders, taverns or abusive police officers, is used most often to denote the probability, rates or *frequency* of future crime, rather than its *seriousness*. In colloquial usage among criminal justice practitioners, these dimensions are often confused or treated as synonymous. When applied to convicted offenders under community supervision, for example, the concept of "high risk" could even mean high probability of technical violations, rather than of committing any new offense at all. This language fails to focus on the critical distinction between high *risk* and high *harm*.

High harm vs high risk Consider an offender who is convicted five times for shoplifting, sentenced to probation, fails to appear for probation meetings, and is then sentenced to prison as a technical violator with no new crime. He may have a high likelihood of a future technical violation, or even theft, but not of committing mass murder. Technical violators may be a prime example of a "high-risk, low-harm" sample. Department stores with repeated shoplifting arrests may be another example, consuming major police resources but preventing little serious harm to the community. What may impress the public most in a democracy is not that an experiment reduces repeat offending but that it prevents serious injuries or death.

Common metric of harm Taking harm reduction seriously requires a common metric of harm across all crimes. Without such a metric of the varying costs of crime, analyses proceed as if all crimes are of equal severity. Decades of public attitude surveys show that they are not (Rossi et al. 1973; Wolfgang et al. 1985). It matters little just which metric is used. What does matter is that a homicide be counted as more harmful to the community than a shoplifting arrest, and that a rape be counted as more harmful than a car theft, and that the differences in magnitude between harm counted reflect the highly consistent differences in public opinions of crime severity.

Needles in the haystack The simplest way to apply the idea of severity to the power few is to limit forecasting analysis to a small number of extremely serious offenses: the extremely serious needles in the haystack of criminal events. This approach, clearly reflected in the "global war on terror," gets much closer to a common metric of harm than does defining the power few as "million dollar Murrays," as Gladwell did, or as the top 500 repeat call addresses, as Sherman (1992) did. The latter definitions of the power few stress frequency over seriousness. That implies two premises that may be wrong. One premise is that high volume and high risk of seriousness are strongly correlated; Lattimore et al. (2004) have recently falsified this claim with respect to youth offenders. The other premise is that simple cost to the taxpayer is a leading metric of harm. When compared with the public's opinion about murder or sex crimes against children, cost seems to matter quite little. In order to prevent the most harm, experiments in criminology can measure harm more meaningfully by stressing seriousness over prevalence or frequency. The most accurate way to do this, if hardest to explain, is by constructing an index of total harm.

A total harm index Despite the fact that the Uniform Crime Index of the Federal Bureau of Investigation (FBI) is merely a sum, not an index, there is a great need for a *crime index* in criminology. A true index would combine all crime events by assigning weights to each classification of crime. In an ideal world, the classifications would be based on extent of injury or ripple effects of injury. For example, a murder may be 50 times more serious than theft (as measured by a public opinion poll), but the total harm may vary based on the number of people who depended on the murder victim. For example, we could ask how many children under 18 years of age the murder victim had: none, three, 12? Research funding, however, would not usually cover the costs of such intense measurement or more sophisticated statistical interpretations of variance in public attitudes on severity. In our current world we could do much good with a crude scaling of each event by the public opinion severity score for its legal category (e.g., robbery), then sum the weighted values for all events to compute the index score for each unit (neighborhood, offender, street address) or total population for each time period. This could be done not only for each city, but also for the entire nation each year. It could also be done for the experimental and control group as the principal outcome measure in each experiment. Finally, it could be the best way to define the power few, in a forecasting model based upon the distribution of total harm likely to be caused in a population.

Low-risk and low-harm probationers Nowhere is this distinction more relevant today than in probation. Since the mid-1990s, when New York City replaced its one-size-fits-all model of supervising probationers with a caseload classified by risk, there has been growing interest in allocating probation cases by the power few solution. Up to 75% of New York City (NYC) probationers are classified as low risk and placed on low-intensity supervision. The caseloads of low-intensity supervisors run up to 600 offenders per officer (Jacobson 2006). This is made possible by palm-print identification of the offenders who report periodically to a computerized kiosk in a probation office lobby, answering questions by computer that an officer used to ask in person. This low intensity frees up most officers to spend much more time

with high risk cases. What it may not do, however, is to allow probation officers to focus on cases with a high risk of high harm. That is because the risk analysis treats all recidivism as equal, rather than focusing on the high-harm needles in the haystack.

