NBER WORKING PAPER SERIES

THE CREDIBILITY REVOLUTION IN EMPIRICAL ECONOMICS: HOW BETTER RESEARCH DESIGN IS TAKING THE CON OUT OF ECONOMETRICS

Joshua Angrist
Jörn-Steffen Pischke

Working Paper 15794
http://www.nber.org/papers/w15794

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2010

We thank Guido Imbens for suggesting this topic and for feedback; Daron Acemoglu, Olivier Blanchard, John Donohue, Isaac Ehrlich, Glenn Ellison, Jeff Grogger, Radha Iyengar, Larry Katz, Alan Krueger, Ethan Ilzetzki, Guido Lorenzoni, Albert Marcet, Aviv Nevo, Alan Manning, Bruce Meyer, Parag Pathak, Gary Solon, and Justin Wolfers for helpful comments and discussions; and the JEP editors, David Autor, James Hines, Charles Jones, and Timothy Taylor for comments on earlier drafts. Remaining errors and omissions are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics
Joshua Angrist and Jörn-Steffen Pischke
NBER Working Paper No. 15794
March 2010
JEL No. C20,C30,C50,C51,D00

ABSTRACT

This essay reviews progress in empirical economics since Leamer's (1983) critique. Leamer highlighted the benefits of sensitivity analysis, a procedure in which researchers show how their results change with changes in specification or functional form. Sensitivity analysis has had a salutary but not a revolutionary effect on econometric practice. As we see it, the credibility revolution in empirical work can be traced to the rise of a design-based approach that emphasizes the identification of causal effects. Design-based studies typically feature either real or natural experiments and are distinguished by their prima facie credibility and by the attention investigators devote to making the case for a causal interpretation of the findings their designs generate. Design-based studies are most often found in the microeconomic fields of Development, Education, Environment, Labor, Health, and Public Finance, but are still rare in Industrial Organization and Macroeconomics. We explain why IO and Macro would do well to embrace a design-based approach. Finally, we respond to the charge that the design-based revolution has overreached.

Joshua Angrist
Department of Economics
MIT, E52-353
50 Memorial Drive
Cambridge, MA 02142-1347
and NBER
angrist@mit.edu

Jörn-Steffen Pischke
CEP
London School of Economics
Houghton Street
London WC2A 2AE
UK
and NBER
s.pischke@lse.ac.uk

Just over a quarter century ago, Edward Leamer (1983) reflected on the state of empirical work in economics. He urged empirical researchers to "take the con out of econometrics" and memorably observed (p. 37): "Hardly anyone takes data analysis seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analysis seriously." Leamer was not alone; Hendry (1980), Sims (1980), and others writing at about the same time were similarly disparaging of empirical practice. Reading these commentaries, we wondered as late-1980s Ph.D. students about the prospects for a satisfying career doing applied work. Perhaps credible empirical work in economics is a pipe dream. Here we address the questions of whether the quality and the credibility of empirical work have increased since Leamer's pessimistic assessment. Our views are necessarily colored by the areas of applied microeconomics in which we are active, but we look over the fence at other areas as well.

Leamer (1983) diagnosed his contemporaries' empirical work as suffering from a distressing lack of robustness to changes in key assumptions—assumptions he called "whimsical" because one seemed as good as another. The remedy he proposed was sensitivity analysis, in which researchers show how their results vary with changes in specification or functional form. Leamer's critique had a refreshing emperor's-new-clothes earthiness that we savored on first reading and still enjoy today. But we're happy to report that Leamer's complaint that "hardly anyone takes anyone else's data analysis seriously" no longer seems justified. Empirical microeconomics has experienced a credibility revolution, with a consequent increase in policy relevance and scientific impact. Sensitivity analysis played a role in this, but as we see it, the primary engine driving improvement has been a focus on the quality of empirical research designs. This emphasis on research design is in the spirit of Leamer's critique, but it did not feature in his remedy.

The advantages of a good research design are perhaps most easily apparent in research using random assignment, which not coincidentally includes some of the most influential microeconometric studies to appear in recent years. For example, in a pioneering effort to improve child welfare, the Progresa program in Mexico offered cash transfers to randomly selected mothers, contingent on participation in prenatal care, nutritional monitoring of children, and the children's regular school attendance (Gertler, 2004 and Schultz, 2004, present some of the main findings). In the words of Paul Gertler, one of the original investigators (quoted in Ayres, 2007, p. 86), "Progresa is why now thirty countries worldwide have conditional cash transfer programs." Progresa is emblematic of a wave of random assignment policy evaluations sweeping development economics (Duflo and Kremer, 2008, provide an overview).

Closer to home, the Moving to Opportunity program, carried out by the U.S. Department of Housing and Urban Development, randomly selected low-income families in Baltimore, Boston, Chicago, Los Angeles, and New York City to be offered housing vouchers specifically limited to low-poverty areas (Kling, Liebman, and Katz, 2007). The program has produced surprising and influential evidence weighing against the view that neighborhood effects are a primary determinant of low earnings by the residents of poor neighborhoods.

Structural econometric parameters, such as the intertemporal substitution elasticity (a labor supply elasticity that measures the response to transitory wage changes), have also been the focus of randomized experiments. For example, Fehr and Goette (2007) randomized the pay of bicycle messengers, offering one group and then another a temporarily higher wage. This cleverly designed study shows how wages affect labor supply in an environment where lifetime wealth is unchanged. The result is dramatic and convincing: holding wealth constant, workers shift hours into high-wage periods, with an implied intertemporal substitution elasticity of about unity.

Such studies offer a powerful method for deriving results that are defensible both in the seminar room and in a legislative hearing. But experiments are time consuming, expensive, and may not always be practical. It's difficult to imagine a randomized trial to evaluate the effect of immigrants on the economy of the host country. However, human institutions or the forces of nature can step into the breach with informative natural or quasi-experiments. For example, in an influential paper, Card (1990a) used the Mariel boatlift from Cuba to Florida, when Cuban émigré's increased Miami's labor force by about 7 percent in a period of three months, as a natural experiment to study immigration. More recently, paralleling the Moving to Opportunity experimental research agenda, Jacob (2004) studied the causal effects of public housing on housing project residents by exploiting the fact that public housing demolition in Chicago was scheduled in a manner unrelated to the characteristics of the projects and their residents.

Like the results from randomized trials, quasi-experimental findings have filtered quickly into policy discussions and become part of a constructive give-and-take between the real world and the ivory tower, at least when it comes to applied microeconomics. Progress has been slower in empirical macro, but a smattering of design-based empirical work appears to be generating a limited though useful consensus on key concerns, such as the causal effect of monetary policy on inflation and output. Encouragingly, the recent financial crisis has spurred an effort to produce credible evidence on questions related to banking. Across most fields (although industrial organization appears to be an exception, as we discuss later), applied economists are now less likely to pin a causal interpretation of the results on

econometric methodology alone. Design-based studies are distinguished by their prima facie credibility and by the attention investigators devote to making both an institutional and a data-driven case for causality.

Accounting for the origins of the credibility revolution in empirical economics is like trying to chart the birth of rock and roll. Early influences are many, and every fan has a story. But from the trenches of empirical labor economics, we see an important impetus for better designs and more randomized trials coming from studies questioning the reliability of econometric evaluations of subsidized government training programs. A landmark here is Lalonde (1986), who compared the results from an econometric evaluation of the National Supported Work demonstration with those from a randomized trial. The econometric results typically differed quite a bit from those using random assignment. Lalonde argued that there is little reason to believe that statistical comparisons of alternative models (specification testing) would point a researcher in the right direction. Two observational studies of training effects foreshadowed the Lalonde results: Ashenfelter (1978) and Ashenfelter and Card (1985), using longitudinal data to evaluate federal training programs without the benefit of a quasi-experimental research design, found it difficult to construct specification-robust estimates. Ashenfelter (1987) concluded that randomized trials are the way to go.

Younger empiricists also began to turn increasingly to quasi-experimental designs, often exploiting variation across U.S. states to get at causal relationships in the fields of labor and public finance. An early example of work in this spirit is Solon (1985), who estimated the effects of unemployment insurance on the duration of unemployment spells by comparing the change in job-finding rates in states that had recently tightened eligibility criteria for unemployment insurance, to the change in rates in states that had not changed their rules. Gruber's (1994) influential study of the incidence of state-mandated maternity benefits applies a similar idea to a public finance question. Angrist (1990) and Angrist and Krueger (1991) illustrated the value of instrumental variables identification strategies in studies of the effects of Vietnam-era military service and schooling on earnings. Meyer's (1995) methodological survey made many applied microeconomists aware of the quasi-experimental tradition embodied in venerable texts on social science research methods by Campbell and Stanley (1963) and Cook and Campbell (1979). These texts, which emphasize research design and threats to validity, were well known in some disciplines, but distinctly outside the econometric canon.[1]

[1] Many of the applied studies mentioned here have been the subjects of critical re-examinations. This back and forth has mostly been constructive. For example, in an influential paper that generated wide-ranging methodological work, Bound, Jaeger, and Baker (1995) argue that the use of many weak instrumental variables

In this essay, we argue that a clear-eyed focus on research design is at the heart of the credibility revolution in empirical economics. We begin with an overview of Leamer's (1983) critique and his suggested remedies, based on concrete examples of that time. We then turn to the key factors we see contributing to improved empirical work, including the availability of more and better data, along with advances in theoretical econometric understanding, but especially the fact that research design has moved front and center in much of empirical micro. We offer a brief digression into macroeconomics and industrial organization, where progress—by our lights—is less dramatic, although there is work in both fields that we find encouraging. Finally, we discuss the view that the design pendulum has swung too far. Critics of design-driven studies argue that in pursuit of clean and credible research designs, researchers seek good answers instead of good questions. We briefly respond to this concern, which worries us little.

