J Quant Criminol (2010) 26:489–500
DOI 10.1007/s10940-010-9117-3

ORIGINAL PAPER

Gold Standard Myths: Observations on the Experimental Turn in Quantitative Criminology

Robert J. Sampson

Published online: 18 September 2010

Criminology has experienced a distinct "experimental turn." The sheer number of quantitative experiments in criminology has increased dramatically in recent decades, accompanied by an influential and institutionalized movement to promote experiments.[1] Evidence of this turn is seen in our journals, funding trends, scholarly awards, proclamations by governmental agencies like the Office of Justice Programs, the emergence of new scholarly divisions in the American Society of Criminology, and separate societies such as the Academy of Experimental Criminology.

The claims for experimental methods have not been modest. Among other things, experiments have been argued to enhance scientific quality, evidence-based policy, causal inference, and even liberty.[2] Criminology is by no means alone in its experimental zeal. Angus Deaton, the former President of the American Economic Association, recently coined the term *randomistas* to describe those economists convinced that adopting the paradigm of the randomized clinical trial (RCT) was the best hope for scientific progress.[3] Claims to superiority are in some sense unsurprising. After all, the common reference to experiments as the "gold standard" quite explicitly invokes a sense of hierarchical or superior knowledge (Cartwright 2007). Observational methods have apparently been shown their lower station in life.

[1] I focus on experiments as classically understood, in particular the model of the randomized clinical trial in medicine that has motivated the experimental movement in criminology.

[2] A number of programmatic statements promoting experimental criminology are available (e.g., Sherman 2009; Weisburd 2010; Weisburd, Mazerolle and Petrosino 2010). Critiques of Sherman's argument that experiments advance liberty recently appeared in *Criminology and Criminal Justice*; see Carr (2010), Hope (2009), Hough (2010) and Tilley (2009).

[3] Deaton (2008). For further debate see the special issue of the *Journal of Economic Perspectives* where Angrist, of instrumental variable fame, refers to the "credibility revolution" in empirical economics (Angrist and Pischke 2010). A number of critics respond. Educational research has also seen a strong experimental push. For an evaluation, see Raudenbush (2008).

R. J. Sampson (✉)
Harvard University, Cambridge, MA, USA
e-mail: rsampson@wjh.harvard.edu

On the occasion of the 25th anniversary of the *Journal of Quantitative Criminology*, it seems appropriate to reflect on the experimental turn and its implications for criminology at large and criminological methods in particular. I do so by unifying a number of otherwise disparate themes that bear on the claims of experimentalists. My assessment is mixed. Quantitative criminology has made great strides and the experimental subfield has done much to improve our knowledge base. I agree with Weisburd's (2010) recent argument, for example, that folklores about experiments being "unethical" or "impractical" are (in most cases) just that. Experiments are an essential part of the methodological tool kit of criminologists, and I would hope to see more, not fewer, field experiments—especially at the group level (more on this below). With creativity and a good question, experiments can be quite exciting. I also think the counterfactual or "potential outcomes" paradigm in the social sciences has been positive, on balance.[4]

But criminological *randomistas* have overreached in their claims and generated their own folklores, or what I think are more appropriately referred to as myths. Experimental myths are more than just stories or part of a tradition—they have become actively institutionalized in the routine workings of criminology. As one example, the popular concept of "evidence based" policy is now widely taken to mean *experimental*-based. The stakes are thus high for criminal justice policy. I argue in this paper that experimentalists have fallen prey to three interrelated myths: (1) randomization solves the causal inference problem, (2) experiments are assumption (or theory) free, and (3) experiments are more policy relevant than observational research. I briefly analyze each of these myths, laying out my case in broad strokes due to space constraints. I conclude, as have others, that experiments are not the gold standard simply because there is no free-standing gold standard. Observational and experimental science should instead be partners in crime. In this spirit, I end by arguing the need to recast policy evaluation and proposing how future research might advance beyond myths and abstract claims to employ pragmatic strategies for addressing the deep challenges that face all methodological approaches.[5]

Myth 1: Randomization Solves the Causal Inference Problem

Claims for RCT superiority are not surprising given that experiments have long been cloaked in the mantle of science, especially the laboratory paradigm of randomization (or investigator control over allocation to treatment). Randomization is said to eliminate "confounding" (omitted variable bias) and therefore is the trump card that the experimentalist plays in the causal inference game (Weisburd 2010). Setting aside issues of practicality (many things of interest to criminologists are not amenable to randomization, including most things "macro") and ethics, randomization is not the scientific panacea it is often made out to be.