Experiments in homicide prevention

The Jerry Lee Center of Criminology at the University of Pennsylvania has recently developed a "high-harm" approach at the request of Philadelphia's Adult Probation and Parole Department (APPD). With over 50,000 active cases at any given time, a 30-year data set showing all clients who are later charged with murder or attempted murder yields an ample haystack in which to find the power few "needles." Using a nonlinear data mining approach to exploratory analysis, Berk et al. (2007) have developed a relatively straightforward model that forecasts murder charges up to 52-times more accurately than does a random sample of the population.

Richard Berk's model includes variables that would not normally be considered by judges or probation officers, largely because they are not based on previous convictions. However, since the most likely killers have long records of prior prosecutions not leading to convictions, these data comprise valuable tools for improving forecasting skills. The most useful variables include age at first prosecution as an adult (legally possible at age 10 years for murder, older ages for other serious offenses). They also include current age, which is the most powerful predictor, even though it has no legal relevance once the offender is 18 years old. With just under 1 in 1,000 APPD clients charged with murder each year, and some 1 in 200 charged with attempted murder, the total harm caused by all clients is great. However, the average harm per client is far greater for those predicted to kill someone.

Rather than putting all adult probationers under one-size-fits-all supervision, the APPD has begun to use Berk's model to place high harm cases under its most intensive supervision. At the same time, it has moved to place thousands of others under less intensive supervision, planning a randomized trial of a "low risk of high harm" model based on the extreme harm approach, rather than on the probability or volume of any offending. If this experiment shows no increase in extreme harm among what we might call the "powerless many," it could pave the way to the first classification based on risk of high harm rather than on risk of any harm.

The Jerry Lee Center's partnership with APPD is also providing the few likely murderers with far more services than could be provided to the full caseload. These services include cognitive behavioral therapy for depression, post-traumatic stress disorder, and offending behavior per se, for which the evidence on repeat offending, in general, is encouraging (Landenberger and Lipsey 2005). The services also include more frequent drug testing, home visits in evening hours after curfew, and a far lower caseload (five to 15 cases per officer) than in any of the intensive supervision experiments. Because the likely killers are also very likely to be killed—with two such clients having been murdered in the first 100 identified by the model—the question of whether close monitoring is punishment or a life-saving service is open to discussion.

A similar approach could be taken to the regulation of taverns where people are repeatedly killed, of the schools where the most extreme violence happens to

teachers or students, and of the businesses where armed robberies repeatedly occur. What treatment will work cannot be known in advance. However, the commonly employed one-size-fits-all approach to taverns, schools, or convenience stores seems unlikely to address the unique features of the high-harm power few. The job of experimental criminology is to cut through the normal curve paradigm to develop and test prevention programs that are as customized to unique causes as are genetically customized pharmaceuticals (Connor 2003). Whether we can do that depends on our answers to the final questions I pose about the power few solutions: whether they are both ethical and politically viable.

Ethics and politics

Gladwell's choice echoes similar ethical dilemmas in criminology. He saw no way to solve the homeless problem without a power few solution that was unfair to all deserving people in need. Criminologists see little chance to protect the community from the power few without unfairly punishing them for crimes they have not yet committed. If the ethics of preventing crime include provision of services rather than punishment, however, many people would accept a power few solution as ethical. If the units of analysis are not offenders but places or collectivities (schools, prisons), a power few solution may appear even more ethical. If the ethics includes fairness to crime victims, the solution could seem ethically compelling. What is ethical could, in turn, be made politically acceptable, one experiment at a time. If experiments can then succeed in finding big effects, then both ethics and politics may be more easily clarified.

Support vs punishment

It matters immensely whether the policies we test provide support or punishment for convicted offenders. Based solely on a statistical forecast placing them in the power few causing future harm, some convicted offenders would be assigned to high-harm groups in error. Increasing punishment for a known percentage of people who are "innocent" of future harm strikes many as grossly unethical. However, providing extra health, rehabilitation and psychological services to people at risk, even if they never commit serious harm, sounds more like a good investment. Allocating resources by place-based or collective units seems even less likely to produce ethical challenges, since they may be cast as benefits to the majority of law-abiding people associated with those places or groups.