The Leamer Critique and His Proposed Remedies

Naive Regressions and Extreme Bounds Analysis

Leamer (1983) presented randomized trials—a randomized evaluation of fertilizer, to be specific—as an ideal research design. He also argued that randomized experiments differ only in degree from nonexperimental evaluations of causal effects, the difference being the extent to which we can be confident that the causal variable of interest is independent of confounding factors. We couldn't agree more. However, Leamer went on to suggest that the best way to use nonexperimental data to get closer to the experimental ideal is to explore the fragility of nonexperimental estimates. Leamer did not advocate *doing* randomized trials or, for that matter, looking for credible natural experiments.

The chief target of Leamer's (1983) essay was naive regression analysis. In fact, none of the central figures in the Leamer-inspired debate had much to say about research design. Rather, these authors (like McAleer, Pagan, and Volker, 1985, and Cooley and LeRoy, 1986, among others) appear to have accepted the boundaries of established econometric practice, perhaps because they were primarily interested in addressing traditional macroeconomic questions using time series data.

After making the tacit assumption that useful experiments are an unattainable ideal, Leamer (1983, but see also 1978, 1985) proposed that the whimsical nature of key assumptions in regression

---

biases some of the estimates reported in Angrist and Krueger (1991). For a recent discussion of weak instruments problems, see our book (Angrist and Pischke, 2009).

analysis be confronted head-on through a process of sensitivity analysis. Sims (1988) threw his weight behind this idea as well. The general heading of sensitivity analysis features an explicitly Bayesian agenda. Recognizing the severe demands of Bayesian orthodoxy, such as a formal specification of priors and their incorporation into an elaborate multivariate framework, Leamer also argued for a more *ad hoc* but intuitive approach called "extreme bounds analysis." In a nutshell, extreme bounds analysis amounts to the estimation of regressions with many different sets of covariates included as controls; practitioners of this approach are meant to report a range of estimates for the target parameter.

The Deterrent Effect of Capital Punishment

We sympathize with Leamer's (1983) view that much of the applied econometrics of the 1970s and early 1980s lacked credibility. To make his point, and to illustrate the value of extreme bounds analysis, Leamer picked an inquiry into whether capital punishment deters murder. This question had been analyzed in a series of influential papers by Isaac Ehrlich, one exploiting time series variation (Ehrlich, 1975a) and one using cross sections of states (Ehrlich, 1977b). Ehrlich concluded that the death penalty had a substantial deterrent effect. Leamer (1983) did not try to replicate Ehrlich's work, but reported on an independent time-series investigation of the deterrence hypothesis using extreme bounds analysis, forcefully arguing that the evidence for deterrence is fragile at best (although Ehrlich and Liu, 1999, disputed this).

It's hard to exaggerate the attention this topic commanded at the time. The U.S. Supreme Court decision in *Furman v. Georgia* (408 U.S. 153 [1972]) had created a de facto moratorium on the death penalty. This moratorium lasted until *Gregg v. Georgia* (428 U.S. 153 [1976]), at which time the high court decided that the death penalty might be allowable if capital trials were bifurcated into separate guilt–innocence and sentencing phases. Gary Gilmore was executed not long after, in January 1977. Part of the intellectual case for restoration of capital punishment was the deterrent effect (against a backdrop of high and increasing homicide rates at that time). Indeed, the U.S. Supreme Court cited Ehrlich's (1975a) paper in its *Gregg v. Georgia* decision reinstating capital punishment.

Ehrlich's work was harshly criticized by a number of contemporaries in addition to Leamer, most immediately Bowers and Pierce (1975) and Passell and Taylor (1977). Ehrlich's results appeared to be sensitive to changes in functional form, inclusion of additional controls, and especially to changes in sample. Specifically, his finding of a significant deterrent effect seemed to depend on observations from the 1960s. The critics argued that the increase in murder rates in the 1960s may have been driven by factors other than the sharp decline in the number of executions during this period. Ehrlich (1975b,

1977a) disputed the critics' claims about functional form and argued that the 1960s provided useful variation in executions that should be retained.

Ehrlich's contemporaneous critics failed to hit on what we think of as the most obvious flaw in Ehrlich's analysis. Like other researchers studying deterrent effects, Ehrlich recognized that the level of the murder rate might affect the number of executions as well as vice versa and that his results might be biased by omitted variables (especially variables with a strong trend). Ehrlich sought to address problems of reverse causality and omitted variables bias by using instrumental variables in a two-stage least squares procedure. He treated the probabilities of arrest, conviction, and execution as endogenous in a simultaneous-equations set-up. His instrumental variables were lagged expenditures on policing, total government expenditure, population, and the fraction of the population nonwhite. But Ehrlich did not explain why these are good instruments, or even how and why these variables are correlated with the right-hand-side endogenous variables.[2]

Ehrlich's work on capital punishment seems typical of applied work in the period about which Leamer (1983) was writing. Most studies of this time used fairly short time series samples with strong trends common to both dependent and independent variables. The use of panel data to control for year and fixed effects—even panels of U.S. states—was still rare. The use of instrumental variables to uncover causal relationships was typically mechanical, with little discussion of why the instruments affected the endogenous variables of interest or why they constitute a "good experiment." In fact, Ehrlich was ahead of many of his contemporaries in that he recognized the need for something other than naive regression analysis. In our view, the main problem with Ehrlich's work was the lack of a credible research design. Specifically, he failed to isolate a source of variation in execution rates that is likely to reveal causal effects on homicide rates.

The Education Production Function

Other examples of poor research design from this time period come from the literature on education production. This literature (surveyed in Hanushek, 1986) is concerned with the causal effect of school inputs, such as class size or per-pupil expenditure, on student achievement. The systematic quantitative study of school inputs was born with the report by Coleman et al. (1966), which (among other things) used regression techniques to look at the proportion of variation in student outputs that can be accounted for in an R^2 sense by variation in school inputs. Surprisingly to many at the time, the

[2] Ehrlich's (1977b) follow-up cross-state analysis did not use two-stage least squares. In later work, Ehrlich (1987, 1996) discussed his choice of instruments and the associated identification problems at greater length.

Coleman report found only a weak association between school inputs and achievement. Many subsequent regression-based studies replicated this finding.

The Coleman Report was one of the first investigations of education production in a large representative sample. It is also distinguished by sensitivity analysis, in that it discusses results from many specifications (with and without controls for family background, for example). The problem with the Coleman report and many of the studies in this mold that followed is that they failed to separate variation in inputs from confounding variation in student, school, or community characteristics. For example, a common finding in the literature on education production is that children in smaller classes tend to do worse on standardized tests, even after controlling for demographic variables. This apparently perverse finding seems likely to be at least partly due to the fact that struggling children are often grouped into smaller classes. Likewise, the relationship between school spending and achievement is confounded by the fact that spending is often highest in a mixture of wealthy districts and large urban districts with struggling minority students. In short, these regressions suffer from problems of reverse causality and omitted variables bias.

Many education production studies from this period also ignored the fact that inputs like class size and per-pupil expenditure are inherently linked. Because smaller classes cannot be had without spending more on teachers, it makes little sense to treat total expenditure (including teacher salaries) as a control variable when estimating the causal effect of class size (a point noted by Krueger, 2003). Finally, the fact that early authors in the education production literature explored many alternative models was not necessarily a plus. In what was arguably one of the better studies of the period, Summers and Wolfe (1977) report only the final results of an exhaustive specification search in their evaluation of the effect of school resources on achievement. To their credit, Summers and Wolfe describe the algorithm that produced the results they chose to report, and forthrightly caution (p. 642) that "the data have been mined, of course." As we see it, however, the main problem with this literature is not data mining, but rather the weak foundation for a causal interpretation of whatever specification authors might have favored.

Other Empirical Work in the Age of Heavy Metal

The 1970s and early 1980s saw rapid growth in mainframe computer size and power. Stata had yet to appear, but magnetic tape jockeys managed to crunch more and more numbers in increasingly elaborate ways. For the most part, however, increased computing power did not produce more credible estimates. For example, the use of randomized trials and quasi-experiments to study education production was rare until fairly recently (a history traced in Angrist, 2004). Other areas of social science saw isolated though ambitious efforts to get at key economic relationships using random assignment. A bright spot

was the RAND Health Insurance Experiment, initiated in 1974 (Manning, Newhouse, Duan, Keeler, and Leibowitz, 1987). This experiment looked at the effects of deductibles and copayments on health care usage and outcomes. Unfortunately, many of the most ambitious (and expensive) social experiments were seriously flawed: the Seattle/Denver and Gary Income Maintenance Experiments, in which the government compared income-support plans modeled on Milton Friedman's idea of a negative income tax, were compromised by sample attrition and systematic income misreporting (Ashenfelter and Plant, 1990; Greenberg and Halsey, 1983). This fact supports Leamer's (1983) contention that the difference between a randomized trial and an observational study is one of degree. Indeed, we would be the first to admit that a well-done observational study can be more credible and persuasive than a poorly executed randomized trial.

There was also much to complain about in empirical macroeconomics. An especially articulate complaint came from Sims (1980), who pointed out that macroeconomic models of that time, typically a system of simultaneous equations, invoked identification assumptions (the division of variables into those that are jointly determined and exogenous) that were hard to swallow and poorly defended. As an alternative to the simultaneous equations framework, Sims suggested the use of unrestricted vector autoregressions (VARs) to describe the relation between a given set of endogenous variables and their lags. But Sims's complaint did not generate the same kind of response that grew out of concerns about econometric program evaluation in the 1980s among labor economists. Macroeconomists circled their wagons but did not mobilize an identification posse.

Sims's argument came on the heels of a closely related and similarly influential stab at the heart of empirical macro known as the Lucas critique. Lucas (1976) and Kydland and Prescott (1977) argued via theoretical examples that in a world with forward-looking optimizing agents, *nothing* can be learned from past policy changes. Lucas held out the hope that we might instead try to recover the empirical response to changes in policy rules by estimating the structural parameters that lie at the root of economic behavior, such as those related to technology or preferences (Lucas saw these parameters as stable or at least policy invariant). But Kydland and Prescott—invoking Lucas—appeared willing to give up entirely on conventional empirical work (1977, p. 487): "If we are not to attempt to select policy optimally, how should it be selected? Our answer is, as Lucas (1976) proposed, that economic theory be used to evaluate alternative policy rules and that one with good operating characteristics be selected." This view helped to lay the intellectual foundations for a sharp turn toward theory in macro, though often informed by numbers via "calibration."