[4] Often called "Rubin causality" after the pioneering work of the statistician Donald Rubin, the counterfactual model has become the dominant conceptual framework for casual inference (Holland 1986). For an excellent book-length treatment, see Morgan and Winship (2007).

[5] In his recent ASC Presidential Address, Clear (2010) provides a strong critique of the experimental paradigm in criminology from a different angle. In this essay, appropriate for the *JQC*, I take experiments on their own terms and address key methodological limitations and resulting implications for both causal knowledge and policy formation. I take no stance on Clear's normative position on how the ASC should engage or promote the content of policy.

For one thing, the classical design of an experiment assumes random samples, the bread and butter of observational research design. Criminologists in the experimental tradition tend to gloss this over and focus on the benefits of randomization rather than the selection mechanisms that produce the samples in the first place. But the only way an RCT is strictly valid with respect to some specific population is if you have a probability sample drawn from that population or if not, that you know the selection mechanism (Imai et al. 2008). Consider that the vast majority of RCTs in medical research, from which experimental criminology draws its inspiration, are based on those who are selectively located in a medical setting (e.g., hospitals). In criminology, think prisons, courts, or perhaps a high crime neighborhood. I am hard pressed to think of many experiments in criminology either with multiple stage randomization—where the subjects (say, probationers) are randomly drawn to receive (randomly) a treatment—or where the selection mechanisms into the convenience sample are fully known. It is somewhat ironic that randomization is touted as the answer to selection bias when a different form of potential bias (selective samples) is typically present in criminological experiments.

Non random or selective samples might not be so bad if treatment effects are homogeneous across people or settings but we know this not to be case (Heckman 2001; Manski 2009). The definition of a treatment effect is thus contingent on the distribution of traits within a population. As Smith (2009, p. 237) astutely notes, this places traditionally non-experimental social science, such as demography, in a strong position to address experimental claims about causality and policy intervention. The same is potentially true of criminology, but I think it is fair to say that treatment effect heterogeneity has been overlooked as a threat to experimental claims.

Another challenge is that compliance with experimental treatments among selected participants is often surprisingly low, requiring explicit yet usually unexamined assumptions in order to make causal inferences. I address this more in the next section with respect to "SUTVA," but I would emphasize here, as have others, that humans exert choice and can accept or reject most treatments, anticipate future outcomes and act on them, and often discuss treatments with fellow subjects. The threat is not a trivial one that can be dismissed by reference to randomization, especially when combined with selection into the sample, typically small Ns, and substantial noncompliance (Heckman 2001; Moffitt 2005; Smith 2009). In the case of noncompliance combined with missing data, which is virtually ubiquitous, the ability to make model free assumptions—one of the proposed virtues of experiments—disappears. With noncompliance it is also the case that randomization does not guarantee exogeneity in the "treatment on treated" (TOT) case. This is because the treatment being randomized (say a housing voucher) is correlated with other changes (e.g., moving) that are relegated to the error term (Keane 2010). Randomization may get you exogeneity in the "intent to treat" case but not without further assumptions that must be imposed.[6]

There is yet a deeper problem. Causal inference is ultimately tied up with causal explanation, which resides at a theoretical level and is not something that comes directly from the data. Data never "speak for themselves"—making sense of causal patterns requires theoretical claims about unobserved mechanisms and social processes no matter

[6] I thank Gary King for discussion of these points. For a practical guide to addressing common misunderstandings in the analysis of experiments, see Imai et al. (2008)). An influential counterfactual approach to noncompliance is to estimate the "local average treatment effect" (LATE), or the treatment effect on compliers, where treatment assignment is used as an instrumental variable for the treatment actually taken (Angrist et al. 1996).

what the experiment *or* statistical method employed (Wikström and Sampson 2006). Causal explanation requires theory, in other words, not a particular method (Heckman 2005). In general, then, no one estimation technique is privileged by "science" in ascertaining causal explanatory knowledge. While common understanding links causality to the experimental (randomization) method, science is catholic on method, and fundamentally is about principles and procedures for the systematic pursuit of knowledge involving the formulation of a problem, the collection of data through observation or experiment, the possibility of replication (science is "public"), and the formulation and testing of hypotheses. This is virtually a dictionary definition of science.

