NBER WORKING PAPER SERIES

SHACKLING THE IDENTIFICATION POLICE?

Christopher J. Ruhm

Working Paper 25320
http://www.nber.org/papers/w25320

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2018

This paper was presented on November 19, 2018 as the Presidential address at the 88th Annual Meeting of the Southern Economic Association, in Washington DC. I thank the University of Virginia Bankard Fund for providing financial support. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2018 by Christopher J. Ruhm. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Shackling the Identification Police?
Christopher J. Ruhm
NBER Working Paper No. 25320
November 2018
JEL No. A11,B4,C50,H0,I0,J0,O0

ABSTRACT

This paper examines potential tradeoffs between research methods in answering important questions versus providing more cleanly identified estimates on problems that are potentially of lesser interest. The strengths and limitations of experimental and quasi-experimental methods are discussed and it is postulated that confidence in the results obtained may sometimes be overvalued compared to the importance of the topics addressed. The consequences of this are modeled and several suggestions are provided regarding possible steps to encourage greater focus on questions of fundamental importance.

Christopher J. Ruhm
Frank Batten School of
Leadership and Public Policy
University of Virginia
235 McCormick Rd.
P.O. Box 400893
Charlottesville, VA 22904-4893
and NBER
ruhm@virginia.edu

Introduction

This paper provides some "big picture" thoughts on tradeoffs in the types of empirical research conducted in the broad area of applied microeconomics, by which I include fields such as labor, health, public, education and environmental economics. Nothing here is necessarily original but, hopefully, it will be thought-provoking.

When I was in graduate school and in my first years thereafter, applied microeconomics was dominated by increasingly sophisticated econometric methods that, to my view, often yielded little confidence in the results because their plausibility rested upon often opaque and possibly incorrect statistical assumptions. Others, with far more expertise than I, have previously expressed similar reservations. For instance, Edward Leamer's famous (1983) article, "Let's Take the Con Out of Econometrics", emphasized the fragility of many empirical estimates to even small changes in model specifications.

Therefore, it was a revelation to me to read David Card's (1990) paper using the Mariel Boatlift as a plausibly exogenous event for studying immigration effects on the wages of low-skilled workers. This approach struck me as clean, transparent and relatively-assumption free. I was not the only alone in feeling this way. This and other studies emphasizing natural experiments led to what Angrist and Pishke (2010) reasonably described as a "credibility revolution" in empirical economics. One result is that all well-trained applied micro-economists now think about identification strategies and have considerable familiarity with "quasi-experimental" methods such as instrumental variable (IV), regression discontinuity (RD), differences-in-differences (DD) and panel data techniques. Taking one step further, "experimental" designs with randomized controlled trials (RCTs) are increasingly viewed (by some) to be the "gold standard" of applied economic research.

Evidence of the rapid growth in these methods can be obtained from Kleven's (2018) textual analysis of more than 4500 National Bureau of Economic Research working papers in the area of public economics published since 1975. He shows that over the five year periods ending in 1980 and 2017 the fraction of papers using the term identification rose from 1% to 46%. Over the same period, the share using quasi-experimental methods, laboratory experiments and RCTs grew from 0% in all three cases to 21%, 7% and 9%. By the five-year period concluding in 2017, DD, RD and event study methods were employed in 22%, 12% and 9% of the papers. Nor are these patterns limited to public economics. Cameron et al.'s (2016) comprehensive review indicates that the vast majority of impact evaluations in

international development now use experimental or quasi-experimental methods.[1] They found that there were an average of seven such evaluations were published annually from 1981-1999. But the numbers rose rapidly thereafter reaching 82 in 2004, 173 in 2008, 296 in 2010, and 377 in 2012 (the last year studied). Similarly, Panhans and Singleton (2017), demonstrate rapid increases in mentions of experimental and quasi-experimental methods (differences-in-differences, regression discontinuity, natural experiment and randomized controlled trials) and instrumental variables since the early 1990s for the top four general economics journals and eleven top field journals.[2]