Our overview of empirical work in the Leamer era focuses on shortcomings. But we should also note that the best applied work from the 1970s and early 1980s still holds up today. A well-cited example is Feldstein and Horioka (1980), which argues that the strong link between domestic savings and investment weighs against the notion of substantial international capital mobility. The Feldstein and Horioka study presents simple evidence in favor of a link between domestic savings and investment, discusses important sources of omitted variables bias and simultaneity bias in these estimates, and tries to address these concerns. Obstfeld's (1995) extensive investigation of the Feldstein and Horioka (1980) framework essentially replicates their findings for a later and longer period.

Why There's Less Con in Econometrics Today

Improvements in empirical work have come from many directions. Better data and more robust estimation methods are part of the story, as is a reduced emphasis on econometric considerations that are not central to a causal interpretation of the main findings. But the primary force driving the credibility revolution has been a vigorous push for better and more clearly articulated research designs.

Better and More Data

Not unusually for the period, Ehrlich (1975a) analyzed a time series of 35 annual observations. In contrast, Donohue and Wolfers (2005) investigate the capital punishment question using a panel of U.S. states from 1934 to 2000, with many more years and richer within-state variation due to the panel structure of the data. Better data often engenders a fresh approach to long-standing research questions. Grogger's (1990) investigation of the deterrent effect of executions on daily homicide rates, inspired by sociologist Phillips (1980), is an example.[3] Farther afield, improvements have come from a rapidly expanding reservoir of micro data in many countries. The use of administrative records has also grown.

Fewer Distractions

Bower's and Pierce (1975) devoted considerable attention to Ehrlich's (1975a) use of the log transformation, as well as to his choice of sample period. Passell and Taylor (1977) noted the potential for omitted variables bias, but worried as much about F-tests for temporal homogeneity and logs. The methodological appendix to Ehrlich's (1977b) follow-up paper discusses the possibility of using a Box–

[3] The decline in the use of time series and the increase in the use of panel data and researcher-originated data are documented for the field of labor economics in table 1 of Angrist and Krueger (1999).

Cox transformation to implement a flexible functional form, tests for heteroskedasticity, and uses generalized least squares. Ehrlich's (1975b) reply to Bowers and Pierce focused on the statistical significance of trend terms in samples of different lengths, differences in computational procedures related to serial correlation, and evidence for robustness to the use of logs. Ehrlich's (1977a) reply to Passell covers the sample period and logs, though he also reports some of his (1977b) cross-state estimates. Ehrlich's rejoinders devoted little attention to the core issue of whether the sources of variation in execution used by his statistical models justify a causal interpretation of his estimates, but Ehrlich's contemporaneous critics did not hit this nail on the head either. Even were the results insensitive to the sample, the same in logs and levels, and the residuals independent and identically distributed, we would remain unsatisfied. In the give and take that followed Ehrlich's original paper, the question of instrument validity rarely surfaced, while the question of omitted variables bias took a back seat to concerns about sample break points and functional form.[4]

As in the exchange over capital punishment, others writing at about the same time often seemed distracted by concerns related to functional form and generalized least squares. Today's applied economists have the benefit of a less dogmatic understanding of regression analysis. Specifically, an emerging grasp of the sense in which regression and two-stage least squares produce average effects even when the underlying relationship is heterogeneous and/or nonlinear has made functional form concerns less central. The linear models that constitute the workhorse of contemporary empirical practice usually turn out to be remarkably robust, a feature many applied researchers have long sensed and that econometric theory now does a better job of explaining.[5] Robust standard errors, automated clustering, and larger samples have also taken the steam out of issues like heteroskedasticity and serial correlation. A legacy of White's (1980a) paper on robust standard errors, one of the most highly cited from the period, is the near death of generalized least squares in cross-sectional applied work. In the interests of replicability, and to reduce the scope for errors, modern applied researchers often prefer simpler estimators though they might be giving up asymptotic efficiency.

---

[4] Hoenack and Weiler's (1980) critical re-examination of Ehrlich (1975a) centered on identification problems, but the alternative exclusion restrictions Hoenack and Weiler proposed were offered without much justification and seem just as hard to swallow as Ehrlich's (for example, the proportion nonwhite is used as an instrument).

[5] For this view of regression, see, for example, White (1980b), Chamberlain's (1984) chapter in the *Handbook of Econometrics*, Goldberger's (1991) econometrics text, or our book (Angrist and Pischke, 2009) for a recent take. Angrist and Imbens (1995) show how conventional two-stage least squares estimates can be interpreted as an average causal effect in models with nonlinear and heterogeneous causal effects.

Better Research Design

Leamer (1983) led his essay with the idea that experiments—specifically, randomized trials—provide a benchmark for applied econometrics. He was not alone among econometric thought leaders of the period in this view. Here is Zvi Griliches (1986, p. 1466) at the beginning of a chapter on data in *The Handbook of Econometrics*: "If the data were perfect, collected from well-designed randomized experiments, there would hardly be room for a separate field of econometrics." Since then, empirical researchers in economics have increasingly looked to the ideal of a randomized experiment to justify causal inference. In applied micro fields such as development, education, environmental economics, health, labor, and public finance, researchers seek real experiments where feasible, and useful natural experiments if real experiments seem (at least for a time) infeasible. In either case, a hallmark of contemporary applied microeconometrics is a conceptual framework that highlights specific sources of variation. These studies can be said to be *design based* in that they give the research design underlying any sort of study the attention it would command in a real experiment.

The econometric methods that feature most prominently in quasi-experimental studies are instrumental variables, regression discontinuity methods, and differences-in-differences-style policy analysis. These econometric methods are not new, but their use has grown and become more self-conscious and sophisticated since the 1970s. When using instrumental variables, for example, it's no longer enough to mechanically invoke a simultaneous equations framework, labeling some variables endogenous and others exogenous, without substantially justifying the exclusion restrictions and as-good-as-randomly-assigned assumptions that make instruments valid. The best of today's design-based studies make a strong institutional case, backed up with empirical evidence, for the variation thought to generate a useful natural experiment.

The Card and Krueger (1992a, b) school quality studies illustrate this and arguably mark a turning point in the literature on education production. The most important problem in studies of school quality is omitted variables bias. On one hand, students who attend better-resourced schools often end up in those schools by virtue of their ability or family background; on the other, weaker students may receive disproportionately more inputs (say, smaller classes). Card and Krueger addressed this problem by focusing on variation in resources at the state-of-birth-by-cohort level, which they link to the economic returns to education estimated at the same level. For example, they used Census data to compare the returns to education for residents of Northern states educated in the North with the returns to education for residents of Northern states educated in more poorly resourced Southern schools**.**

The Card and Krueger papers show that the economic returns to schooling are higher for those from states and cohorts with more resources (controlling for cohort and state fixed effects and for state of residence). They implicitly use state-level variation in education spending as a natural experiment: aggregation of individual data up to the cohort/state level is an instrumental variables procedure where the instruments are state-of-birth and cohort dummy variables. (In Angrist and Pischke, 2009, we explain why aggregation in this way works as an instrumental variable.) State-by-cohort variation in the returns to schooling is unlikely to be driven by selection or sorting, because individuals do not control these variables. State-by-cohort variation in school resources also appears unrelated to omitted factors such as family background. Finally, Card and Krueger took advantage of the fact that school resources increased dramatically in the South when the Southerners in their sample were school age. The Card and Krueger school quality studies are not bullet proof (Heckman, Layne-Farrar, and Todd, 1996 offer a critique), but their findings on class size (the strongest set of results in Card and Krueger, 1992a) have been replicated in other studies with good research designs.

Angrist and Lavy (1999) illustrate the regression discontinuity research design in a study of the effects of class size on achievement. The regression discontinuity approach can be used when people are divided into groups based on a certain cutoff score, with those just above or just below the cutoff suddenly becoming eligible for a different treatment. The Angrist–Lavy research design is driven by the fact that class size in Israel is capped at 40, so a cohort of 41 is usually split into two small classes, while a cohort of 39 is typically left in a single large class. This leads to a series of notional experiments: comparisons of schools with enrollments just above and below 40, 80, or 120, in which class sizes vary considerably. In this setting, schools with different numbers of students may be quite similar in other characteristics. Thus, as school enrollment increases, a regression capturing the relationship between number of students and academic achievement should show discontinuities at these break points. The Angrist–Lavy design is a version of what is known as the "fuzzy" regression discontinuity design, in which the fuzziness comes from the fact that class size is not a deterministic function of the kinks or discontinuities in the enrollment function. Regression discontinuity estimates using Israeli data show a marked increase in achievement when class size falls.[6]

[6] Fuzzy regression discontinuity designs are most easily analyzed using instrumental variables. In the language of instrumental variables, the relationship between achievement and kinks in the enrollment function is the reduced form, while the change in class size at the kinks is the first stage. The ratio of reduced form to first-stage effects is an instrumental variable estimate of the causal effect of class size on test scores. Imbens and Lemieux (2008) offer a practitioners' guide to the use of regression discontinuity designs in economics.

The key assumption that drives regression discontinuity estimation of causal effects is that individuals are otherwise similar on either side of the discontinuity (or that any differences can be controlled using smooth functions of the enrollment rates, also known as the "running variable," that determine the kink points). In the Angrist–Lavy study, for example, we would like students to have similar family backgrounds when they attend schools with grade enrollments of 35–39 and 41–45. One test of this assumption, illustrated by Angrist and Lavy (and Hoxby, 2000) is to estimate effects in an increasingly narrow range around the kink points; as the interval shrinks, the jump in class size stays the same or perhaps even grows but the estimates should be subject to less and less omitted variables bias. Another test, proposed by McCrary (2008), looks for bunching in the distribution of student background characteristics around the kink. This bunching might signal strategic behavior—an effort by some families, presumably not a random sample, to sort themselves into schools with smaller classes. Finally, we can simply look for differences in mean pre-treatment characteristics around the kink.