It follows that the lab sciences are no more scientific than say, astronomy or evolutionary biology. Stanley Lieberson, the author of *Making It Count: The Improvement of Social Research and Theory* (1985), has argued that the social sciences have tried to mimic a classical physics-like focus on determinism where instead we should think more like evolutionary biologists. Darwin's theory was constructed not in a lab or in an experiment but by "drawing rigorous conclusions based on observational data rather than true experiments" and "an ability to absorb enormous amounts of diverse data into a relatively simple system" (Lieberson and Lynn 2002, p. 1). Analogous to evolutionary theorists, criminologists are often concerned with causal processes that take on historical and institutional dimensions that range over long periods of time (sometimes decades) and that are not amenable to randomization. Criminological social inquiry thus requires a more flexible conception of causality than that offered by the RCT paradigm.

Myth 2: Experiments are Assumption (Theory) Free

One often hears that experiments are preferable to observational studies because they require fewer assumptions. But as I have already alluded to, experiments just require *different* assumptions. I would argue that criminologists, more so than disciplines such as economics, seem to make assumptions implicit rather than explicit, and in the world of crime and social behavior, some of the crucial assumptions are directly contrary to available evidence.

Enter "SUTVA." Possibly the most inelegant phrase to have entered the social science lexicon, the "stable unit treatment value assumption" (SUTVA) underlies all experiments. Although not necessarily of great consequence for research on fertilizers or possibly taking a pill, SUTVA is foundational to criminology because at bottom it refers to assumptions about social interactions and what is known in the statistical literature as "interference" (Rosenbaum 2007; Sobel 2006). The essential idea of SUTVA is that potential outcomes for any unit of an experiment are independent of the treatment assignment of any other unit or population member under study. Human response to treatments like "Hawthorne" or "John Henry" effects (when participants in the control group alter their behavior purposely because of the experiment) are thus ruled out, as, more problematically, are any forms of social interaction. I quote here a summary in words of a very dense, technical literature:

> "If, through processes of information diffusion, norm formation, leadership, endogenous reinforcement, or competition in tournaments, social interactions are important for outcomes, units effects are inherently undefined because, in this case, outcomes for any particular unit i depend on the number or distribution of treated units j in the population, and the standard interpretation of any parameter estimate no longer applies" (Gangl 2010, p. 40).

The implications of SUTVA are profound. Processes of contagion, information diffusion, jealousy, and learning are staples of criminology and the social sciences in general. Raudenbush (2008), for example, notes how the very nature of educational processes leads to violations of SUTVA, calling into question the experimental turn in that discipline as well. Yet with few exceptions the experimental literature in criminology, and apparently education, has not tackled interference in a straightforward way. Consider just a few examples of the pervasiveness of the problem for criminology. Berk (2005) notes that placing a substantial number of rival gang members in the same boot camp could dramatically alter the nature of the treatment and the subsequent response (p. 421). In another, "suppose a drug treatment program was randomly provided to some youth facilities and not others. The potential benefits of the drug treatment program for individuals could be enhanced or reduced because of peer pressure among juveniles, all of whom were participating in the same program (p. 421)." Not only are these plausible social processes, many are predicted by existing criminological theories.

Consider as another example the random assignment of housing vouchers. The Moving to Opportunity (MTO) experiment randomly assigned housing vouchers to an experimental and control group in five cities in the mid 1990s. Much has been written about MTO and a recent debate ensued in the *American Journal of Sociology* which is relevant to my present argument (Clampet-Lundquist and Massey 2008; Ludwig et al. 2008; Sampson 2008). The MTO selected poor families (non randomly) at origin in a small geographic area of concentrated poverty. I have studied the Chicago MTO site where less than half took up (or complied with) the experiment and moved. My central point for present purposes is that social interactions among MTO participants would constitute statistical "interference" in violation of the "stable unit treatment value assumption" underlying experiments. Ludwig et al. argue that the interference concern is mitigated by the small sample of MTO participants in destination neighborhoods. But my concern is the interactions induced by the dense clustering in the *origin* neighborhoods.

Suppose, for example, that I was an MTO complier and used my voucher to move further away from poverty. If I or my children had friends or family back in the old neighborhood who were voucher eligible, and we complained about the hardship of moving or, alternatively, expressed enthusiasm about the new neighborhood, these social interactions could have influenced the moving decisions, destinations or other outcomes of those in our network—especially given the multiyear window of randomization (and thus lease-up) and the fact that many compliers drifted back to poor neighborhoods. Sobel (2006, pp. 1400–1401) provides other examples. Ludwig et al. assume that the intervention had no effect on noncompliers and that interference is unlikely because "fully 55%" reported no friends in baseline neighborhoods (p. 155). But this means that 45%, or nearly half, *did* have friends in the neighborhood (35% had family). My calculations also reveal that approximately 20% of the core residential population at baseline was in MTO, a nontrivial level of saturation. Moreover, social influences do not just derive from friends—acquaintances and "weak ties" may be just as important. Add to all this the fact that there were major events going on in the housing projects in and around the MTO site related to housing, such as the widely reported demolition of the nearby Robert Taylor Homes (with over 20,000 residents). It is hard to imagine that social interactions about moving would have been absent, especially among those receiving financial assistance in the way of vouchers. Other interference scenarios are plausible given that migration is a social process driven by preexisting networks.