Sentenced criminals in most US states are provided very little opportunities for classic rehabilitation (Ruth and Reitz 2003). As a scarce resource, rehabilitation can be allocated on the basis of equal probability or seriousness of risk. It is not obvious which principle Americans (or Britons, or Israelis or Japanese) would prefer. However, it is not obvious that a risk-based allocation of services would be unethical. Nor is it obvious that such a principle would be as offensive in the context of criminal rehabilitation as Gladwell argues it is for the homeless. Perhaps the ethics may depend on the content of the treatment rather than on the principle of its allocation.

In the ethics of regulation, Braithwaite (2002) has developed a model of sanctioning called a regulatory pyramid. For both ethical and practical reasons, he proposes low levels of severity of sanctions for most violations, which are generally first violations. As repeated or more serious violations are detected, more severe penalties are to be imposed. At the peak of the pyramid is the most extreme penalties reserved for the most serious infractions or record of infractions.

In the ethics of harm reduction by support, an inverted pyramid makes more sense. This pyramid would allocate resources of preventive support or surveillance, rather than penalties of differential severity. A "pyramid of support" would allocate scarce resources in a harm-generating population to create the most equitable benefits for reducing human suffering. The few cases at the peak of the pyramid that will cause the most harm would receive the most resources. Just as in the homicide experiments, resources would decline in proportion to declining seriousness of likely harm.

The pyramid of support may help practitioners to visualize Gladwell's empirical claim: the more resources, the better the *potential* for a big harm-reducing result (depending on whether the right treatment is chosen for the specific population). More precisely, Figure 1 is a moral principle of resource allocation: the more harm a unit may cause, the greater the share that unit should receive of the total resources available to deal with that kind of harm. Experimental criminology can go well beyond that moral starting point, testing many theories of how best to use those extra resources. By starting with a predictive analysis that leads us to the most harmful cases (e.g., Berk et al. 2007), we can clearly justify the inverted pyramid of resource allocation. We may even be able to discover how to use those extra resources most effectively.

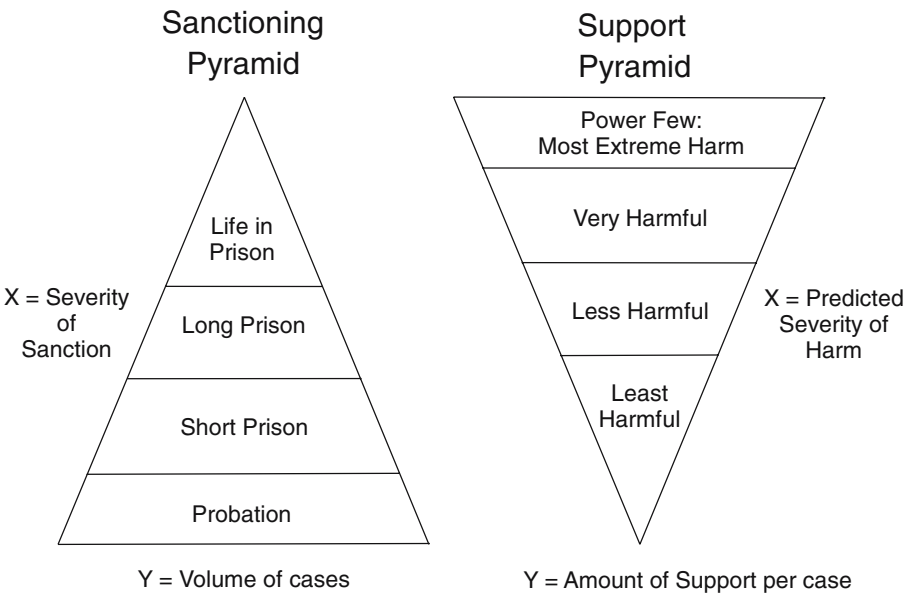

Fig. 1 How a pyramid of support compares to a sanctioning pyramid

What remains of Gladwell's dilemma is that some low-harm offenders may deserve a chance for rehabilitation on some other grounds. Perhaps they are more sincerely repentant or have done more to try to repair the harm they have done to their victims. The pyramid of support does not preclude these additional preferences. However, it might balance them against other considerations, such as whether any effective ways of preventing harm by the most harmful can be discovered. In this sense it is important to distinguish between principles of *experimentation* and principles of *ongoing policy*. The moral basis for *discovering* whether harm can be minimized is very different from the moral basis of *hoping* that harm may be minimized. From an experimental standpoint, there is no reason to argue that resource allocations should follow potential harm if the resource allocations do not work.