In a recent paper, Urqiola and Verhoogen (2009) exploit enrollment cutoffs like those used by Angrist and Lavy in a sample from Chile. The Chilean data exhibit an enticing first stage, with sharp drops (discontinuities) in class size at the cutoffs (multiples of 45). But household characteristics also differ considerably across these same kinks, probably because the Chilean school system, which is mostly privatized, offers both opportunities and incentives for wealthier students to attend schools just beyond the cutoffs. The possibility of such a pattern is an important caution for users of regression discontinuity methods, though Urqiola and Verhoogen note that the enrollment manipulation they uncover for Chile is far from ubiquitous and does not arise in the Angrist–Lavy study. A large measure of the attraction of the regression discontinuity design is its experimental spirit and the ease with which claims for validity of the design can be verified.

The last arrow in the quasi-experimental quiver is differences-in-differences, probably the most widely applicable design-based estimator. Differences-in-differences policy analysis typically compares the evolution of outcomes in groups affected more and less by a policy change. The most compelling differences-in-differences-type studies report outcomes for treatment and control observations for a period long enough to show the underlying trends, with attention focused on how deviations from trend relate to changes in policy. Figure 1, from Donohue and Wolfers (2005), illustrates this approach for the death penalty question. This figure plots homicide rates in Canada and the United States for over half a century, indicating periods when the death penalty was in effect in the two countries. The point of the figure is not to focus on Canada's consistently lower homicide rate, but instead to point out that Canadian and U.S. homicide rates move roughly in parallel, suggesting that America's sharp changes in death penalty policy were of little consequence for murder. The figure also suggests that the deterrent effect of

capital punishment would have to be large to be visible against the background noise of yearly fluctuations in homicide rates.

Paralleling the growth in quasi-experimental experiment designs, the number and scope of real experiments has increased dramatically, with a concomitant increase in the quality of experimental design, data collection, and statistical analysis. While 1970s-era randomized studies of the negative income tax were compromised by misreporting and differential attrition in treatment and control groups, researchers today give these concerns more attention and manage them more effectively. Such problems are often solved by a substantial reliance on administrative data, and a more sophisticated interpretation of survey data when administrative records are unavailable.

A landmark randomized trial related to education production is the Tennessee STAR experiment. In this intervention, more than 10,000 students were randomly assigned to classes of different sizes from kindergarten through third grade. Like the negative income tax experiments, the STAR experiment had its flaws. Not all subjects contributed follow-up data and some self-selected into smaller classes after random assignment. A careful analysis by Krueger (1999), however, shows clear evidence of achievement gains in smaller classes, even after taking attrition and self-selection into account.[7]

Economists are increasingly running their own experiments as well as processing the data from experiments run by others. A recent randomized trial of a microfinance scheme, an important policy tool for economic development, is an ambitious illustration (Banerjee, Duflo, Glennerster, and Kinnan, 2009). This study evaluates the impact of offering small loans to independent business owners living in slums in India. The Banerjee et al. study randomizes the availability of microcredit across over 100 Indian neighborhoods, debunking the claim that realistic and relevant policy interventions cannot be studied with random assignment.

With the growing focus on research design, it's no longer enough to adopt the language of an orthodox simultaneous equations framework, labeling some variables endogenous and others exogenous, without offering strong institutional or empirical support for these identifying assumptions. The new emphasis on a credibly exogenous source of variation has also filtered down to garden-variety regression

[7] A related development at the forefront of education research is the use of choice lotteries as a research tool. In many settings where an educational option is over-subscribed, allocation among applicants is by lottery. The result is a type of institutional random assignment, which can then be used to study school vouchers, charter schools, and magnet schools (for example, Rouse, 1998, who looks at vouchers).

estimates, in which researchers are increasingly likely to focus on sources of omitted variables bias, rather than a quixotic effort to uncover the "true model" generating the data.[8]

More Transparent Discussion of Research Design

Over 65 years ago, Haavelmo submitted the following complaint to the readers of *Econometrica* (1944, p. 14): "A design of experiments (a prescription of what the physicists call a 'crucial experiment') is an essential appendix to any quantitative theory. And we usually have some such experiment in mind when we construct the theories, although—unfortunately—most economists do not describe their design of experiments explicitly."

In recent years, the notion that one's identification strategy—in other words, research design—must be described and defended has filtered deeply into empirical practice. The query "What's your identification strategy?" and others like it are now heard routinely at empirical workshops and seminars. Evidence for this claim comes from the fact that a full text search for the terms "empirical strategy," "identification strategy," "research design," or "control group" gets only 19 hits in Econlit from 1970–1989, while producing 742 hits from 1990–2009. We acknowledge that just because an author uses the term "research design" does not mean that he or she has a good one! Moreover, some older studies incorporate quality designs without using today's language. Still, the shift in emphasis is dramatic and reflects a trend that's more than semantic.

Good designs have a beneficial side effect: they typically lend themselves to a simple explanation of empirical methods and a straightforward presentation of results. The key findings from a randomized experiment are typically differences in means between treatment and controls, reported before treatment (to show balance) and after treatment (to estimate causal effects). Nonexperimental results can often be presented in a manner that mimics this, highlighting specific contrasts. The Donohue and Wolfers (2005) differences-in-differences study mentioned above illustrates this by focusing on changes in American law as a source of quasi-experimental variation and documenting the parallel evolution of outcomes in treatment and control groups in a comparison of the U.S. and Canada.

---

[8] The focus on omitted variables bias is reflected in a burgeoning literature using matching and the propensity score as an alternative (or complement) to regression. In the absence of random assignment, such strategies seek to eliminate observable differences between treatment and control groups, with little or no attention devoted to modelling the process determining outcomes. See Imbens and Wooldridge (2009) for an introduction.

Whither Sensitivity Analysis?

Responding to what he saw as the fragility of naive regression analysis, Leamer (1983) proposed extreme bounds analysis, which focuses on the distribution of results generated by a variety of specifications. An extreme version of extreme bounds analysis appears in Sala-i-Martin's (1997) paper reporting two million regressions related to economic growth. Specifically, in a variation on a procedure first proposed in this context by Levine and Renelt (1992), Sala-i-Martin computes two million of the many possible growth regressions that can be constructed from 62 explanatory variables. He retains a fixed set of three controls (GDP, life expectancy, and the primary school enrollment rate in 1960), leaving 59 possible "regressors of interest." From these 59, sets of three additional controls are chosen from 58 while the 59th is taken to be the one of interest. This process is repeated until every one of the 59 possible regressors of interest has played this role in equations with all possible sets of three controls, generating 30,857 regressions per regressor of interest. The object of this exercise is to see which variables are robustly significant across specifications.

Sala-i-Martin's (1997) investigation of extreme bounds analysis must have been fun. Happily, however, this kind of agnostic specification search has not emerged as a central feature of contemporary empirical work. Although Sala-i-Martin succeeds in uncovering some robustly significant relations (the "fraction of the population Confucian" is a wonderfully robust predictor of economic growth), we don't see why this result should be taken more seriously than the naive capital punishment specifications criticized by Leamer. Are these the right controls? Are six controls enough? How are we to understand sources of variation in one variable when the effects of three others, arbitrarily chosen, are partialed out? Wide-net searches of this kind offer little basis for a causal interpretation.

Design-based studies typically lead to a focused and much narrower specification analysis, targeted at *specific* threats to validity. For example, when considering results from a randomized trial, we focus on the details of treatment assignment and the evidence for treatment-control balance in pre-treatment variables. When using instrumental variables, we look at whether the instrument might have causal effects on the outcome in ways other than through the channel of interest (in simultaneous equations lingo, this is an examination of the exclusion restriction). With differences-in-differences, we look for group-specific trends, since such trends can invalidate a comparison of changes across groups. In a regression discontinuity design, we look at factors like bunching at the cutoff point, which might suggest that the cutoff directly influenced behavior. Since the nature of the experiment is clear in these designs, the tack we should take when assessing validity is also clear.

Mad About Macro

In an essay read to graduating University of Chicago economics students in 1988, Robert Lucas (1988) described what, as he sees it, economists do. Lucas used the specific question of the connection between monetary policy and economic depression to frame his discussion, which is very much in the experimentalist spirit: "One way to demonstrate that I understand this connection—I think the only really convincing way—would be for me to engineer a depression in the United States by manipulating the US money supply."

Ruling out such a national manipulation as immoral, Lucas (1988) describes how to create a depression by changing the money supply at Kennywood Park, an amusement park near Pittsburgh that is distinguished by stunning river views, wooden roller coasters, and the fact that it issues its own currency. Lucas's story is evocative and compelling (the Kennywood allegory is a version of Lucas, 1973). We're happy to see a macroeconomist of Lucas's stature use an experimental benchmark to define causality and show a willingness to entertain quasi-experimental evidence on the effects of a change in the money supply. Yet this story makes us wonder why the real world of empirical macro rarely features design-based research.

Many macroeconomists have abandoned traditional empirical work entirely, focusing instead on "computational experiments," as described in this journal by Kydland and Prescott (1996). In a computational experiment, researchers choose a question, build a (theoretical) model economy, "calibrate" the model so that its behavior mimics the real economy along some key statistical dimensions, and then run a computational experiment by changing model parameters (for example, tax rates or the money supply rule) to address the original question. The last two decades have seen countless studies in this mold, often in a dynamic stochastic general equilibrium framework. Whatever might be said in defense of this framework as a tool for clarifying the implications of economic models, it produces no direct evidence on the magnitude or existence of causal effects. An effort to put reasonable numbers on theoretical relations is harmless and may even be helpful. But it's still theory.

