METHODS FOR RESEARCH REVIEW[1]

January 9, 2021

Ben Feigenberg, University of Illinois at Chicago
Dylan Fitzpatrick, University of Chicago
Jens Ludwig, University of Chicago
Morgan Williams, Jr., New York University

I. Overview

The University of Chicago Crime Lab is a non-profit, non-partisan faculty-led research center established 12 years ago to partner with the public sector and use the power of data and data science to improve the fairness and effectiveness of the criminal justice system. Examples include, for instance, helping implement reforms at the Chicago Police Department that were recommended by the US Department of Justice Civil Rights Division, the Chicago Mayor's Police Accountability Task Force, and are written into the consent decree between the Illinois Attorney General and CPD, as well as working on a number of community-led public safety initiatives in partnership with local NGOs such as Heartland Alliance, Institute for Nonviolence Chicago, Lawndale Christian Legal Center North Lawndale Employment Network, Children's Home & Aid and Youth Advocate Programs.

The Crime Lab is pleased to partner with the Council on Criminal Justice (CCJ) Policing Task Force by summarizing the available data on different police-reform topics. The goal of the Crime Lab research reports is to summarize as objectively as possible what the available empirical literature can, and can't, tell us about the potential desirable effects and any unintended consequences of different changes to policing policy or practices. This evidence is intended to help inform the task force's deliberations and its policy recommendations.

This short memo lays out our approach to the evidence and how to summarize it in our research reports. In brief summary:

- The available empirical research on policing reform varies enormously in quality and quantity, depending on the particular reform topic
- We use the same standards of evidence for drawing conclusions across all topics. These standards are discussed in detail in section III below, and aspire to mirror the approach employed by National Academy of Sciences research committee reports.
- For each policing reform topic we prioritize for discussion the best studies within the relevant literature for that topic. Given the variation in quality and quantity of research on each topic, the number and types of studies we discuss will also vary across topics.

II. The Causal Inference Challenge

The goal of police reform is to improve policing outcomes (more fairness in receipt of enforcement actions or police use of force, less use of force overall, etc.). Police reforms can only accomplish that goal if they are causally related to these outcomes. To understand whether a

[1] Please direct comments or questions to Jens Ludwig (jludwig@uchicago.edu)

given type of police reform is indeed related to policing outcomes, we need to have some comparison of what happens in a department with and without the reform in place.

The key challenge in isolating the causal effect of some change in policing policy or practice on outcomes is that police departments often have some degree of control over what they're doing and when they do it:

- The most common way that the news media (and typically policymakers as well) tries to figure out whether some reform has changed outcomes is by comparing what happened within a department after the reform is in place to what the department was like beforehand (what social scientists would call a *time-series* or *before-after design*). The problem with that sort of comparison is that lots of other things are changing over time as well within a jurisdiction, besides the reform. In particular we run the risk of confounding the effect of the policy or practice with whatever might be changing in the jurisdiction to cause it to want to change what it is doing.
- Alternatively, if we compare outcomes of two or more departments at a point in time that have different police policies or practices (a *cross-section regression*), how do we know whether we are seeing the effect of the policy or practice, or instead seeing the effect of whatever other things differ across the police agencies that cause some to choose to adopt a given policy or practice and other agencies to eschew it? It's natural to think we can just account for these other sort of confounding factors by measuring and then statistically controlling for them. But jurisdictions vary from one another in countless ways; in a complex social system like a city or state, we are unlikely to ever fully measure all potentially relevant confounders.

Solving this problem is enormously difficult and is in many ways *the* key challenge for all of social science. It is easy for practitioners to look at this challenge and think this is just 'academics being academics,' looking for perfection or niceties without regard to real-world imperatives or the need for concrete action. But the data themselves have shown us that failure to adequately account for confounding in the real world can lead to estimates that are not only of the wrong size, but of the wrong sign – that is, when social science estimates are wrong they might not only be a little bit off, they might mistakenly even tell us to do the opposite of the thing that will actually make the world better.[2]

The difficulty – and importance – of solving this problem can be seen by analogy in another area where lives are at stake: medicine. The global COVID-19 pandemic of 2020 is on track to infecting 100 million people by the time it's done, with at least 2 million deaths and millions more having lost their livelihoods as a result of pandemic-related economic disruptions. Governments all over the world were under enormous pressure to come up with a cure. But every government in the world resisted approval of any cures until they were subject to evidence

[2] A convincing early empirical demonstration of how large the bias can be from misleading social science estimates, and how they can even be of the wrong sign, comes from the ground-breaking study from our University of Chicago colleague Robert LaLonde (1986). That finding has now been replicated within many other subsequent 'within-study comparisons.' While it is the case that sometimes the bias can be modest, we have no way to tell in any given application what the magnitude of the bias is. That makes it hard to know whether we're in even the right ballpark of the truth. This is not ideal for policy purposes.

from a randomized controlled trial (RCT). These RCTs enroll large numbers of people and randomly assign some to the new medical treatment being studied and others to status quo treatment (or placebo if there is no status quo treatment). By essentially flipping a coin to decide whether someone receives the new intervention, the treatment and control groups will be similar on average with respect to all other determinants of health outcomes – except for receipt of the new medical treatment. This is how we isolate the effects of that treatment.