What remains of the concerns about punishment is a major question to address in the process of developing experiments with individual offenders. What status should be given to offenders in deciding whether what we call support is what they think of as a punishment? If we say that cognitive behavioral therapy is a supportive way to help them think more constructively, but they say that coming in to the probation office more frequently for such therapy is a punishment, who is right? In other contexts, courts have denied that even pre-trial detention is punishment (Rehnquist 1979), let alone rehabilitation. However, that legal distinction may not satisfy the ethical concern.

What may be more relevant here is the ethical argument of the Federal Judicial Center (1981) that *experiments* in *how* to administer sanctions do not need offender consent if they have already been found guilty of a crime, so long as the intent and content of the treatments tested are not known to be adverse in their effects on the offender or the community. So long as there is equal reason to believe that it might (or might not) be of benefit, the moral state of *equipoise* justifies the research. Then we would know whether offenders who are assigned to such counseling, on average, live 10 years' longer and are shot 40% less often; we might have a strong moral argument that the treatment is as much for their own good as for the community's, and that their conviction of a criminal offense means they have forfeited their right to decide whether to accept such beneficial support.

Offenders vs other units

Moral dilemmas about fairness of resource allocation are often transformed when the unit of analysis is shifted from individuals to collectivities. Readers who disagree with the foregoing discussions of experiments with offenders might still accept the idea of power few solutions to crime and justice across collectivities. Inequalities across collectivities are so great that many people choose to set aside such questions as truly ethical. From differential funding per student across public school districts, to differential homicide rates across neighborhoods, differential allocations of resources based on need (or on wealth) may strike people as fair in general. They may only say differentials are unfair if differentials take resources away from them personally.

A major reallocation of police patrols from low-crime areas to high-crime hot spots, for example, was unanimously approved by the Minneapolis City Council in 1987 (Sherman and Weisburd 1995). The grounds for the decision were that the entire community would benefit, including the people who lived on low-crime blocks but were more likely to be robbed in commercial areas in high-crime hot spots.

Now apply this principle to Gladwell's choice. If Gladwell had framed the question as to whether all of the homeless people concentrated in downtown Denver would receive the same treatment, but not the homeless people scattered in the suburbs, he may not have had such a powerful example of something that "feels" unethical. Recast in collectivity terms, the moral sting of differential resource allocation may disappear. A comprehensive analysis of the moral principles of group vs individual resource allocation is beyond the scope of this paper. However, if the pyramid of support is ethical at the individual level, it should be even more ethical at the collective level.

Fairness to crime victims

The ultimate ethical justification for power few solutions may be their consequences for individual victims, rather than collectivity rates of crimes. Even the most principled analysis of offender rights cannot ignore the rights of victims and potential victims. Gladwell's analysis completely fails, for example, to consider the harm done by the power few homeless to police officers with whom they fight, to pedestrians they shout at or threaten while panhandling (begging), to retailers who suffer their vomiting on the doorstep of a business. Other homeless people may deserve Cadillac treatment, but victims of the power homeless may have a more compelling case. Why should anyone be exposed to a problem that a government *can* solve but does not solve because it is unfair to someone who does not benefit by the solution? Why, one could ask, is it not unfair to the victims of the power few to ignore them in order to deal with lower-harm units in the population?

The true dilemma Gladwell points us to is the question not of *whether* to be unfair, but of *to whom* to be unfair. Even if it is unfair to other convicted criminals if we concentrate support for the power few, it may well be fair to the victims of terrible crimes that could be committed by the power few. Equal distribution of rehabilitation services across all offenders, or of police services across all hot spots, or of training of prison guards across all prisons, may result in unequal distribution of harm to the victims of the power few. If there is no way to find a fair solution to this problem, the best solution may be to add the utilitarian principle of least total harm for the greatest number.