Some rays of sunlight poke through the grey clouds of dynamic stochastic general equilibrium. One strand of empirical macro has turned away from modeling outcome variables such as GDP growth, focusing instead on the isolation of useful variation in U.S. monetary and fiscal policy. A leading contribution here is Romer and Romer (1989), who, in the spirit of Friedman and Schwartz (1963), reviewed the minutes of Federal Reserve meetings in an attempt to isolate events that look like good monetary policy "experiments." Their results suggest that monetary contractions have significant and

long-lasting effects on the real economy. Later, in Romer and Romer (2004), they produced similar findings for the effects of policy shocks conditional on the Fed's own forecasts. [9]

The Romers' work is design-based in spirit and, for the most part, in detail. Although a vast literature models Federal Reserve decision making, until recently, surprisingly few studies have made an institutional case for policy experiments as the Romers' do. Two recent monetary policy studies in the Romer spirit, and perhaps even closer to the sort of quasi-experimental work we read and do, are Richardson and Troost (2009), who exploit regional differences in Fed behavior during the Depression to study liquidity effects, and Velde (2009), who describes the results of an extreme monetary experiment much like the one Lucas envisioned (albeit in eighteenth-century France). Romer and Romer (2007) use methods similar to those they used for money to study fiscal policy, as do Ramey and Shapiro (1998) and Barro and Redlick (2009), who investigate the effects of large fiscal shocks due to wars.

The literature on empirical growth has long suffered from a lack of imagination in research design, but here too the picture has recently improved. The most influential design-based study in this area has probably been Acemoglu, Johnson, and Robinson (2001), who argue that good political institutions are a key ingredient in the recipe for growth, an idea growth economists have entertained for many decades. The difficulty here is that better institutions might be a luxury that richer countries can enjoy more easily, leading to a vexing reverse causality problem. Acemoglu et al. (2001) try to overcome this problem by using the differential mortality rates of European settlers in different colonies as an instrument for political institutions in the modern successor countries. Their argument goes like this: where Europeans faced high mortality rates, they couldn't settle, and where Europeans couldn't settle, colonial regimes were more extractive, with little emphasis on property rights and democratic institutions. Where European immigrants could settle, they frequently tried to emulate the institutional set-up of their home countries, with stronger property rights and more democratic institutions. This approach leads to an instrumental variables strategy where the instrument for the effect of institutions on growth is settler mortality. [10]

---

[9] Angrist and Kuersteiner (2007) implement a version of the Romer and Romer (2004) research design using the propensity score and an identification argument cast in the language of potential outcomes commonly used in microeconometric program evaluation.

[10] Albouy (2008) raises concerns about the settler mortality data that Acemoglu, Johnson, and Robinson (2001) used to construct instruments. See Acemoglu, Johnson, and Robinson (2006) for a response to earlier versions of Albouy's critique.

Acemoglu, Johnson, and Robinson (2001) are in the vanguard of a wave of promising research on the sources of economic growth using a similar style. Examples include Bleakley (2007), who looks at the effect of hookworm eradication on income in the American South; and Rodrik and Wacziarg (2005) and Persson and Tabellini (2008), who investigate interactions between democracy and growth using differences-in-differences type designs.

With these examples accumulating, macroeconomics seems primed for a wave of empirical work using better and more clearly articulated designs. Ricardo Reis, a recently tenured macroeconomist at Columbia University, observed in the wake of the 2008 financial crisis: "Macroeconomics has taken a turn towards theory in the last 10–15 years. Most young macroeconomists are more comfortable with proving theorems than with getting their hands on any data or speculating on current events."[11] The charge that today's macro agenda is empirically impoverished comes also from older macro warhorses like Mankiw (2006) and Solow (2008). But the recent economic crisis, fundamentally a macroeconomic and policy-related affair, has spawned intriguing design-based studies of the crisis' origins in the mortgage market (Keys, Mukherjee, Seru, and Vig, 2010; Bubb and Kaufman, 2009). The theory-centric macro fortress appears increasingly hard to defend.

Industrial Disorganization

An important question at the center of the applied industrial organization agenda is the effect of corporate mergers on prices. One might think, therefore, that studies of the causal effects of mergers on prices would form the core of a vast micro-empirical literature, the way hundreds of studies in labor economics have looked at union relative wage effects. We might also have expected a large parallel literature evaluating merger policy, in the way that labor economists have looked at the effect of policies like right-to-work laws. But it isn't so. In a recent review, Ashenfelter, Hosken, and Weinberg (2009) found only about 20 empirical studies evaluating the price effects of consummated mergers directly; for example, Borenstein (1990) compares prices on airline routes out of hubs affected to differing degrees by mergers. Research on the aggregate effects of merger policy seems to be even more limited; see the articles by Baker (2003) and Crandall and Winston (2003) in this journal for a review and conflicting interpretations.

[11] As quoted by Justin Wolfers' (2008) in his New York Times column "Freakonomics" (<http://freakonomics.blogs.nytimes.com/2008/03/31/more-on-the-missing-macroeconomists/>).

The dominant paradigm for merger analysis in modern academic studies, sometimes called the "new empirical industrial organization," is an elaborate exercise consisting of three steps: The first estimates a demand system for the product in question, often using the discrete choice/differentiated products framework developed by Berry, Levinsohn, and Pakes (1995). Demand elasticities are typically identified using instrumental variables for prices; often, the instruments are prices in other markets (as in Hausman, 1996). Next, researchers postulate a model of market conduct, say, Bertrand–Nash price-based competition between different brands or products. In the context of this model, the firms' efforts to maximize profits lead to a set of relationships between prices and marginal costs for each product, with the link provided by the substitution matrix estimated in the initial step. Finally, industry behavior is simulated with and without the merger of interest.

Nevo (2000) uses this approach to estimate the effect of mergers on the price of ready-to-eat breakfast cereals in a well-cited paper. Nevo's study is distinguished by careful empirical work, attention to detail, and a clear discussion of the superstructure of assumptions upon which it rests. At the same time, this elaborate superstructure should be of concern. The postulated demand system implicitly imposes restrictions on substitution patterns and other aspects of consumer behavior about which we have little reason to feel strongly. The validity of the instrumental variables used to identify demand equations—prices in other markets—turns on independence assumptions across markets that seem arbitrary. The simulation step typically focuses on a single channel by which mergers affect prices—the reduction in the number of competitors—when at least in theory a merger can lead to other effects like cost reductions that make competition tougher between remaining producers. In this framework, it's hard to see precisely which features of the data drive the ultimate results.

Can mergers be analyzed using simple, transparent empirical methods that trace a shorter route from facts to findings? The challenge for a direct causal analysis of mergers is to use data to describe a counterfactual world in which the merger didn't occur. Hastings (2004) does this in a study of the retail gasoline market. She analyzes the takeover of independent Thrifty stations by large vertically integrated station owner ARCO in California, with an eye to estimating the effects of this merger on prices at Thrifty's competitors. Hastings' research design specifies a local market for each station: treatment stations are near a Thrifty station, control stations are not. She then compares prices around the time of the merger using a straightforward differences-in-differences framework.

A drawback of the Hastings (2004) analysis is that it captures the effects of a merger on Thrifty's competitors, but not on the former Thrifty stations. Still, it seems likely that anticompetitive effects would turn up at any station operating in affected markets. We therefore see the Hastings approach as a fruitful

change in direction. Her estimates have clear implications for the phenomenon of interest, while their validity turns transparently on the quality of the control group, an issue that can be assessed using pre-merger observations to compare price trends. Hastings's paper illustrates the power of this approach by showing almost perfectly parallel price trends for treatment and control stations in two markets (Los Angeles and San Diego) in pre-treatment months, followed by a sharp uptick in Thrifty competitor pricing after the merger.[12]

For policy purposes, of course, regulators must evaluate mergers before they have occurred; design-based studies necessarily capture the effects of mergers after the fact. Many new empirical industrial organization studies forecast counterfactual outcomes based on models and simulations, without a clear foundation in experience. But should antitrust regulators favor the complex, simulation-based estimates coming out of the new empirical industrial organization paradigm over a transparent analysis of past experience? At a minimum, we'd expect such a judgment to be based on evidence showing that the simulation-based approach delivers reasonably accurate predictions. As it stands, the proponents of this work seem to favor it as a matter of principle.

So who can you trust when it comes to antitrust? Direct Hastings (2004)–style evidence, or structurally derived estimates as in Nevo (2000)? We'd be happy to see more work trying to answer this question by contrasting credible quasi-experimental estimates with results from the new empirical industrial organization paradigm. A pioneering effort in this direction is Hausman and Leonard's (2002) analysis contrasting "direct" (essentially, differences-in-differences) and "indirect" (simulation-based) estimates of the equilibrium price consequences of a new brand of toilet paper. They evaluate the economic assumptions underlying alternative structural models (for example, Nash–Bertrand competition) according to whether the resulting structural estimates match the direct estimates. This is reminiscent of Lalonde's (1986) comparison of experimental and nonexperimental training estimates, but instead of contrasting model-based estimates with those from a randomized trial, the direct estimates are taken to provide a benchmark that turns on fewer assumptions than the structural approach. Hausman and Leonard conclude that one of their three structural models produces estimates "reasonably similar" to the direct estimates. Along the same lines, Peters (2006) looks at the predictive value of structural analyses of airline mergers, and finds that structural simulation methods yield poor predictions of post-merger ticket

[12] As with most empirical work, Hastings's (2004) analysis has its problems and her conclusions may warrant qualification. Taylor, Kreisle, and Zimmerman (2007) fail to replicate Hastings' findings using an alternative data source. Here as elsewhere, however, a transparent empirical strategy facilitates replication and constructive criticism.

prices. Likewise, Ashenfelter and Hosken (2008) compare differences-in-differences-type estimates of the effects of the breakfast cereals merger to those reported by Nevo (2000). Ashenfelter and Hoskens conclude that transparently identified design-based results differ markedly from those produced by the structural approach. Weinberg and Hosken (2009) similarly report a poor match between structural and quasi-experimental estimates of the effects of mergers on the prices of motor oil and syrup.