The resulting bad news is that there is no statistical "fix" for SUTVA violations absent more assumptions that again require theory—overcoming interference requires us to

characterize and specify how it actually occurs. Otherwise causal estimates combine disparate and possibly countervailing effects in the study population (Sobel 2006).[7] As Berk notes (2005), one would need a plausible model of the interference, (observational) data with which to estimate its key features, and a statistical procedure that would produce consistent estimates. There is a growing statistical literature making advances on interference in randomized experiments (Hudgens and Halloran 2008; Rosenbaum 2007). Group randomized trials have been proposed as one solution but they require us to assume stability or lack of interference among the groups, such as among classrooms or neighborhoods. Raudenbush (2008) notes the problematic nature of this assumption for schools, where student mobility is considerable. Migration or social interactional patterns that are not a function of spatial proximity but instead span multiple neighborhoods or the entire city (Graif and Sampson 2010) cast doubt on SUTVA for neighborhood trials as well. Analogous to classrooms that are nested in schools and therefore an educational social structure, neighborhoods are embedded in larger communities and a metropolitan structure of inequality (Sampson 2008). More generally, social structure is by its very nature a threat to SUTVA. This should not be surprising, because randomization in effect describes a nonsocial world.

The larger lesson seems clear. While observational research is often sharply criticized for making "heroic" assumptions that cannot be proven, experiments are far from home free: they just have to make different assumptions and most of these cannot be proven either. In the MTO case, for example, assuming a lack of interactions between participating families that make up 20% of the original neighborhood and the other residents is not any more plausible to me than many of the assumptions invoked in observational research. Depending on the behavior, it may even be that experiments make more heroic assumptions than observational studies.

Myth 3: Experiments are More Policy Relevant than Observational Studies

Although it is widely (but not uniformly) accepted that the internal validity of experiments is strong, effective policy is fundamentally about *external* validity. External validity is probably not the best descriptor of the problem, however. Social science confronts contextual variations all the time (Weisburd 2010). It is not the location or population differences so much as that once a policy takes effect the rules of the game change, possibly inducing system level changes. So let us assume away the assumption and randomization concerns and consider what I will call the "policy transfer" rather than external validity problem. The gap between internal validity and policy transfer constitutes a major challenge to the idea that policy relevance necessarily favors experimental evidence. This is a general problem that requires knowledge about social behavior.

The idea is that once you shift from a specific study result or experiment to an institutionalized policy, new social processes and reactions take root. Suppose a randomized controlled trial suggested that busing black children to better resource-endowed white schools in the same (segregated) city improved educational performance. Suppose further this is a valid causal result. How policy relevant is the study? Is it transferable to a macro

[7] Technically, SUTVA means that the causal effect estimate in MTO is the difference between the average treatment effect and the "spillover" effect on the untreated (Sobel 2006, p. 1405). Not only are spillover effects are of great substantive interest, policy inferences could be led significantly astray by not distinguishing these two components of the treatment effect.

level policy, which is what matters? After all, we want to improve the lives of many children, not just a handful. It depends. If white families choose to withdraw from public schools because they do not want their children schooled with blacks in any proportion greater than some low threshold, then the composition of the system would rapidly shift and at some point the policy falls apart as a new equilibrium is set. Sadly, white flight was a reality in many US cities and resulted in nearly all minority schools with a lower tax base and fewer school resources. A new social system was thus created that changed the very conditions presumably producing the causal effect in the original experiment. This is not to say busing is "bad," only that unlike fertilizer plots, human reactions to policies are not necessarily aligned to the inferences drawn from the underlying "micro" experiment (or observational study). In some cities, moreover, integration can and did seem to work, with busing producing different (heterogeneous) effects. The relevant point is that social processes, such as segregation, sorting, and tipping points are relevant to the transference of the inference. The same is true for programs such as mixed-income housing. Transfer effects and conditions of extrapolation must be asked of all policy recommendations.