Conclusion

The major problem for experimental criminology is to find more and bigger benefits of the strategies we test for preventing harm. One direction we can go in search of that goal is to focus our efforts on the "power few" units of places, offenders, communities, schools, prisons, police agencies and victims where the greatest total harm is found. Simply reallocating resources to the power few does not guarantee success, as several experiments have already shown. However, big effects in experimental criminology do appear more likely when such re-allocations are attempted. Reallocation may be a necessary, if not sufficient, condition for discovering what works.

All this is hypothetical grist for a future systematic review or for further experiments using power few samples. Rather than focusing on a given treatment, the review could focus on experiments themselves, asking whether they are more likely to have large effect sizes when samples are selected for high-risk or high-harm characteristics. On the premise that there may not yet be enough experiments in criminology with strong programs, statistically powerful samples, and strong integrity of implementation, it may be equally valuable to do more such tests. Either way, experimental criminology will be faced with a public dialogue about the key moral and political issues at stake.

Malcolm Gladwell's choice between solving problems and being true to our principles leaves out the major role of experimental social science. It is only when we can consistently demonstrate an ability to solve problems that the choice even becomes an issue. Simply conducting the experiments places us on a different moral plane from a permanent policy. Decisions about such policies would arguably depend on how much of a benefit can be demonstrated, and for whom, with what tradeoffs among victims, offenders, collectivities and communities. Until and unless the crucial experiments can be done, a deliberative democracy may have too little information about which to deliberate. So long as the experiments themselves can be designed to test hypotheses that would spread benefits widely, experimental criminology will be on firm ethical and political ground. And if there is any doubt of our willingness to say that any particular application of the strategy doesn't work, we can always cite the "power few" research of Joan McCord.

References

Abrahamse, A., Ebener, P., Greenwood, P., Fitzgerald, N., & Kosin, T. (1991). An experimental evaluation of the Phoenix repeat offender program. *Justice Quarterly, 8*(2), 141–168.

Barry, J. (2004). *The great influenza: The epic story of the deadliest plague in history.* NY: Viking.

Berk, R., Campbell, A., Klap, R., & Western, B. (1992). The deterrent effects of arrest in incidents of domestic violence: A Bayesian analysis of four field experiments. *American Sociological Review, 57*, 698–708.

Berk, R., Sherman, L., Barnes, G., Kurtz, E., & Ahlman, L. (2007). Forecasting murder within a population of probationers and parolees: A high stakes application of statistical learning. Working Paper, Jerry Lee Center of Criminology, University of Pennsylvania, http://www.crim.upenn.edu/murder.pdf.

Braga, A. (2001). The effects of hot spots policing on crime. *Annals of the American Academy of Political and Social Science, 578*, 104–125.

Braithwaite, J. (2002). *Restorative justice and responsive regulation.* NY: Oxford University Press.

Chalmers, I. (2003). Trying to do more good than harm in policy and in practice: The role of rigorous, transparent, up-to-date evaluations. *Annals of the American Academy of Political and Social Science, 589*, 22–40.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum & Associates.

Connor, S. (2003). Glaxo chief: Our drugs do not work on most patients. *The Independent* (UK), 8 December 2003, p. 1.

Doll, R. (1998). Controlled trials: The 1948 watershed. *BMJ, 317*(7167), 1217–1220 (31 October 1998).

Eck, J., & Spelman, W. (1987). *Problem solving: Problem-oriented policing in Newport News.* Washington, DC: Police Executive Research Forum.

Eck, J., Clarke, R., & Guerette, R. (2007). Risky facilities: Crime concentrations in homogeneous sets of establishments and facilities. *Crime Prevention Studies, 21*, 255–264.