A good structural model might tell us something about economic mechanisms as well as causal effects. But if the information about mechanisms is to be worth anything, the structural estimates should line up with those derived under weaker assumptions. Does the new empirical industrial organization framework generate results that match credible design-based results? So far, the results seem mixed at best. Of course, the question of which estimates to prefer turns on the quality of the relevant quasi-experimental designs and our faith in the ability of a more elaborate theoretical framework to prop up a weakly identified structural model. We find the empirical results generated by a good research design more compelling than the conclusions derived from a good theory, but we also hope to see industrial organization move towards stronger and more transparent identification strategies in a structural framework.

Has the Research Design Pendulum Swung Too Far?

The rise of the experimentalist paradigm has provoked a reaction, as revolutions do. The first counterrevolutionary charge raises the question of external validity—the concern that evidence from a given experimental or quasi-experimental research design has little predictive value beyond the context of the original experiment. The second charge is that experimentalists are playing small ball while big questions go unanswered.

External Validity

A good research design reveals a particular truth, but not necessarily the whole truth. For example, the Tennessee STAR experiment reduced class sizes from roughly 25 to 15. Changes in this range need not reveal the effect of reductions from 40 students to 30. Similarly, the effects might be unique to the state of Tennessee. The criticism here—made by a number of authors including Heckman (1997); Rosenzweig and Wolpin (2000); Heckman and Urzua (2009); and Deaton (2009)—is that in the quest for internal validity, design-based studies have become narrow or idiosyncratic.

Perhaps it's worth restating an obvious point. Empirical evidence on any given causal effect is always local, derived from a particular time, place, and research design. Invocation of a superficially general structural framework does not make the underlying variation or setting more representative. Economic theory often suggests general principles, but extrapolation of causal effects to new settings is always speculative. Nevertheless, anyone who makes a living out of data analysis probably believes that heterogeneity is limited enough that the well-understood past can be informative about the future.

A constructive response to the specificity of a given research design is to look for more evidence, so that a more general picture begins to emerge. For example, one of us (Angrist) has repeatedly estimated the effects of military service, with studies of veterans of World War II, the Vietnam era, the first Gulf War, and periods in between. The cumulative force of these studies has some claim to external validity—that is, they are helpful in understanding the effects of military service for those who served in any period and therefore, hopefully, for those who might serve in the future. In general, military service tends to depress civilian earnings, at least for whites, a finding that is both empirically consistent and theoretically coherent. The primary theoretical channel by which military service affects earnings is human capital, particularly in the form of lost civilian experience. In a design-based framework, economic theory helps us understand the picture that emerges from a constellation of empirical findings, but does not help us paint the picture. For example, the human capital story is not integral to the validity of instrumental variable estimates using draft lottery numbers as instruments for Vietnam-era military service (as in Angrist, 1990). But human capital theory provides a framework that reconciles larger losses early in a veteran's career (when experience profiles tend to be steeper) with losses dissipating after many years (as shown in Angrist and Chen, 2008).

The process of accumulating empirical evidence is rarely sexy in the unfolding, but accumulation is the necessary road along which results become more general (Imbens, 2009, makes a similar point). The class size literature also illustrates this process at work. Reasonably well-identified studies from a number of advanced countries, at different grade levels and subjects, and for class sizes ranging anywhere from a few students to about 40, have produced estimates within a remarkably narrow band (Krueger, 1999; Angrist and Lavy, 1999; Rivkin, Hanushek, and Kain, 2005; Heinesen, forthcoming). Across these studies, a ten-student reduction in class size produces about a 0.2 to 0.3 standard deviation increase in individual test scores. Smaller classes do not always raise test scores, so the assessment of findings should be qualified (see, e.g., Hoxby, 2000). But the weight of the evidence suggests that class size reductions generate modest achievement gains, albeit at high cost.

Applied micro fields are not unique in accumulating convincing empirical findings. The evidence on the power of monetary policy to influence the macro economy also seems reasonably convincing. As we see it, however, the most persuasive evidence on this point comes not from elaborate structural models, which only tell us that monetary policy does or does not affect output depending on the model, but from credible empirical research designs, as in some of the work we have discussed. Not surprisingly, the channels by which monetary policy affects output are less clear than the finding that there is an effect. Questions of why a given effect appears are usually harder to resolve than the questions of whether it appears or how large it is. Like most researchers, we have an interest in mechanisms as well as causal effects. But inconclusive or incomplete evidence on mechanisms does not void empirical evidence of predictive value. This point has long been understood in medicine, where clinical evidence of therapeutic effectiveness has for centuries run ahead of the theoretical understanding of disease.

Taking the "Econ" out of Econometrics too?

Related to the external validity critique is the claim that the experimentalist paradigm leads researchers to look for good experiments, regardless of whether the questions they address are important. In an engaging account in *The New Republic*, Scheiber (2007) argued that young economists have turned away from important questions like poverty, inequality, and unemployment to study behavior on television game shows. Scheiber quotes a number of distinguished academic economists who share this concern. Raj Chetty comments: "People think about the question less than the method . . . so you get weird papers, like sanitation facilities in Native American reservations." James Heckman is less diplomatic: "In some quarters of our profession, the level of discussion has sunk to the level of a *New Yorker* article."

There is no shortage of academic triviality. Still, Scheiber's (2007) critique misses the mark because he equates triviality with narrowness of context. For example, he picks on DellaVigna and Malmendier (2006), who look at the attendance and renewal decisions of health club members, and on Conlin, O'Donoghue, and Vogelsang (2007), who study catalog sales of winter clothing. Both studies are concerned with the behavioral economics notion of present-oriented biases, an issue with far-reaching implications for economic policy and theory. The market for snow boots seems no less interesting in this context than any other retail market, and perhaps more so if the data are especially good. We can look to these design-based studies to validate the findings from more descriptive empirical work on bigger-ticket items. For example, DellaVigna and Paserman (2005) look for present-oriented biases in job search behavior.

In the empirical universe, evidence accumulates across settings and study designs, ultimately producing some kind of consensus. Small ball sometimes wins big games. In our field, some of the best research designs used to estimate labor supply elasticities exploit natural and experimenter-induced variation in specific labor markets. Oettinger (1999) analyzes stadium vendors' reaction to wage changes driven by changes in attendance, while Fehr and Goette (2007) study bicycle messengers in Zurich who, in a controlled experiment, received higher commission rates for one month only. These occupations might seem small and specialized, but they are no less representative of today's labor market than the durable manufacturing sector that has long been of interest to labor economists.

These examples also serve to refute the claim that design-based empirical work focuses on narrow policy effects and cannot uncover theoretically grounded structural parameters that many economists care about. Quasi-experimental labor supply studies such as Oettinger (1999) and Fehr and Goette (2007) try to measure the intertemporal substitution elasticity, a structural parameter that can be derived from a stochastic dynamic framework. Labor demand elasticities, similarly structural, can also be estimated using quasi-experiments, as in Card (1990b), who exploits real wage variation generated by partial indexation of union contracts.

Quasi-experimental empirical work is also well suited to the task of contrasting competing economic hypotheses. The investigations of present-oriented biases mentioned above focus on key implications of alternative models. In a similarly theoretically-motivated study, Karlan and Zinman (2009) try to distinguish moral hazard from adverse selection in the consumer credit market using a clever experimental design involving two-stage randomization. First, potential borrowers were offered different interest rates before they applied for loans. Their initial response to variation in interest rates is used to gauge adverse selection. Some of the customers who took loans were then randomly given rates lower than the rates initially offered. This variation is used to identify moral hazard in a sample where everyone has already committed to borrow.

What about grand questions that affect the entire world or the march of history? Nunn (2008) uses a wide range of historical evidence, including sailing distances on common trade routes, to estimate the long-term growth effects of the African slave trade. Deschênes and Greenstone (2007) use random year-to-year fluctuations in temperature to estimate effects of climate change on energy use and mortality. In a study of the effects of foreign aid on growth, Rajan and Subramanian (2008) construct instruments for foreign aid from the historical origins of donor–recipient relations. These examples and many more speak eloquently for the wide applicability of a design-based approach. Good research designs complement good questions. At the same time, in favoring studies that feature good designs, we accept an incremental

approach to empirical knowledge in which well-designed studies get the most weight while other evidence is treated as more provisional.

Conclusion

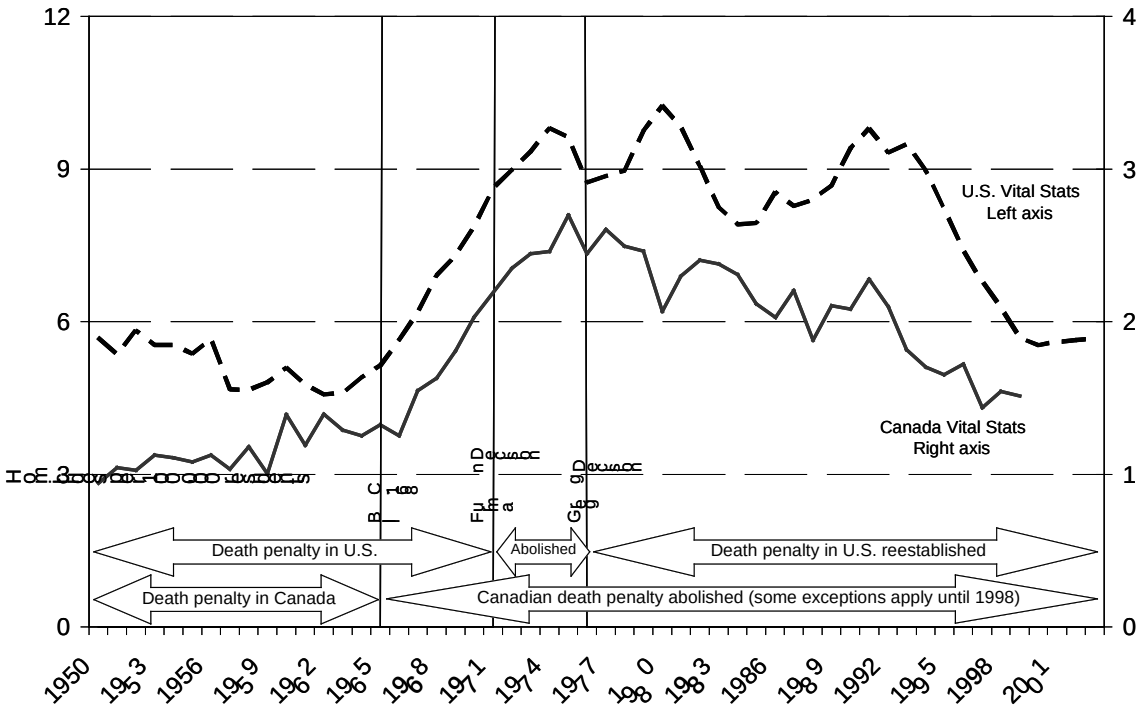
Leamer (1983) drew an analogy between applied econometrics and classical experimentation, but his proposal for the use of extreme bounds analysis to bring the two closer is not the main reason why empirical work in economics has improved. Improvement has come mostly from better research designs, either by virtue of outright experimentation or through the well-founded and careful implementation of quasi-experimental methods. Empirical work in this spirit has produced a credibility revolution in the fields of labor, public finance, and development economics over the past 20 years. Design-based revolutionaries have notched many successes, putting hard numbers on key parameters of interest to both policy makers and economic theorists. Imagine what could be learned were a similar wave to sweep the fields of macroeconomics and industrial organization.