The increased use of experimental and quasi-experimental methods has facilitated major contributions to empirical research across a wide variety of fields. However, I am concerned that the "identification police" are becoming a too powerful force in our profession. I use the term identification police as a short-hand to indicate an almost exclusive focus on research designs (generally experimental and quasi-experimental) that provide clean identification. My worry is that excessive reliance on such methods may move us away from examining issues that are of fundamental significance but for which unambiguous causal inference is more difficult to obtain. There are often likely to be tradeoffs, both direct and indirect, between the confidence in and importance of the results obtained from empirical research.[3] Disproportionate attention to the former may lead to us to focus on inquiries that, while clearly identified, may sometimes be of secondary value, whereas other critical issues may be neglected.

Limitations of Experimental and Quasi-experimental Approaches

The limitations of experimental and quasi-experimental approaches have been discussed previously in considerable detail (e.g. Heckman and Smith 1995; Keane 2010; Deaton and Cartwright 2018). One concern is that the average treatment effects (ATE) in small RTC's may be imprecisely estimated, even if they are unbiased. Another is that the estimated treatment effects may poorly indicate the ATE of interest. For instance, the local average

[1] Over the period examined (1990-2012), 66% of the evaluations were based on RCTs while 17%, 8% and 2% employed DD, IV and RD methods.

[2] They also indicate differences in the use of these strategies across fields. For example, they highlight the frequency of DD methods in health economics, RD estimates in public economics and RCTs in development economics.

[3] I am not claiming that this framing is novel. For instance, after drafting this paper, I became aware of lessons for graduate students by Amy Finkelstein in 2006 where she described an article production possibility frontier with the two arguments being "interest/importance of question" and "how convincingly can you answer this question" (see econ.lse.ac.uk/staff/spischke/phds/Amy%20Finkelstein%20IAP%20talk%202006.ppt).

treatment effect (LATE) obtained from IV models, may differ markedly from the population-wide ATE. This relates to the more general issue that clean identification may yield estimates that are internally but not externally valid. Also, these methods may reveal *what* the treatment effect is without telling us *why*, which limits the usefulness of the findings without the addition of theory or structure.[4] Finally, the reliability of estimated treatment effects may be jeopardized if there are post-randomization differences in RTC's (e.g. because of differential attrition and placebo or Hawthorne effects) or because of the poor quality of many "natural experiments", whereby the treatment and control groups are actually dissimilar.[5]

My emphasis here, however, is less about potential shortcomings of experimental and quasi-experimental techniques per se – all methods have limitations – than to the possibility that excessive reliance on them leads to an overemphasis on confidence in the results obtained at the cost of failing to investigate important questions ill-suited to the use of these techniques. To some extent, this may reflect the aforementioned distinction between internal and external validity. For instance, successful IV analyses may provide an unbiased and hopefully precise measure of a local average treatment effect but only limited information on overall policy effects of broader interest. Consider Angrist and Krueger's (1991) use of birthdates as a source of exogenous variation in the ages at which compulsory schooling ends, in their IV analysis of the effects of education on earnings. The resulting LATE indicates the returns to completing approximately the 10th or 11th grade of schooling, versus not doing so, but provides little information on the gains from higher levels of education that are probably of greater current interest.[6]

But to reiterate, my concern is more wide-ranging than this. It is that the methods most well suited to answer important economic issues are likely to vary with the specific question examined and that an overemphasis on causal inference will sometime be counterproductive. Research problems are generally best addressed using heterogeneous set of research approaches, often supplementing or substituting for experimental and quasi-experimental

---

[4] This point is sometimes made to justify the "superiority" of structural modeling but the real issue is that "black-box" analyses that do not indicate mechanisms have limited applicability. Many different methods – including carefully designed but purely descriptive investigations – can sometimes help to get inside the "black-box".