Farrell, G. (1995). Preventing repeat victimization. In M. Tonry & D. Farrington (Eds.), *Building a safer society. Crime and justice 19* (pp 469–534). Chicago: University of Chicago Press.
Farrington, D. (1983). Randomized experiments on crime and justice. In N. Morris & M. Tonry (Eds.), *Crime and justice 4* (pp. 257–308). Chicago: University of Chicago Press.
Farrington, D. (2006). Key longitudinal-experimental studies in criminology. *Journal of Experimental Criminology, 2*, 121–141.
Farrington, D., & Welsh, B. (2005). Randomized experiments in criminology: What have we learned in the last two decades? *Journal of Experimental Criminology, 1*, 9–38.
Feder, L., & Dugan, L. (2002). A test of the efficacy of court-mandated counseling for domestic violence offenders: The Broward experiment. *Justice Quarterly, 19*(2), 343–375.
Federal Judicial Center. (1981). *Experimentation in the law.* Washington, DC: Federal Judicial Center.
Forst, B., Lucianovic, J., & Cox, S. (1978). *What happens after arrest? A court perspective of police operations in the District of Columbia.* Washington, DC: INSLAW.
Gladwell, M. (2006). Million dollar Murray: Why problems like homelessness may be easier to solve than to manage. *The New Yorker*, 13 February 2006.
Gottfredson, G. (1987). Peer group interventions to reduce the risk of delinquent behavior: A selective review and a new evaluation. *Criminology, 25*, 1001–1043.
Hill, A. B. (1990). Memories of the British streptomycin trial in tuberculosis. *Controlled Clinical Trials, 11*, 77–79.
Independent Commission on the Los Angeles Police Department. (1991). *Report.* Los Angeles, CA: City of Los Angeles.
Jacobson, M. (2006). Lecture, Jerry Lee Center of Criminology, University of Pennsylvania. (March).
Johnson, S. (2006). *The ghost map: The story of London's most terrifying epidemic—and how it changed science, cities, and the modern world.* N.Y.: Riverhead Books (Penguin Press).
Knightly, P., Potter, E., & Wallace, M. (1979). *Suffer the children*: *The story of thalidomide*. New York, NY: Viking Press.
Kock, R. (1999). *The 80/20 principle: The secret to success by achieving more with less.* New York: Doubleday.
Landenberger, N., & Lipsey, M. (2005). The positive effects of cognitive-behavioral programs for offenders: A meta-analysis of factors associated with effective treatment. *Journal of Experimental Criminology, 1*, 454–476.
Lattimore, P., MacDonald, J., & Piquero, A. (2004). Studying the characteristics of arrest frequency among paroled youthful offenders. *Journal of Research in Crime and Delinquency, 41*, 37–57.
MacKenzie, D. (2006). *What works in corrections.* Cambridge: Cambridge University Press.
MacKenzie, D., & Hickman, L. (1998). *What works in corrections (Report submitted to the State of Washington Legislature Joint Audit and Review Committee).* College Park, MD: Department of Criminal Justice and Criminology, University of Maryland.
Martin, S., & Sherman, L. (1986). Selective apprehension: A police strategy for repeat offenders. *Criminology, 24*(1), 155–174.
McCord, J. (1978). A thirty-year followup of treatment effects. *American Psychologist, 33*(3), 284–289.
McCord, J. (2003). Cures that harm: Unanticipated outcomes of crime prevention programs. *Annals of the American Academy of Political and Social Science, 587*, 16–30.
Olds, D., Henderson, C., Cole, R., Eckenrode, J., Kitzman, H., Luckey, D., et al. (1998). Long-term effects of nurse home visitation on children's criminal and antisocial behavior: 15-year follow-up of a randomized controlled trial. *Journal of the American Medical Association, 280*, 1238–1244.
Pagani, L., Tremblay, R., Vitaro, F., & Parent, S. (1998). Does preschool help prevent delinquency in boys with a history of perinatal complications? *Criminology, 36*(2), 245–268.
Pate, A., & Hamilton, E. (1992). Formal and informal deterrents to domestic violence: The Dade County spouse assault experiment. *American Sociological Review, 57*, 691–697.
Pate, T., McCullough, J., Bowers, R., & Ferrara, A. (1976). *Kansas City peer review panel: An evaluation report*. Washington, DC: Police Foundation.
Rehnquist, W. (1979). *Bell v. Wolfish. US Supreme Court* (441U.S. 520).
Rossi, P., Waite, E., Bose, C., & Berk, R. (1973). The seriousness of crimes: Normative structure and individual differences. *American Sociological Review, 39*, 224–237.
Ruth, H., & Reitz, K. (2003). *The challenge of crime: Rethinking our response.* Cambridge, MA: Harvard University Press.
Schweinhart, L. (2005) *The Perry preschool study through age 40: Summary, conclusions and frequently asked questions.* Ypsilanti: High/Scope Educational Research Foundation.