We thank Guido Imbens for suggesting this topic and for feedback; Daron Acemoglu, Olivier Blanchard, John Donohue, Isaac Ehrlich, Glenn Ellison, Jeff Grogger, Radha Iyengar, Larry Katz, Alan Krueger, Ethan Ilzetzki, Guido Lorenzoni, Albert Marcet, Aviv Nevo, Alan Manning, Bruce Meyer, Parag Pathak, Gary Solon, Matt Weinberg, and Justin Wolfers for helpful comments and discussions; and the JEP editors—David Autor, James Hines, Charles Jones, and Timothy Taylor—for comments on earlier drafts. Remaining errors and omissions are our own.

Figure 1

Homicide Rates and the Death Penalty in the United States and Canada

(U.S. and Canada rates on the left and right y*-axes, respectively)*



Source: Donohue and Wolfers (2005).

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review*, 91(5): 1369–1401.

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2006. "Reply to the Revised (May 2006) Version of David Albouy's 'The Colonial Origins of Comparative Development: An Investigation of the Settler Mortality Data.'" Available at: http://econ-www.mit.edu/faculty/acemoglu/paper.

Albouy, David Y. 2008. "The Colonial Origins of Comparative Development: An Investigation of the Settler Mortality Data." NBER Working Paper 14130.

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review*, 80(3): 313–36.

Angrist, Joshua D. 2004. "Education Research Changes Tack." *Oxford Review of Economic Policy*, 20(2): 198–212.

Angrist, Joshua D., and Stacey Chen. 2008. "Long-term Economic Consequences of Vietnam-Era Conscription: Schooling, Experience and Earnings." IZA Discussion Paper 3628.

Angrist, Joshua D., and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association*, 90(430): 431–42.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106(4): 976–1014.

Angrist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics*, vol. 3, ed. O. Ashenfelter and D. Card. 1277–1366. Amsterdam: North-Holland.

Angrist, Joshua D., and Guido Kuersteiner. 2007. "Semiparametric Causality Tests Using the Policy Propensity Score." NBER Working Paper 10975.

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics*, 114(2): 533–75.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricists Companion*. Princeton: Princeton University Press.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics*, 60(1): 47–57.

Ashenfelter, Orley. 1987. "The Case for Evaluating Training Programs with Randomized Trials." *Economics of Education Review,* 6(4): 333–38.

Ashenfelter, Orley, and David Card. 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics*, 67(4): 648–60.

Ashenfelter, Orley, and Daniel Hosken. 2008. "The Effect of Mergers on Consumer Prices: Evidence from Five Selected Case Studies." NBER Working Paper 13859.

Ashenfelter, Orley, Daniel Hosken, and Matthew Weinberg. 2009. "Generating Evidence to Guide Merger Enforcement?" NBER Working Paper 14798.

Ashenfelter, Orley, and Mark W. Plant. 1990. "Nonparametric Estimates of the Labor-Supply Effects of Negative Income Tax Programs." *Journal of Labor Economics*, 8(1, Part 2): S396–S415.

Ayres, Ian. 2007. *Super Crunchers.* New York: Bantam Books.

Baker, Jonathon B. 2003. "The Case for Antitrust Enforcement." *Journal of Economic Perspectives*: 17(4): 27–50.

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. 2009. "The Miracle of Microfinance? Evidence from a Randomized Evaluation." Unpublished manuscript, MIT Department of Economics, May.

Barro, Robert J., and Charles J. Redlick. 2009. "Macroeconomic Effects from Government Purchases and Taxes." NBER Working Paper 15369.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica*, 63(4): 841–90.

Bleakley, Hoyt. 2007. "Disease and Development: Evidence from Hookworm Eradication in the American South." *Quarterly Journal of Economics*, 122(1): 73–117.

Borenstein, Severin. 1990. "Airline Mergers, Airport Dominance, and Market Power." *American Economic Review*, 80(2): 400–404.

Bound, John, David Jaeger, and Regina 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.

Bowers, William J., and Glenn L. Pierce. 1975. "The Illusion of Deterrence in Isaac Ehrlich's Research on Capital Punishment." *Yale Law Journal,* 85(2): 187–208.

Bubb, Ryan, and Alex Kaufman. 2009. "Securitization and Moral Hazard: Evidence from a Lender Cutoff Rule." Federal Reserve Bank of Boston Public Policy Discussion Paper No. 09-5.

Campbell, Donald, and Julian Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Card, David. 1990a. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review*, 43(2): 245–57.

Card, David. 1990b. "Unexpected Inflation, Real Wages, and Employment Determination in Union Contracts." *American Economic Review*: 80(4): 669–88.

Card, David, and Alan B. Krueger. 1992a. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy*, 100(1): 1–40.

Card, David, and Alan B. Krueger. 1992b. "School Quality and Black–White Relative Earnings: A Direct Assessment." *Quarterly Journal of Economics*, 107(1): 151–200.

Chamberlain, Gary. 1984. "Panel Data." In *Handbook of Econometrics*, vol. 2, ed. Zvi Griliches and Michael D. Intriligator, 1248–1318. Amsterdam: North-Holland.

Coleman, James S., et al. 1966. *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.

Conlin, Michael, Ted O'Donoghue, and Timothy J. Vogelsang. 2007. "Projection Bias in Catalog Orders." *American Economic Review*, 97(4): 1217–1249.

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis for Field Settings.* Chicago: Rand McNally.

Cooley, Thomas F., and Stephen F. LeRoy. 1986. "What Will Take the Con Out of Econometrics? A Reply to McAleer, Pagan, and Volker." *American Economic Review*, 76(3): 504–507.

Crandall, Robert W., and Clifford Winston. 2003. "Does Antitrust Policy Improve Consumer Welfare? Assessing the Evidence." *The Journal of Economic Perspectives*, 17(4): 3–26.

Deaton, Angus. 2009. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development." NBER Working Paper 14690.

DellaVigna, Stefano, and Ulrike Malmendier. 2006. "Paying Not to Go to the Gym." *American Economic Review*, 96(3): 694–719.

DellaVigna, Stefano, and Daniele Paserman. 2005. "Job Search and Impatience." *Journal of Labor Economics*, 23(3): 527–88.

Deschênes, Olivier, and Michael Greenstone. 2007. "Climate Change, Mortality, and Adaptation: Evidence from Annual Fluctuations in Weather in the US." NBER Working Paper 13178.

Donohue, John J., and Justin Wolfers. 2005. "Uses and Abuses of Empirical Evidence in the Death Penalty Debate." *Stanford Law Review*, vol. 58, pp. 791–845.

Duflo, Esther, and Michael Kremer. 2008. "Use of Randomization in the Evaluation of Development Effectiveness." In *Evaluating Development Effectiveness*, World Bank Series on Evaluation and Development, vol. 7, pp. 93–120. Transaction Publishers.

Ehrlich, Isaac. 1975a. "The Deterrent Effect of Capital Punishment: A Question of Life and Death." *American Economic Review*, 65(3): 397–417.

Ehrlich, Isaac. 1975b. "Deterrence: Evidence and Inference." *Yale Law Journal*, 85(2): 209–27.

Ehrlich, Isaac. 1977a. "The Deterrent Effect of Capital Punishment: Reply." *American Economic Review*, 67(3): 452–58.

Ehrlich, Isaac. 1977b. "Capital Punishment and Deterrence: Some Further Thoughts and Additional Evidence." *Journal of Political Economy*, 85(4): 741–88.

Ehrlich, Isaac. 1987. "On the Issue of Causality in the Economic Model of Crime and Law Enforcement: Some Theoretical Considerations and Experimental Evidence." *American Economic Review*, 77(2): 99–106.

Ehrlich, Isaac. 1996. "Crime, Punishment, and the Market for Offenses." *Journal of Economic Perspectives*, 10(1): 43–67.

Ehrlich, Isaac, and Zhiqiang Liu. 1999. "Sensitivity Analyses of the Deterrence Hypothesis: Let's Keep the Econ in Econometrics." *Journal of Law & Economics*, 42(1): 455–87.

Fehr, Ernst, and Lorenz Goette. 2007. **"**Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment." *American Economic Review*, 97(1): 298–317.

Feldstein, Martin, and Charles Horioka. 1980. "Domestic Saving and International Capital Flows." *Economic Journal*, 90(358): 314–29.

Friedman, Milton, and Anna J. Schwartz. 1963. *A Monetary History of the United States, 1867–1960*. Princeton: Princeton University Press for the National Bureau of Economic Research.

Gertler, Paul. 2004. "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment." *American Economic Review*, 94(2): 336–41.

Goldberger, Arthur S. 1991. *A Course in Econometric*s. Cambridge, MA: Harvard University Press.