[5] A related issue receiving recent attention are the potential limitations in common tests of the "parallel trends" assumption needed when estimating DD models (Jaeger, Joyce, and Robert 2018; Kahn-Lang and Lang 2018).

[6] Concerns have also been raised that the estimated LATE in this application might be biased because of a weak instruments problem (Bound, Jaeger, and Baker 1995) and since the composition of births may not randomly distributed over the calendar year (Buckles and Hungerman 2013).

techniques with non-experimental approaches. This will be true as research on given topic accumulates over time but also, in some cases, applies to individual projects. These alternative approaches may be quite varied, sometimes including highly technical methods (e.g. estimation of dynamic discrete choice structural models) but with useful information also often frequently provided using "simple" descriptive, correlational and reduced-form regression analyses of observational data.[7]

Research Questions Without Clean Identification

A charter member of the identification police will likely rank empirical methods from extremely useful to worthless (or even harmful) in the following order: randomized controlled trials, instrumental variable and regression discontinuity applications, differences-in-differences and panel data methods, propensity score and other matching estimators, regression, unconditional means and descriptive analyses.

Such a ranking has merit in some situations but is entirely misleading in others. For instance, RCT's are often viewed as the "gold-standard" but contain fundamental limitations for many questions. They poorly indicate the treatment effects of interest when general equilibrium impacts of large-scale changes differ sharply from the partial equilibrium effects of small scale interventions. They provide limited information on the long-run effects of sustained interventions because it is usually financially prohibitive to run RCTs for lengthy periods of time. And randomization is simply not possible for many critical questions. Quasi-experimental methods can at times address some of these problems. For instance, observational or administrative data will often be available for lengthy periods and at relatively low cost.[8] However, these approaches often contain other limitations, as already discussed.

To get other perspectives on these issues, I sent out the following query on social media (Facebook and Twitter) and email: "I would like to get your best examples of IMPORTANT microeconomic questions (in labor/health/public/environmental/education etc.) where clean identification is difficult or impossible to obtain." Responses included the following.[9]

---

[7] I put "simple" in quotes because I believe that the expertise and experience required to implement these methods well is often drastically underestimated and receives too little attention in the profession.

[8] Distinguishing between partial and general equilibrium effects may still be difficult but can sometimes be accomplished with ingenious quasi-experimental designs. For example, see Finkelstein's (2007) analysis of the effects of the introduction of Medicare on hospital spending.

[9] Some responses have been edited for clarity or brevity. I thank Charles Courtemanche, Arindrajit Dube, Donald Fullerton, James Harrigan, Garth Heutel, Stephen Holland, Matthew Kraft, Thomas McGuire, Maya Rossin-Slater and Diane Schanzenbach for supplying these examples.

- Effects of trade liberalization on the distribution of real wages.
- Contributions of location, preferences, local policy decisions and luck to geographic differences in morbidity and mortality rates.
- Effects of the school climate and work environment on teacher and student outcomes.
- Importance of norms on firms' wage setting.
- Extent to which economic factors explain the rise in obesity.
- Impact of family structure on child outcomes.
- Effects of inequality, child abuse and domestic violence on later life outcomes.
- Social cost of a ton of SO_2 emissions.
- Effect of race on healthcare use.
- Effect of climate change on agricultural productivity.

I concur that none of subjects these can be fully addressed using experimental approaches although, in some cases, quasi-experimental methods might provide useful information.. This suggests that the identification police have become so influential that the term "clean identification" may be synonymous with RCT's for at least some segments of our profession.