In criminology, for example, one can hypothesize that segregation, legal cynicism, and other social processes would condition the influence of mandatory arrest policies for domestic violence if taken to scale. The influential Minneapolis experiment may be valid, but transference to city-wide or national policy entails a process that goes well beyond the original experimental inference. Again this is not merely about external validity, such as whether the Minneapolis experiment holds in another city (in fact, replications were highly discrepant, yielding null, positive, and negative findings). I am referring to the more fundamental micro–macro link: what happens *even in Minneapolis* if a new policy is institutionalized based on the experiment? As in aggregation bias, homology of processes across levels cannot be assumed. Once implemented as formal policy, the causal parameter estimate of the underlying experiment confronts an altered social structure and a cultural world of values and politics. Carr (2010) makes a similar point in noting the unintended consequences of stop and frisk policies based on experimental evidence.

The economist Robert Lucas was awarded the Nobel Laureate for his work making an even more radical argument, suggesting that studies based on the past (and hence under different policy regimes) are problematically related to evaluating what will happen under a new policy. He argues for structural models that invoke a specific theory of how individual behavior changes as a function of traditional economic factors such as incentives and preferences. More broadly, Heckman points out the tension between the goals of experiments and structural models: "Under assumptions that ensure that it produces valid answers, the randomized experimental method bypasses the need to specify elaborate behavioral models. However, this makes experimental evidence an inflexible vehicle for predicting outcomes in environments different from those used to conduct the experiment" (Heckman 1992, p. 227). My point is that there is an apparent tradeoff between goals of assessing whether a specific program "works" and the more general application of a behavioral or social theory to policy change and thus whether and why a policy based on a particular study "will work." The implementation of policy that comes after an experiment will always occur with interdependent and forward thinking agents, in neighborhood or other contextual settings, and with unintended consequences—including, perhaps, human subjects acting in ways in direct opposition to what the treatment intends (Manski 2009). Policies reflect values, moreover, and not everyone in the population may share these values.

Finally, I think it is instructive to ponder an irony in the common appeal that experimental criminologists make to the medical model's policy relevance. One gets the distinct

impression from this intellectual move that gains in medicine or public health have come about primarily from RCT experiments. But this is not the case. The CDC (see http://www.cdc.gov/mmwr/preview/mmwrhtml/00056796.htm) has recently noted how the life expectancy of Americans has increased dramatically, lengthening by greater than 30 years since 1900 alone. The consensus is that about 25 years of this gain owes to advances in public health policy. The top ten list that follows (sorted alphabetically) is based on the opportunity for prevention and the impact on death, illness, and disability in the United States.

- Control of infectious diseases
- Decline in deaths from coronary heart disease and stroke
- Family planning
- Fluoridation of drinking water
- Healthier mothers and babies
- Motor-vehicle safety
- Recognition of tobacco use as a health hazard
- Safer and healthier foods
- Safer workplaces
- Vaccination

What is remarkable about this list is the dominance of observation-based public health interventions. The Salk vaccine is the most famous example of field trial or experimental success. But motor vehicle laws, safe foods, control of infectious disease, and reduction in smoking account for the lion's share of increases in longevity. To take one of the biggest health transformations, no crucial experiment was needed to conclude that smoking causes cancer. Consensus was achieved after sustained rigorous research of the old fashioned observational variety, with resulting transformative declines in the prevalence of smoking in the US. Population policies (e.g., on fertility, infant mortality) are also great successes in many countries around the world, with many of these policies derived from basic non-experimental research (see Smith 2009). If we seek a medical model as a guide for criminology, perhaps we should think more like those in public health. In fact, it is likely that criminology's top-ten list of policy-induced sources of the great American crime decline is also dominated by non-experimentally motivated changes (e.g., improved medical response and treatment; deterrence/incapacitation from increases in incarceration; community or problem-oriented policing, COMPSTAT).

Implications and New Directions

No method can be said to constitute the gold standard simply because "there is no gold standard" (Cartwright 2007, p. 11). By this logic I am not making a negative claim that experiments are uniquely flawed or problematic. Quite the contrary—no method is universally superior and the assertion of a gold standard is merely a rhetorical device. (Perhaps one could say there is a gold standard method if there was a corresponding gold standard question, but the latter does not exist.) The choice of method depends on the theoretical question and the nature of the phenomena under study, neither of which fall on a hierarchy. The hard truth is that we have little choice but to adapt in creative ways to the limitations that confront all social science inquiry.