Sherman, L. (1992). Attacking crime: Police and crime control. In M. Tonry & N. Morris (Eds.), *Modern policing: Crime and justice 15* (pp 159–230).
Sherman, L. (2000). Reducing incarceration rates: The promise of experimental criminology. *Crime and Delinquency, 46*(3), 299–314.
Sherman, L. (2006). To develop and test: The inventive difference between evaluation and experimentation. *Journal of Experimental Criminology, 2*(3), 393–406.
Sherman, L., & Smith, D. (1992). Crime, punishment and stake in conformity: Legal and informal control of domestic violence. *American Sociological Review, 57*, 680–690.
Sherman, L., & Weisburd, D. (1995). General deterrent effects of police patrol in crime "hot spots": A randomized, controlled trial. *Justice Quarterly, 12*, 625–648.
Sherman, L., & Strang, H. (2007). *Restorative justice: The evidence.* London: Smith Institute.
Sherman, L., Gartin, P. R., & Buerger, M. E. (1989). Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology, 27*, 27–55.
Sherman, L., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., & Bushway, S. (1997). *Preventing crime: What works, what doesn't, what's promising.* Washington, D.C.: US Department of Justice, Office of Justice Programs.
Sherman, L., Strang, H., Barnes, G., & Woods, D. (2000). *Recidivism patterns in the Canberra Reintegrative Shaming Experiments (RISE).* Canberra: Australian Institute of Criminology.
Simon, H. (1955). On a class of skew distribution functions. *Biometrika, 42*, 425–440.
Smith, J. (1990). *Patenting the sun: Polio and the salk vaccine.* N.Y.: William Morrow and Co.
Tröhler, U. (2003). James Lind and scurvy: 1747 to 1795. The James Lind Library (www.jameslindlibrary.org).
Uglow, J. (2003). *The lunar men.* NY: Farar, Straus & Giroux.
Wagner, P. (2005). Eric Cadora shows how incarceration is concentrated in particular Brooklyn neighborhoods. *Prisoners of the Census*, 24 January 2005. Accessed 8 August 2007 at http://www.prisonersofthecensus.org/news/2005/01/24/cadora/.
Weisburd, D., Maher, L., & Sherman, L. (1988). Contrasting crime-general and crime-specific theory of crime: The case of hot spots of crime. In W. Laufer & F. Adler (Eds.), *Advances in criminological theory 4* (pp 45–70). New Brunswick, NJ: Trans-Action Books.
Weisburd, D., Petrosino, A., & Mason, G. (1993). Design sensitivity in criminal justice experiments. In M. Tonry (Ed.), *Crime and justice: A review of research 17* (pp 337–379). Chicago: University of Chicago Press.
Weisburd, D., Lum, C., & Yang, S. (2001). When can we conclude that treatments or programs 'don't work? *Annals of the American Academy of Political and Social Science, 587*, 31–48.
Weisburd, D. Bushway, S., Lum, C., & Yang, S. (2004). Trajectories of crime at places: A longitudinal study of street segments in the City of Seattle. *Criminology, 42*, 283–322.
Wolfgang, M., Figlio, R., & Sellin, T. (1972). *Delinquency in a birth cohort.* Chicago: University of Chicago Press.
Wolfgang, M., Figlio, R., Tracy, P., & Singer, S. (1985). *The national survey of crime severity.* Washington, D.C.: U.S. Government Printing Office.
Zipf, G. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.* Cambridge, MA: Addison-Wesley Press.

Lawrence W. Sherman is the Director of the Jerry Lee Center of Criminology at the University of Pennsylvania and Wolfson Professor of Criminology at Cambridge University. His research activities include designing and replicating randomized experiments in crime prevention, primarily by varying criminal justice responses, including his randomized trials on restorative justice in collaboration with the Australian National University's Regulatory Institutions Network.