For a more concrete indication of the value and limitations of experimental and quasi-experimental approaches, consider the case of the fatal drug epidemic, which is possibly the most serious public health problem in the United States today. To provide brief background, the number of US drug deaths increased from 16,849 in 1999 to 63,632 in 2016 (Hedegaard, Warner, and Miniño 2017) and have been the leading cause of injury deaths since 2009 (Paulozzi 2012). The rise in overdose mortality is believed to have been initially fueled by enormous increases in the availability of prescription opioids, with more recent growth dominated by heroin and fentanyl.[10] However, some researchers argue that the underlying causes are economic and social decline (rather than supply-factors) that have particularly affected disadvantaged Americans (Case and Deaton 2017).

What role can different methodological approaches play in increasing our understanding of this issue? RCTs could be designed to test certain short-term interventions – such as comparing the efficacy of specific medication-assisted treatment options for drug addicts – but probably have limited broader applicability since randomization will not be practical for most potential policies and longer-term effects will be difficult to evaluate. Quasi-experimental methods have provided useful information on specific interventions such as

[10] This is partially the result of a shift to illegal drugs following the release of an abuse-resistant formulation of OxyContin (Jones, Mack, and Paulozzi 2013; Hedegaard, Warner, and Miniño 2017; Alpert, Powell, and Pacula 2018; Evans, Lieber, and Power 2018).

the effects of prescription drug monitoring programs (Dowell et al. 2016; Buchmueller and Carey 2018) and on the effects of other policies, like the legalization of medical marijuana (Chu 2015; Powell, Pacula, and Jacobson 2018; Bradford et al. 2018). However, the challenges of using these strategies should not be understated since the results often depend on precise characteristics of the policies and the timing of implementation, which may be difficult to ascertain in practice.[11] Moreover, while the estimated policy impacts are often reasonably large, they are dwarfed by the overall increase in fatal drug overdoses. Efforts to understand the root causes of the drug epidemic are therefore likely to be resistant to clean identification and instead require an "all of the above" approach using experimental and quasi-experimental methods where possible, but also the accumulation evidence from a variety of data sources and techniques, including descriptive and regression analyses that in isolation may fail to meet desired standards of causal inference but, hopefully, can be combined with other investigations to provide a compelling preponderance of evidence.

The relationship between smoking and lung cancer provides a striking example of in important question that was "answered" using strategies that would be viewed as unacceptable today by the identification police. The understanding of tobacco use as a major causal factor was not based upon RCTs involving humans but rather resulted the accretion of evidence from a wide variety of sources including: bench science, animal experiments and epidemiological evidence from non-randomized prospective and retrospective studies (Proctor 2012). Quasi-experimental evidence was eventually provided (e.g. from analyses of changes in tobacco taxes) but long after the question had been largely resolved.

To summarize, clean identification strategies will frequently be extremely useful for examining the partial equilibrium effects of specific policies or outcomes – such as the effects of reducing class sizes from 30 to 20 students or the consequences of extreme deprivation in-utero – but will often be less successful at examining the big "what if" questions related to root causes or effects of major changes in institutions or policies. This is not to say that such topics are never amenable to experimental or quasi-experimental approaches. One reason Card's (1990) Mariel boatlift paper has been so influential is because it addresses a broad policy question – the labor market effects of immigration. Another example is Manning et al.'s (1987) analysis of the RAND health insurance experiment which, among other aspects, provided important evidence on the price elasticity of demand for medical

[11] Horowitz et al. (2018) provide an excellent example and analysis of these problems in the context of estimating the effects of prescription drug monitoring programs.

care.[12] However, these examples may be the exception rather than the rule. In many cases, cleanly identified research strategies answer relatively narrow questions.

Model

Consider a simple model showing possible tradeoffs between the importance of research questions and the confidence in the findings obtained from studies examining them. Assume that a research project yields social benefit (S) according to:

$$S(I, A)$$

where I indicates importance of the question and A is accuracy of the result. With subscripts indicating partial derivatives, $S_I, S_A \geq 0$ and $S(0, A) = S(I, 0) = 0$. Thus, social benefit increases with both importance and accuracy, while research that is unimportant or completely inaccurate supplies no social benefit. In the pathological case where $A < 0$, the research has a negative social benefit.