It is true that I have posited three myths about experiments, but only to counteract overreaching in the claims of experimentalists and to lay the groundwork for consideration

of constructive ways to retire them. It bears emphasis that the three myths are social constructions that are not inherent to the experimental method itself. In the spirit of progress, my concluding argument is threefold: (1) users of experimental methods need to directly address fundamental assumptions and limitations that come into play when we are in a social world, (2) observational and experimental methods can be integrated to make science better (basic research), and (3), observational and experimental methods can be integrated to make policy better.

Consider that the counterfactual revolution has spawned not just increased use of experimental methods, but improved statistical and conceptual tools that explicitly rely on the potential outcomes model of causal inference (Morgan and Winship 2007). Indeed, there is a direct theoretical affinity between experiments and counterfactual models for analyzing observational data. Recent advances have been both theoretical and empirical. In statistical theory I would point to Sobel's (2006) elaboration of the causal parameters estimated in the MTO experiment, the work by Hong and Raudenbush (2008) for time-varying treatments in education, and the work of Rosenbaum (2007) on interference generally. In terms of empirical applications, we have seen multiple advances in propensity scoring, matching, fixed effects, regression discontinuity, and inverse proportional treatment weighting. These statistical techniques have been recruited to improve our ability to estimate causal parameters with observational data. As a result, "naïve regressions" are now widely recognized as flawed, and more attention is being paid to basic conditions necessary for causal inference in observational data, such as covariate balance and keeping inferences within the "support of the data."

A recent example in this journal is the use of trajectory methods combined with matching to estimate the effect of imprisonment on future crime (Nieuwbeerta et al. 2009). Using longitudinal data, careful specification of assumptions, use only of data that support causal analysis, and theoretically guided hypotheses, the authors make what I considerable credible inferences about the effect of imprisonment. Recent papers have used similar time-varying methods to estimate the effects of marriage and gangs on crime (Haviland and Nagin 2005; Sampson et al. 2006). These studies have their own limitations, of course, but I am not convinced that experimental data, if they exist, would necessarily do any better. How would we randomize marriage? Imprisonment? If we did a gang experiment, can we rule out social influence (interference)? How? Are the assumptions required to address interference better or worse than the assumptions of the observational studies? Suppose we can get "exogenous" variation in imprisonment, such as in the case of random variation in judge behavior induced by court dockets in one jurisdiction for one crime for one time period, as recently published in *Criminology* (Green and Wink 2010). Is the result better than the time-varying observational study of the effects of imprisonment by Nieuwbeerta et al. (2009) for an entire country? It is not clear from the method alone and many questions have to be asked, including some of those I raised above, before an answer could even be attempted. The logic of this essay implies that under some conditions, observational research may be more appropriate than an experiment based on randomization. But rather than make claims ex ante, a healthier approach to criminological science is to proceed along a dual track of statistical and experimental research, matching methods to substantive questions rather than starting with a favored method.

Another promising approach is to combine methods whenever possible and compare results *within the same study*. Ultimately it is an empirical question as to which method "does better" and under what social conditions. Where we can directly compare it is not the case that experiments invariably outperform observational or statistical methods, as is commonly argued by experimentalists. For example, it was thought to be the case in labor

economics that statistical methods got it wrong about work training programs as a result of the influential argument of LaLonde (1986). But that was 25 years ago and some have claimed observational results similar to the experiment when examined under stringent econometric modeling (e.g., difference in difference models; sensitivity-based propensity modeling). Heckman (1992, p. 213) reports that when proper sensitivity tests are performed on the LaLonde work data, "they eliminate all but the nonexperimental models that reproduce the inference obtained by experimental methods." In criminology, Berk and colleagues compared the results from a randomized clinical trial evaluating strategies for parole and probation supervision among low risk offenders to results from a regression discontinuity "quasi-experiment." They report an equally strong outcome: "the results from the two approaches are effectively identical" (Berk et al. 2010, p. 191). A more general result across a wider range of outcomes is found in a study led by Thomas Cook, who literally wrote the book on experimental design with the late Donald Campbell. Assessing twelve within-study comparisons, Cook et al. (2008, p. 745) find strong similarities in causal results between randomized experiments and both quasi-experimental (e.g., regression discontinuity) and observational studies with well specified matching and careful consideration of selection processes. They conclude that the correspondence in causal findings "contradicts the monolithic pessimism emerging from past reviews of the within-study comparison literature".