Of course in policy applications, RCTs are not always feasible. We turn next to how we deal with that challenge in practice in our research reports.

III. Hierarchy of evidence

A. RCTs

For the reasons described above, we put the greatest weight on RCT evidence when it is available. Within the set of RCTs that are available, we put more weight on those that:

- Have relatively larger sample sizes, so that the estimated impacts are less vulnerable to being slightly off due to sampling variability (that is, we have more precise estimates)
- Demonstrate that randomization was carried out correctly – that is, show that the randomized treatment and control groups on average have similar characteristics at the time of random assignment itself
- Document and report the nature of the ‘dosage’ of the intervention being studied that is received by the treatment versus control group
- Have low rates of attrition in the collection of follow-up outcome measures, and similar levels of attrition between treatment and control groups. Even if a RCT succeeds in creating two groups that are comparable at baseline, if (say) we have follow-up outcome data on only half the treatment group but all of the controls, the people or places for whom we actually can measure outcomes may no longer be comparable.
- Present adequate information about the statistical inference procedures to allow readers to know whether the results are likely to be due to chance or not, and how large of an impact can or can’t be ruled out by the available data. This includes reporting standard errors or confidence intervals, so it is possible to make judgments about the risk of concluding an intervention is not effective when in fact the problem is the estimates have large uncertainty intervals around them (that is, the risk of ‘false negatives’). This also includes adequately accounting for the number of separate statistical estimates that have been carried out, since without adjustment for this, when we generate enough statistical estimates one is sure to be statistically significant even if just by pure chance (put differently, this is the risk of ‘false positives’).

B. Natural experiments

Of course RCTs are not always available in the policy world. A major development over the last several decades has been the increased emphasis on finding sharp shifts in what nature does out in the world – so-called “natural experiments” – that, ideally, mimic in the real world something like the idealized randomized experiment we would wish to carry out in practice if we could. Angrist and Pischke (2010) term this the *credibility revolution* within the social sciences.

A canonical example in the area of policing is the study of the Clinton-era COPS community policing program by economists William Evans and Emily Owens (2006). The natural experiment at the heart of the study is the distribution of federal government COPS grants that went to some jurisdictions but not others. The authors show that while the jurisdictions that received more COPS funding had different levels of crime at the time the funding was distributed than the places that got less COPS funding, the two groups of jurisdictions had similar trends in crime up to that point. So the study was able to isolate the effects of the COPS funding on policing outcomes by comparing trends between 'treatment' and 'control' jurisdictions from the pre-COPS to post-COPS period.

We prioritize natural experiment studies that demonstrate that the treatment and control jurisdictions are similar at baseline, either in levels or in pre-trends depending on the specific research design that is being employed.

The distinction between the 'natural experiment' approach and the traditional panel data approach that we discuss next is that with the natural experiment we have some clear understanding of the reason why some jurisdictions adopt a policy change at the time they do, and others do not. For example, we know that police funding changes at a given time because this jurisdiction received a COPS grant from the US Department of Justice. That allows us to make a judgment about the plausibility that the change is unrelated to other confounding factors that might be changing at the same time.

C. Panel data studies

For many years, the work-horse research design in the social science study of policing (and almost every other topic for that matter) was the attempt to combine both cross-section and time-series designs by collecting data on multiple jurisdictions, measured at different points in time, in a so-called *panel data study*. The intuition is that two jurisdictions might be systematically different in ways that lead them to have very different levels of the outcomes of interest year after year. But if they have similar trends in outcomes over time, we can look to see if there is a difference or 'break' in trends across the different places once one adopts a new policing policy or practice and the other does not.[3]

The minimum necessary condition for any panel data research design to be credible is evidence that the 'treatment' and 'control' jurisdictions have similar trends in outcomes even before the 'treatment' areas adopt whatever new policy or program is being studied. So we strongly prefer panel-data studies that at the very least provide evidence that the jurisdictions being compared have similar pre-trends in outcomes.

Even then, there remains the risk that the 'treatment' jurisdictions choose to adopt the new policy or practice at a particular time specifically in response to some change in social conditions that are difficult to measure in data, which would lead us to confound the effect of the policy or practice of interest with that unmeasured time-varying confounder. Within the crime literature, for example, the limitations of these panel-regression designs have been discussed in exhaustive detail within the context of analysis of the effects of gun carrying laws on crime outcomes (see

[3] The same logic obviously holds when there are more than two jurisdictions in the panel dataset.

for example Lott and Mustard, 1997, Lott, 1998, Wellford and Pepper, 2004, Donohue, Aneja and Weber, 2019). States seem to have been adopting concealed-carry laws in part to respond to recent, local changes in crime trends, with the consequence that empirical estimates of the effects of these laws wind up being extremely sensitive to even minor changes in estimation approach.