Accuracy is generally unknown *ex ante*, but individuals have priors on it.[13] The expected social benefit of the project (E) is therefore determined by:

$$E(I, C)$$

where $C \in [0,1]$, denotes the confidence in the results. If $C = 0$ the research adds no new knowledge and if $C = 1$, the research findings are correct with certainty.[14]

The optimization problem for an individual maximizing the expected social benefit of a research project, subject to a time constraint, is:

$$\max_{I,C} E(I, C) \quad \textit{subject to} \quad t_I I + t_C C = T, \tag{1}$$

[12] Interestingly, the findings of both studies have recently been challenged. Borjas (2017) argues that Card's results are overturned when focusing on the wages of high school dropouts, who he views as the most relevant comparison group; but this conclusion has been disputed by other researchers (Peri and Yasenov 2018; Clemens and Hunt 2017). Aron-Dine et al. (2013) highlight potential nonrandom assignment to insurance plans and differential participation and reporting rates across them in creating biased results in Manning et al.'s analysis RAND health insurance experiment; they also caution against attempting to derive a single estimated demand elasticity from the complicated nonlinear health insurance contracts contained in the experiment.

[13] Importance is also not fully known *ex ante* but this source of uncertainty is not focused upon here.

[14] If $C < 0$, the research reduces understanding (e.g. deliberately distortionary research). This possibility is ruled out here.

where t_I and t_C, respectively, are the time prices of research investments in importance and confidence and with T being the total time available. First-order conditions (FOC) imply investing time such that

$$\frac{E_I}{t_I} = \frac{E_C}{t_C} \,, \tag{2}$$

which requires choosing I^* and C^* such that the marginal expected social benefits of time spent on importance and confidence are equalized. One implication is that if the time price of confidence falls (e.g. because technology makes it faster to conduct RCT's or get information on policy changes with discontinuous effects), optimal expected social benefit and confidence (E^* and C^*) increase, with uncertainty about whether I^* rises or falls. This may provide one reason why the popularity of "credible" research designs has grown in recent years. Also, if the number of research projects is endogenous, reductions time prices will increase the optimal quantity of investigations, as well as the expected social benefit of each project.[15]

The model to this point emphasizes the tradeoff between confidence and importance of the research results because time spent on one detracts from that available for the other. However, the could also be more direct tradeoffs. Assume that confidence varies systematically with the method of analysis. Define research methods using a continuous variable $M \in [0,1]$, where $M = 0$ refers to strategies yielding no confidence and $M = 1$ for those providing complete confidence in the accuracy of the result. Further assume that confidence is higher for methods with cleaner identification: thus M might be highest for well-designed RCT's, somewhat lower for quasi-experimental designs and lower still for non-experimental analyses of observational data. However, there may be a direct tradeoff if, *ceteris paribus*, methods providing more confidence refer to questions of lesser importance (e.g. because the well-estimated $LATE$ from an IV model is of less general interest than the ATE from a less well identified regression analysis covering a broader range of the treatment variable).

We can incorporate this into the expected social benefit function as:

---

[15] For instance, an individual will choose to conduct N^* projects of identical scope (i.e. the same $[I, C]$ levels) by solving the optimization problem:

$$\max_{N,I,C} N \times E(I,C) \textit{ subject to } N \times (t_I I + t_C C) = \tilde{T} \,,$$

where $\tilde{T}$ is now a global time constraint. FOC for choosing I^* and C^* for each individual project are still as shown in equation (2). In addition, a reduction in time prices relaxes the overall budget constraint allowing for higher levels of the reduced price input and in the expected social benefit of each research project, as well as an increase in the number of projects conducted.

$$E(I(M), C(M)),$$

where $C_M \geq 0$ but $I_M < 0$. The optimization problem then becomes:

$$\underset{I,C,M}{max}\, E(I(M), C(M)) \quad subject\ to \quad t_I I + t_C C = T. \tag{3}$$