It thus seems that the hypothetical biases of observational research that are warned about in abstract examples (e.g., Weisburd 2010, pp. 5–6) do not necessarily pan out in careful research. The hypothesis that I draw is that that the alignment of results across experimental and observational methods is not only possible but will increase as a function of (a) the quality of empirical measures (e.g., to provide better matching), (b) comprehensive specification by a substantive theory of the outcome, (c) the substantive and empirical modeling of selection into treatment,[8] and (d) the rigor of the employed statistical model. At the least it would seem that a research program in quantitative criminology is in order to address this hypothesis and to calibrate the conditions under which convergence is most likely to occur, and vice versa.

Rethinking Quantitative Policy Evaluation

Finally, I would argue that we need new kind of policy paradigm, where the policy transfer problem, treatment-effect heterogeneity, and theories about behavioral change under different policy regimes are integrated with the experimental evaluation of "what works." In so doing we need to recognize that an unbiased causal inference is not the only goal (or the most important even) and broaden our conception of what constitutes evidence. Establishing causal mechanisms is as much about theory and observation as it is about estimating a single parameter in a statistical model *or* experiment (see also Sampson 2008; Sampson Forthcoming). Put simply, there is no "theory free" policy inference of the sort that politicians seem to wish for. Criminologists would thus do well to resist the urge (or demand) to seek technical policy fixes.

Experiments themselves also need to be recast. Most experiments in criminology and social science generally have privileged micro or individual-level interventions. But micro-level inference is not inherent to experiments and randomization can be exploited at the population level. There are now supra-individual interventions in economic development

[8] For a recent argument on the specification of "selection bias" as a social and causal process, see Sampson and Sharkey (2008).

(Deaton, 2008) and public health (Kamo et al. 2008; Sikkema et al. 2000) and there are place-based interventions in policing (Weisburd et al. 2006) that may be paving the way for a more robust population-level consideration of causality. Linking "macro" interventions with observational knowledge from the long literature on communities and crime seems an especially promising direction, as does the design of new community-level and perhaps societal interventions.

With these moves the field of criminology can be a leader by transcending attempts at methodological hegemony and taking a more pluralistic stance on the nature of what counts as evidence. Criminologists should at the least dispense with the use of the "gold standard" language (even if in quotes!) and get on with the hard business of doing good research. Berk (2005) suggests that experiments are better conceived as the "bronze" standard, a similar but different concession. To my mind it is better to jettison the Olympic metaphors altogether.

Acknowledgments I thank Gary King, Carly Knight, John Laub, Steve Raudenbush, P–O Wikström, and Chris Winship for their feedback.

References

Angrist JD, Pischke J-S (2010) The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. J Econ Perspect 24:3–30

Angrist J, Imbens G, Rubin D (1996) Identification of causal effects using instrumental variables. J Am Stat Assoc 91:328–336

Berk R (2005) Randomized experiments as the bronze standard. J Exp Criminol 1:417–433

Berk R, Barnes G, Ahlman L, Kurtz E (2010) When second best is good enough: a comparison between a true experiment and a regression discontinuity quasi-experiment. J Exp Criminol 6:191–208

Carr PJ (2010) The problem with experimental criminology: a response to Sherman's 'evidence and liberty'. Criminol Crim Justice 10:2–10

Cartwright N (2007) Are RCTs the gold standard? Biosocieties 2:11–20

Clampet-Lundquist S, Massey DS (2008) Neighborhood effects on economic self-sufficiency: a reconsideration of the moving to opportunity experiment. Am J Sociol 114:107–143

Clear T (2010) Policy and evidence: the challenge to the American society of criminology: 2009 presidential address to the American society of criminology. Criminology 48:1–25

Cook TD, Shadish WR, Wong VC (2008) Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. J Policy Anal Manage 27(4):724–750

Deaton A (2008) Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development. The Keynes Lecture, British Academy, London

Gangl M (2010) Causal inference in sociological research. Annual Review of Sociology Forthcoming

Graif C, Sampson RJ (2010) Inter-neighborhood networks and the structure of urban residential mobility. Harvard University, Department of Sociology, Cambridge

Green DP, Wink D (2010) Using random judge assignments to estimate the effects of incarceration and probation on recidivism among drug offenders. Criminology 48:357–387

Haviland AM, Nagin DS (2005) Causal inferences with group based trajectory models. Psychometrika 70(3):557–578

Heckman JJ (1992) Randomization and social policy evaluation. In: Manski CF, Garfinkel I (eds) Evaluating welfare and training programs. Harvard University Press, Cambridge, pp 210–229

Heckman JJ (2001) Accounting for heterogeneity, diversity and general equilibrium in evaluating social programs. Econ J 111:654–699