D. Research designs we do not consider as evidence

As implied by the discussion above, we do not consider as credible evidence:

- Cross-section regressions
- Simple pre-post comparisons
- Panel data regressions that do not provide evidence of similarity in pre-trends in the outcomes being examined
- Natural experiment designs that do not provide credible evidence of similarity in pre-trends in levels or pre-trends of outcomes (depending on the design), and that do not have a conceptually plausible source of exogenous identifying variation
- Randomized controlled trials that cannot establish that randomization was done correctly, that the treatment group received adequate ‘dosage’ of the intervention above and beyond what the control group received, and/or that suffers from large amounts of sample attrition in the follow up outcome measurement (or large differences in attrition between treatment and control groups). We do still incorporate RCTs that adequately account for statistical inference (for example doesn’t present standard errors or fails to account for multiple testing issues), although they receive less weight when drawing our conclusions.

E. Drawing on evidence from analogous settings

In situations where high-quality evidence on the police reform of interest is limited, we look to closely related topics that are not necessarily focused on policing specifically but for which there may be a richer base of evidence (for instance, if there is limited evidence on the effects of implicit bias training with police, we consider evidence on the effects of implicit bias training that have been carried out for other occupations and in other contexts). We include a discussion in these cases about how that adjacent evidence relates to the primary topic of interest.

IV. Presentation of results in the research reports

While the relative quality of these research designs is not a function of the particular topic being investigated, the attention paid in each report to studies of varying rigor is determined by the quality and breadth of the associated literature. At one extreme, for topics where a large number of high-quality studies exist (randomized controlled trials or compelling natural experiments), we focus our attention almost exclusively on these studies. In these cases, we may mention specific research studies based on less rigorous research designs when these studies are particularly well known and highly cited or are representative of a larger body of existing research employing similar methods. However, our ultimate conclusions regarding the causal impact of the program or policy being studied will be determined by findings from the subset of studies that employ the most rigorous evidentiary standards.

For topics on which the existing literature is more limited or there are no or few studies employing the most rigorous methodological designs, we include a more detailed discussion of findings from the full range of research designs employed in the existing literature. In these cases, we synthesize the prior findings and discuss the consistency of research conclusions based on the available literature. However, since the bias associated with prior estimates from work that employs less rigorous methodologies may be large (even when findings are consistent), we are reluctant to draw strong conclusions from this past work and will note this in our summary.

For the subset of topics where little or no evidence is available based on the most rigorous study designs, we will summarize all existing work when the related literature is relatively limited. In cases where there is a large body of work that employs less rigorous methodologies, we will typically reference findings from only the most well-known and highly cited studies associated with each available research design, but we will summarize more generally the range of findings associated with each research design employed in past work. To the extent possible, we will also comment on potential sources of bias and the efforts undertaken by researchers to address identification concerns.

We also try to pay special attention to understanding the source of conflicting findings, or null findings. A study might find no effect because the policy or practice doesn't work, or because it was not actually implemented or delivered with any sort of real intensity during the study period – that is, to revert back to medical analogy the 'dosage' may be low. Variability in 'dosage' across places could similarly help us understand variation in impacts across studies, if they are carried out in different locations or time periods. And of course there may be other differences across places besides dosage that moderate the effects of policies and practices on outcomes. We pay careful attention to each of these issues in our reviews.

Finally, in cases where there is a large enough body of high-quality studies where it makes sense to quantitatively summarize the literature's findings, we rely on estimate-averaging across studies. An alternative approach to this sort of quantification is meta-analysis, which essentially statistically models the variation across studies in impact estimates using the features of each of the studies themselves as the explanatory variables (see for example Lipsey and Wilson, 2001, or Borenstein et al. 2011). The goal is to enable readers to understand how specific features of a study affect the average study impact estimate, where the variation caused by a study feature essentially comes from reading off a coefficient within the meta-analytic regression. Our approach instead is to present the simple average of study impacts, and then show how this simple average changes as the composition of the studies changes defined by different rules about what study features are used for inclusion or exclusion from the study sample. Our goal with this approach is improved intuition and transparency for readers.

REFERENCES

Angrist, Joshua D. and Jorn-Steffen Pischke (2010) “The credibility revolution in empirical economics: How better research design is taking the con out of econometrics.” Journal of Economic Perspectives. 24(2): 3-30.

Borenstein, Michael, Larry V. Hedges, Julian PT Higgins, and Hannah R. Rothstein. *Introduction to meta-analysis*. John Wiley & Sons, 2011.

Donohue, John J., Abhay Aneja and Kyle D. Weber (2019) “Right-to-carry laws and violent crime: A comprehensive assessment using panel data and a state-level synthetic controls analysis.” Journal of Empirical Legal Studies.

LaLonde, Robert J. (1986) “Evaluating the econometric evaluations of training programs.” American Economic Review. 76(4): 604-620.

Lipsey, Mark W., and David B. Wilson. *Practical meta-analysis*. SAGE publications, Inc, 2001.

Lott, John R. and David B. Mustard (1997) “Crime, deterrence and right-to-carry concealed handguns.” Journal of Legal Studies. 1-68.

Lott, John R. (1998) More Guns, Less Crime. Chicago: University of Chicago Press.

Wellford, Charles and John V. Pepper (2004) Firearms and Violence: A Critical Review. Washington, DC: National Academies Press.