In addition to the FOC already described, the researcher optimally chooses M^* to achieve:

$$-\frac{C_M}{I_M} = \frac{E_I}{E_C}. \tag{4}$$

This implies that M^* rises when: 1) the confidence gains from using robust methods of causal inference increase; 2) the loss of importance from employing these techniques declines; and 3) the marginal social benefits of confidence are, *ceteris paribus*, high relative to those for importance.[16]

Next consider the case where decision-makers (e.g. journal editors) overweight confidence according to:

$$J(I, C) = E(\alpha I, C),$$

where $0 < \alpha < 1$. A researcher attempting to maximize professional success then faces the optimization problem:

$$\underset{I,C}{max}\, E(\alpha I, C) \quad subject\ to \quad t_I I + t_C C = T. \tag{5}$$

First-order conditions imply choosing C^* and I^* such that

$$\alpha \frac{E_I}{t_I} = \frac{E_C}{t_C}. \tag{6}$$

(6) implies that confidence and importance will, respectively, exceed and fall short of social optimum. When incorporating methods directly into the optimization problem (as in equation 3), the additional FOC is:

$$-\frac{C_M}{I_M} = \alpha \frac{E_I}{E_C}. \tag{7}$$

[16] The last result occurs because diminishing returns imply that higher investments in confidence will be needed to reduce E_C to a specified level.

which implies that methods will be overly weighted towards those that provide clean identification.

Where Do We Go From Here?

My concern is that overemphasis on the use of clean identification strategies, that provide a high level of confidence in the results, may sometimes come at the cost of examining less important questions. A potential downside is that the study of critical issues may be left to other disciplines whose practitioners are typically less well-trained than economists to think in terms of causal inference.

Have the identification police become too powerful? The answer to this question is subjective and open to debate. However, I believe that it is becoming increasingly difficult to publish research on significant questions that lack sufficiently clean identification and, conversely, that research using quasi-experimental and (particularly) experimental strategies yielding high confidence but on questions of limited importance are more often being published. In talking with PhD students, I hear about training that emphasizes the search for discontinuities and policy variations, rather than on seeking to answer questions of fundamental importance. At professional presentations, experienced economists sometimes mention "correlational" or "reduced-form" approaches with disdain, suggesting that such research has nothing to add to the cannon of applied economics.

I do not claim that these concerns are universal. For instance, the top general journals certainly emphasize the need for important contributions and some outlets focus on overarching questions addressed using a multiplicity of approaches that are heavy in data description and breadth of analysis, but often using weaker causal designs. However, the individuals publishing these pieces often are well-established and, I suspect, would have been ill-advised do so earlier in their careers. Given the current incentives, I would certainly caution current graduate students or newly minted Ph.D.'s against conducting such research, since it would place at risk their own career advancement. However, it seems worth considering whether it is desirable for the next generation of scholars to be discouraged from attempting to answer these crucial but less well-identified questions.

If there is agreement that we are currently sacrificing importance by worshipping at the altar of confidence, what should be done? I have no solutions but conclude by suggesting some possibilities.

The first is to explicitly emphasize the centrality of the questions we attempt to answer, rather than the tools used to do so. Causal inference is extremely useful but does not define

applied microeconomic analysis. Obviously, there will be heterogeneity in the weights placed on importance versus confidence as arguments in the research social welfare function and it would be useful to have discussion and debate on these issues in academic journals and through panels at professional meetings or in other venues.[17]

A second step is to increase the incentives for research that addresses substantively significant topics. For instance, journal editors could further emphasize that importance of the contribution, rather than confidence in the findings alone, will be a key component of acceptance decisions. Again, there will be variation in the importance-confidence tradeoff across academic outlets and editors could clearly state the social welfare function used by their journal. Many journals already ask referees discuss the contribution made by the manuscript but these rankings could be made more comprehensive and explicit. Similarly, paper submitters could be strongly encouraged to emphasize the significance of the research question and findings. One potentially useful change would be to urge researchers to shorten or eliminate the "literature review" at the beginning of the paper but (as is routine in some other disciplines) to provide a much more detailed conclusion section that, rather than simply repeating the results, places them in the context of prior findings and clearly indicates the importance of the new knowledge obtained.