Heckman JJ (2005) The scientific model of causality. Sociol Methodol 35:1–97

Holland P (1986) Statistics and causal inference. J Am Stat Assoc 81:945–970

Hong G, Raudenbush SW (2008) Causal inference for time-varying instructional treatments. J Educ Behav Stat 33(3):333–362

Hope T (2009) The illusion of control: a response to professor sherman. Criminol Crim Justice 9:125–134

Hough M (2010) Gold standard or fool's gold: the pursuit of certainty in experimental criminology. Criminol Crim Justice 10:11–32
Hudgens MG, Halloran ME (2008) Toward causal inference with interference. J Am Stat Assoc 103: 832–842
Imai K, King G, Stuart E (2008) Misunderstandings among experimentalists and observationalists about causal inference. J R Stat Soc Ser A 171(Part 2):481–502
Kamo N, Carlson M, Brennan RT, Earls F (2008) Young citizens as health agents: use of drama in promoting community efficacy for hiv/aids. Am J Public Health 98:201–204
Keane MP (2010) A structural perspective on the experimental school. J Econ Perspect 24:47–58
LaLonde R (1986) Evaluating the econometric evaluations of training programs with experimental data. Am Econ Rev 76(4):604–620
Lieberson S (1985) Making it count: the improvement of social research and theory. University of California Press, Berkeley
Lieberson S, Lynn F (2002) Barking up the wrong branch: scientific alternatives to the current model of sociological science. Annu Rev Sociol 28:1–19
Ludwig J, Liebman JB, Kling JR, Duncan GJ, Katz LF, Kessler RC et al (2008) What can we learn about neighborhood effects from the moving to opportunity experiment? A comment on Clampet-Lundquist and Massey. Am J Sociol 114:144–188
Manski CF (2009) Diversified policy choice with partial knowledge of policy effectiveness, 10th annual jerry lee crime prevention symposium. University of Maryland Inn and Conference Center, Adelphi
Moffitt R (2005) Remarks on the analysis of causal relationships in population research. Demography 42: 91–108
Morgan S, Winship C (2007) Counterfactuals and causal inference: methods and principles for social research. Cambridge University Press, New York
Nieuwbeerta P, Nagin DS, Blokland A (2009) Assessing the impact of first-time imprisonment on offenders' subsequent criminal career development: a matched samples comparison. J Quant Criminol 25: 227–257
Raudenbush S (2008) Advancing educational policy by advancing research on instruction. Am Educ Res J 45(1):206–230
Rosenbaum PR (2007) Interference between units in randomized experiments. J Am Stat Assoc 102: 191–200
Sampson RJ (2008) Moving to inequality: neighborhood effects and experiments meet social structure. Am J Sociol 114:189–231
Sampson RJ (Forthcoming) Neighborhood effects: social structure and community in the American city. University of Chicago Press: Chicago
Sampson RJ, Sharkey P (2008) Neighborhood selection and the social reproduction of concentrated racial inequality. Demography 45:1–29
Sampson RJ, Laub JH, Wimer C (2006) Does marriage reduce crime? A counterfactual approach to within-individual causal effects. Criminology 44:465–508
Sherman LW (2009) Evidence and liberty: the promise of experimental criminology. Criminol Crim Justice 9:5–28
Sikkema K, Kelly JA, Winett RA, Solomon LJ, Cargill VA, Roffman RA et al (2000) Outcomes of a randomized community-level hiv prevention intervention for women living in 18 low-income housing developments. Am J Public Health 90:57–63
Smith HL (2009) Causation and its discontents. In: Engelhardt H, Kohler H-P, Fürnkranz-Prskawetz A (eds) Causal analysis in population studies. Springer, New York, pp 233–242
Sobel M (2006) What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. J Am Stat Assoc 101:1398–1407
Tilley N (2009) Sherman vs. Sherman: realsim vs. rhetoric. Criminol Crim Justice 9:135–144
Weisburd D (2010) Justifying the use of non-experimental methods and disqualifying the use of randomized controlled trials: challenging the folklore in evaluation research in crime and justice. J Exp Criminol 6:209–227
Weisburd D, Wyckoff L, Ready J, Eck J, Hinkle J, Gajewski F (2006) Does crime just move around the corner? A controlled study of spatial displacement and diffusion of crime control benefits. Criminology 44:549–592
Weisburd D, Mazerolle L, Petrosino A (2010). The academy of experimental criminology: advancing randomized trials in crime and justice. http://www.crim.upenn.edu/aec/AECCriminologist417.doc
Wikström P-O, Sampson RJ (eds) (2006) The explanation of crime: context, mechanisms, and development. Cambridge University Press, Cambridge