"EVIDENCE-BASED POLICY" SHOULD REFLECT A HIERARCHY OF EVIDENCE

Jennifer L. Doleac

Janeen Buck Willison and I agree on one important point: We need a lot more research on wrap-around services and prisoner reentry programs more broadly. Perhaps it is not surprising that that is the conclusion drawn by two researchers, despite our different takes on existing evidence. But this is the beauty of social science: We can use theory and existing evidence to formulate new hypotheses, and test those new hypotheses in future studies. There is always more to learn.

But as we review current and future evidence, it is important to acknowledge that some studies produce more accurate estimates than others. Interpreting evidence requires more than simply comparing the number of studies finding benefits with the number finding null or detrimental results. Results should be weighted based on where a study falls in a hierarchy of evidence: raw correlational analyses near the bottom, outranked by studies with rich control variables, then by studies using matched comparison groups, then by studies using natural experiments to avoid selection bias (e.g., studies using sound difference-in-difference, regression discontinuity, and instrumental variable designs), then randomized controlled trials (RCTs) at the top.[1] As we move up the hierarchy, we are more sure that the methods identify the causal effect of a program, separate from pre-existing trends or pre-existing differences between participants and non-participants.

Practitioners sometimes use the phrase "evidence-based policy" (EBP) to describe a program supported by evidence that falls anywhere on this hierarchy, as if any evidence at all is sufficient. This can easily lead them astray. In the reentry space, very few programs are supported by evidence at the natural experiment or RCT tier; the vast majority of studies on reentry programs use rich controls or matched comparison groups (hereafter referred to as "non-experimental methods") to identify causal effects. This is partly because such designs are more convenient: they don't require up-front planning or changing how a program is implemented. But this convenience comes at a cost: participants may be different from non-participants in ways that researchers cannot control for, resulting in selection bias.

Non-experimental methods can produce unbiased results if selection into a program based on unobservable factors (such as motivation or initiative) is low. As researchers, we may differ in our priors about how much selection on unobservables will matter in a given context: if important unobservables are correlated with observable characteristics, then using rich controls should limit selection bias. But it is difficult to know how successful such methods are until someone does a study that more plausibly avoids selection bias, by using a natural experiment or RCT. Only then can we compare the results across methods and see the importance of selection bias in a given context.

There are good reasons to be skeptical of studies based on non-experimental methods in the prisoner reentry context. Reentry programs typically screen applicants based on motivation to change or good behavior and may require that applicants actively volunteer. This means that the treatment group will likely be positively-selected on these qualities (motivation, behavior, and initiative). These personal qualities (typically unobservable to the researcher) surely improve reentry outcomes on their own—that is, those who wind up in the treatment group would probably have done better than those in the comparison group, even without the program. Non-participants who look similar to the participants based on observable characteristics (criminal history, age, gender) may still be different on these unobservable dimensions. For this reason, studies that try to reduce selection bias with controls or matched comparison groups may still be biased toward finding beneficial program effects.[2]

How does all of this relate to our current discussion of wrap-around services in the prisoner reentry context? My Point essay reviewed several studies that find null or detrimental effects of such programs. In her Counterpoint essay, Ms. Buck Willison discusses several studies that find beneficial effects of such programs. She attributes the mixed results in the literature to variation in program details and implementation. I see a more fundamental difference: the methods used in the evaluation itself. Studies that suggest wrap-around services are effective are typically based on matched comparison groups. Studies that find no effect or detrimental effects are typically based on RCTs.

Rather than complicating the picture, the recent RCTs have clarified it: they demonstrate the importance of selection bias in this context. The RCT results demonstrate that simply matching participants and non-participants on observable characteristics is not enough to avoid selection bias: the treatment groups are positively selected in ways that are not observable to researchers, and this biases studies toward finding beneficial program effects. Such studies should therefore be heavily discounted in the prisoner reentry context.

WHERE DO WE GO FROM HERE?

The best evidence on wrap-around services and similar holistic approaches to improving reentry outcomes suggest that these programs don't benefit participants—and may be actively detrimental. Such programs are extremely common across the country. Given this, practitioners should reconsider their current programming, and acknowledge that they might be wrong about the successes of even highly-praised efforts.

Those who are in a position to steer reentry programming should push for more RCTs of existing and future programs. Many practitioners are understandably reluctant to withhold promising interventions from a control group. But in this case, it is possible, even likely, that their most prized programs are doing at least as much harm as good. Given this uncertainty, it is completely ethical to provide a variety of less-intensive programs to random subsets of offenders to see if those alternatives achieve more success. This is the best way to figure out which programs work best, and for whom.

RCTs are often infeasible in the criminal justice context due to safety concerns. In these cases, phased roll-outs of new programs can provide useful natural experiments (e.g., implementing a program gradually by facility or housing unit, or using strict eligibility cutoffs such as age or risk score). Practitioners and academic researchers should work together to design evaluations that are both feasible and rigorous. This requires more work, to be sure, but evidence-based policy depends on it. As discussed above and in my Point essay, the popular fall-back option of implementing a program and then finding a comparison group after the fact simply cannot be trusted in this context.

JENNIFER L. DOLEAC is an Associate Professor of Economics at Texas A&M University, 4228 TAMU, College Station, TX 77843 (e-mail: jdoleac@tamu.edu).

[1] Even within these categories, different contexts and identification strategies might tell us about different populations. For instance, a regression discontinuity design identifies the effect of a program on the marginal participant, which might be different from the average effect on all participants.

[2] Even in studies based on natural experiments and RCTs, selective attrition can still bias results. Some studies randomly assign access to a program, but then compare the subset of individuals who actually participated in or completed the program with the full control group. The participants and completers will be a positively-selected group, thus reintroducing selection bias. Studies that use self-reported measures as outcomes also risk introducing selection bias during the follow-up period. This is because researchers will inevitably be unable to find and survey some participants. If non-respondents are not a random subset of the initial treatment and control groups, then this selective non-response will bias results. A straightforward way to avoid this problem is to use administrative data to measure outcomes: data on arrests, convictions, incarcerations, employment, social service receipt, and mortality are typically available (though take more work to access). However, many existing studies still use self-reported measures of outcomes like employment and recidivism to measure program effects.