Third, the central role of questions, rather than tools, could receive increased emphasis in economics graduate programs, although this will be difficult to achieve without corresponding changes in professional research incentives. One suggestion would be to include at least one or two outstanding papers using purely descriptive methods (not even regressions!) as required readings in graduate field courses.

Should we shackle the identification police? Absolutely not! But nor should we ignore our profession's primary responsibility to address substantively important questions, even when clean identification is elusive.

[17] These discussions are already occurring although, to my view, they have often been somewhat sterile addressing, for example, the superiority of structural versus reduced-form methods.

References

Alpert, Abby, David Powell, and Rosalie Liccardo Pacula. 2018. "Supply-Side Drug Policy in the Presence of Substitutes: Evidence from the Introduction of Abuse-Deterrent Opioids." *American Economic Journal: Economic Policy* 10 (4): 1–35. doi:10.3386/w23031.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106 (4): 979–1014.

Angrist, Joshua D, and Jörn-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. doi:10.1257/jep.24.2.3.

Aron-Dine, Aviva, Liran Einav, and Amy Finkelstein. 2013. "The RAND Health Insurance Experiment, Three Decades Later." *Journal of Economic Perspectives* 27 (1): 197–222. doi:10.1257/jep.27.1.197.

Borjas, George J. 2017. "The Wage Impact of the Marielitos: A Reappraisal." *ILR Review* 70 (5): 1077–1110. doi:10.1177/0019793917692945.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90 (430): 443–50. doi:10.1080/01621459.1995.10476536.

Bradford, Ashley C., W. David Bradford, Amanda Abraham, and Grace Bagwell Adams. 2018. "Association Between US State Medical Cannabis Laws and Opioid Prescribing in the Medicare Part D Population." *JAMA Internal Medicine* 178 (5): 667–72. doi:10.1001/jamainternmed.2018.0266.

Buchmueller, Thomas C, and Colleen Carey. 2018. "The Effect of Prescription Drug Monitoring Programs on Opioid Utilization in Medicare." *American Economic Journal: Economic Policy*. doi:10.3386/w23148.

Buckles, Kasey S., and Daniel M. Hungerman. 2013. "Season of Birth and Later Outcomes: Old Questions, New Answers." *Review of Economics and Statistics* 95 (3): 711–24. doi:10.1162/REST_a_00314.

Cameron, Drew B., Anjini Mishra, and Annette N. Brown. 2016. "The Growth of Impact Evaluation for International Development: How Much Have We Learned?" *Journal of Development Effectiveness* 8 (1): 1–21. doi:10.1080/19439342.2015.1034156.

Card, David. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market."

Industrial and Labor Relations Review 43 (2): 245. doi:10.2307/2523702.

Case, Anne, and Angus Deaton. 2017. "Mortality and Morbidity in the 21st Century." *Brookings Papers on Economic Activity*.

Chu, Yu-Wei Luke. 2015. "Do Medical Marijuana Laws Increase Hard-Drug Use?" *Journal of Law and Economics* 58 (2): 481–517. doi:10.1086/684043.

Clemens, Michael A., and Jennifer Hunt. 2017. "The Labor Market Effects of Refugee Waves: Reconciling Conflicting Results." 23433. National Bureau of Economic Research Working Paper No. 23433.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science and Medicine*. doi:10.1016/j.socscimed.2017.12.005.

Dowell, Deborah, Kun Zhang, Rita K. Noonan, and Jason M. Hockenberry. 2016. "Mandatory Provider Review and Pain Clinic Laws Reduce the Amounts of Opioids Prescribed and Overdose Death Rates." *Health Affairs* 35 (10): 1876–83. doi:10.1377/hlthaff.2016.0448.

Evans, William N, Ethan Lieber, and Patrick Power. 2018. "How the Reformulation of OxyContin Ignited the Heroin Epidemic." *Review of Economics and Statistics*, no. Forthcoming.

Finkelstein, Amy. 2007. "The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare." *Quarterly Journal of Economics* 122 (1): 1–37. doi:10.1162/qjec.122.1.1.

Heckman, James J, and Jeffrey A Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9 (2): 85–110. doi:10.1257/jep.9.2.85.

Hedegaard, Holly, Margaret Warner, and Arialdi M Miniño. 2017. "Drug Overdose Deaths in the United States, 1999-2016." *NCHS Data Brief*, no. 294.

Horowitz, Jill, Corey S. Davis, Lynn S. McClellan, Rebecca S. Fordon, and Ellen Meara. 2018. "The Problem of Data Quality in Analyses of Opioid Regulation: The Case of Prescription Drug Monitoring Programs." National Bureau of Economic Research Working Paper No. 24947.

Jaeger, David A., Theodore Joyce, and Kaestner Robert. 2018. "Does Reality TV Induce Real Effects? On the Questionable Association Between 16 and Pregnant and Teenage Childbearing." *Journal of Business & Economic Statistics* forthcoming.

Jones, Christopher M, Karin A Mack, and Leonard J Paulozzi. 2013. "Pharmaceutical Overdose Deaths, United States, 2010." *JAMA* 309 (7): 657–59.

doi:10.1001/jama.2013.272.

Kahn-Lang, Ariella, and Kevin Lang. 2018. “The Promise and Pitfalls of Differences-In-Differences: Reflections on ‘16 and Pregnant’ and Other Applications.” National Bureau of Economic Research Working Paper No. 24857.

Keane, Michael P. 2010. “Structural vs. Atheoretic Approaches to Econometrics.” *Journal of Econometrics* 156 (1): 3–20. doi:10.1016/j.jeconom.2009.09.003.

Kleven, Henrik J. 2018. “Language Trends in Public Economics.” Princeton, NJ.

Leamer, Edward E. 1983. “Let’s Take the Con out of Econometrics.” *American Economic Review* 73 (1): 31–43. doi:10.2307/1803924.

Manning, W G, J P Newhouse, N Duan, E B Keeler, a Leibowitz, and M S Marquis. 1987. “Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment.” *The American Economic Review* 77 (3): 251–77. doi:http://www.aeaweb.org/aer/.

Panhans, Matthew T, and John D Singleton. 2017. “The Empirical Economist’s Toolkit: From Models to Methods.” *History of Political Economy* 49 (Supplement). Duke University Press: 127–57.

Paulozzi, Leonard J. 2012. “Prescription Drug Overdoses: A Review.” *Journal of Safety Research*. doi:10.1016/j.jsr.2012.08.009.

Peri, Giovanni, and Vasil Yasenov. 2018. “The Labor Market Effects of a Refugee Wave: Synthetic Control Method Meets the Mariel Boatlift.” *Journal of Human Resources*, 0217_8561R1. doi:10.3368/jhr.54.2.0217.8561R1.

Powell, David, Rosalie Liccardo Pacula, and Mireille Jacobson. 2018. “Do Medical Marijuana Laws Reduce Addictions and Deaths Related to Pain Killers?” *Journal of Health Economics* 58: 29–42. doi:10.1016/j.jhealeco.2017.12.007.

Proctor, Robert N. 2012. “The History of the Discovery of the Cigaretteelung Cancer Link: Evidentiary Traditions, Corporate Denial, Global Toll.” *Tobacco Control* 21 (2): 87–91. doi:10.1136/tobaccocontrol-2011-050338.