Greenberg, David, and Harlan Halsey. 1983. "Systematic Misreporting and Effects of Income Maintenance Experiments on Work Effort: Evidence from the Seattle–Denver Experiment." *Journal of Labor Economics*, 1(4): 380–407.

Griliches, Zvi. 1986. "Economic Data Issues." In *Handbook of Econometrics*, vol. 3, ed. Zvi Griliches and Michael D. Intriligator, 1465–1514. Amsterdam: North-Holland.

Grogger, Jeffrey. 1990. "The Deterrent Effect of Capital Punishment: An Analysis of Daily Homicide Counts." *Journal of the American Statistical Association*, 85(410): 295–303.

Gruber, Jonathan. 1994. "The Incidence of Mandated Maternity Benefits." *American Economic Review*, 84(3): 662–41.

Haavelmo, Trygve. 1944. "The Probability Approach in Econometrics." *Econometrica*, 12(Supplement): *1–115.*

Hanushek, Eric A. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature*, 24(3): 1141–77.

Hastings, Justine S. 2004. "Vertical Relationships and Competition in Retail Gasoline Markets: Empirical Evidence from Contract Changes in Southern California." *American Economic Review*, 94(1): 317–28.

Hausman, Jerry A. 1996. "Valuation of New Goods under Perfect and Imperfect Competition." In *The Economics of New* Goods, ed. Timothy F. Bresnahan and Robert J. Gordon, 209–247. Chicago: National Bureau of Economic Research.

Hausman, Jerry A., and Gregory K. Leonard. 2002. "The Competitive Effects of a New Product Introduction: A Case Study." *Journal of Industrial Economics*, 50(3): 237–63.

Heckman, James J. 1997. "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." *Journal of Human Resources*, 32(3): 441–62.

Heckman, James J., and Sergio Urzua. 2009. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." NBER Working Paper 14706.

Heckman, James J., Anne Layne-Farrar, and Petra Todd. 1996. "Does Measured School Quality Really Matter?" In *Does Money Matter?: The Effect of School Resources on Student Achievement and Adult Success*, ed. Gary Burtless, 192–289. Washington, DC: Brookings Institution Press.

Heinesen, Eskil. Forthcoming. "Estimating Class-Size Effects Using Within-School Variation in Subject-Specific Classes." *Economic Journal*.

Hendry, David F. 1980. "Econometrics—Alchemy or Science?" *Economica*, 47(188): 387–406.

Hoenack, Stephen A., and William C. Weiler. 1980. "A Structural Model of Murder Behavior and the Criminal Justice System." *American Economic Review*, 70(3): 327–41.

Hoxby, Caroline M. 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics*, 115(4): 1239–85.

Imbens, Guido W. 2009. "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." NBER Working Paper 14896.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics*, 142(2): 615–35.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Development." *Journal of Economic Literature*, 47(1): 5–86.

Jacob, Brian A. 2004. "Public Housing, Housing Vouchers and Student Achievement: Evidence from Public Housing Demolitions in Chicago." *American Economic Review*, 94(1): 233–58.

Karlan, Dean, and Jonathan Zinman. 2009. "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment." *Econometrica* 77(6), 1993-2008.

Keys, Benjamin, Tanmoy Mukherjee, Amit Seru, and Vikrant Vig. 2010. "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans." *Quarterly Journal of Economics*, 125(1): 307–62.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica*, 75(1): 83–119.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics*, 114(2): 497–532.

Krueger, Alan B. 2003. "Economic Considerations and Class Size." *Economic Journal*, 113(485): F34–F63.

Kydland, Finn E., and Edward C. Prescott. 1977. "Rule Rather than Discretion: The Inconsistency of Optimal Plans." *Journal of Political Economy*, 85(3): 473–92.

Kydland, Finn E., and Edward C. Prescott. 1996. "The Computational Experiment: An Econometric Tool**."** *Journal of Economic Perspectives*, 10(1): 69–85.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*, 76(4): 604–620.

Leamer, Edward. 1978. *Specification Searches: Ad Hoc Inference with Non Experimental Data*. New York: John Wiley and Sons.

Leamer, Edward. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review*, 73(1): 31–43.

Leamer, Edward. 1985. "Sensitivity Analyses Would Help." *American Economic Review*, 75(3): 308–313.

Levine, Ross, and David Renelt. 1992. "A Sensitivity Analysis of Cross-Country Growth Regressions." *American Economic Review*, 82(4): 942–63.

Lucas, Robert E. 1973. "Some International Evidence on Output–Inflation Tradeoffs." *American Economic Review*, 63(3): 326–34

Lucas, Robert E. 1976. "Econometric Policy Evaluation: A Critique." In *Carnegie- Rochester Conference Series on Public Policy*, vol. 1, pp. 19–46.

Lucas, Robert E. 1988. "What Economists Do." Unpublished.

Mankiw, Gregory N. 2006. "The Macroeconomist as Scientist and Engineer." *Journal of Economic Perspectives*, 20(4): 29–46.

Manning, Willard G., Joseph P. Newhouse, Naihua Duan, Emmett B. Keeler, and Arleen Leibowitz. 1987. "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment." *American Economic Review*, 77(3): 251–77.

McAleer, Michael, Adrian R. Pagan, Paul A. Volker. 1985. "What Will Take the Con Out of Econometrics?" *American Economic Review*, 75(3): 293–307.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics*, 142(2): 698–714.

Meyer, Bruce D. 1995. "Natural and Quasi-Experiments in Economics." *Journal of Business & Economic Statistics*, 13(2): 151–61.

Nevo, Aviv. 2000. "Mergers with Differentiated Products: The Case of the Ready-to-Eat Cereal Industry." *The RAND Journal of Economics*, 31(3): 395–42**.**

Nunn, Nathan. 2008. "The Long-Term Effects of Africa's Slave Trades." *Quarterly Journal of Economics:* 123(1): 139–76.

Obstfeld, Maurice. 1995. "International Capital Mobility in the 1990s." In *Understanding Interdependence: The Macroeconomics of the Open Economy*, ed. Peter B. Kenen, 201–261. Princeton: Princeton University Press.

Oettinger, Gerald S. 1999. "An Empirical Analysis of the Daily Labor Supply of Stadium Vendors." *Journal of Political Economy*, 107(2): 360–92.

Passell, Peter, and John B. Taylor. 1977. "The Deterrent Effect of Capital Punishment: Another View." *American Economic Review*, 67(3): 445–51.

Persson Torsten, and Guido Tabellini. 2008. "The Growth Effect of Democracy: Is it Heterogeneous and How Can it be Estimated?" in E. Helpman (ed.), *Institutions and Economic Performance*, Cambridge, MA: Harvard University Press.

Peters, Craig. 2006. "Evaluating the Performance of Merger Simulation: Evidence from the US Airline Industry." *Journal of Law and Economics*, 49(2): 627–49.

Phillips, David P. 1980. "The Deterrent Effect of Capital Punishment: New Evidence on an Old Controversy." *American Journal of Sociology*, 86(1): 139–48.

Rajan, Raghuram G., and Arvind Subramanian. 2008. "Aid and Growth: What Does the Cross-Country Evidence Really Show?" *Review of Economics and Statistics*, 90(4): 643–65.

Ramey, Valerie, and Matthew D. Shapiro. 1998. "Costly Capital Reallocation and the Effects of Government Spending." *Carnegie-Rochester Conference Series on Public Policy*, 48(1): 145–94.

Richardson, Gary, and William Troost. 2009. "Monetary Intervention Mitigated Banking Panics during the Great Depression: Quasi-Experimental Evidence from a Federal Reserve District Border, 1929–1933." *Journal of Political Economy*, 117(6): 1031–73.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica*, 73(2): 417–58.

Rodrik, Dani, and Romain Wacziarg. 2005. "Do Democratic Transitions Produce Bad Outcomes?" *The American Economic Review* 95(2): 50-55.

Romer, Christina D., and David H. Romer. 1989. "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz." *NBER Macroeconomics Annual*, vol. 4, pp. 121–70.

Romer, Christina D., and David H. Romer. 2004. "A New Measure of Monetary Shocks: Derivation and Implications." *American Economic Review*, 94(4): 1055–1084.

Romer, Christina D., and David H. Romer. 2007. "The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks." NBER Working Paper 13264.

Rosenzweig, Mark R., and Kenneth I. Wolpin. 2000. "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature*, 38(4): 827–74.

Rouse, Cecilia. 1998. "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program." *Quarterly Journal of Economics*, 113(2): 553–602.

Sala-i-Martin, Xavier. 1997. "I Just Ran Two Million Regressions." *American Economic Review*, 87(2): 178–83.

Scheiber, Noam. 2007. "Freaks and Geeks. How Freakonomics Is Ruining the Dismal Science." *The New Republic*, April 2, pp. 27–31.

Schultz, T. Paul. 2004. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics*, 74(1): 199–250.

Sims, Christopher A. 1980. "Macroeconomics and Reality." *Econometrica*, 48(1): 1–48.

Sims, Christopher A. 1988. "Uncertainty across Models." *American Economic Review*, 78(2): 163–67.

Solon, Gary. 1985. "Work Incentive Effects of Taxing Unemployment Insurance." *Econometrica*, 53(2): 295–306.

Solow, Robert. 2008. "The State of Macroeconomics." *Journal of Economic Perspectives*, 22(1): 243–249.

Summers, Anita A., and Barbara L. Wolfe. 1977. "Do Schools Make a Difference?" *American Economic Review*, 67(4): 639–52.

Taylor, Christopher, Nicholas Kreisle, and Paul Zimmerman. 2007. "Vertical Relationships and Competition in Retail Gasoline Markets: Comment." The Federal Trade Commission, Bureau of Economics Working Paper 291.

Urquiola, Miguel, and Eric Verhoogen. 2009. "Class-size Caps, Sorting, and the Regression-Discontinuity Design." *American Economic Review*, 99(1): 179–215.

Velde, Francois. 2009. "Chronicles of a Deflation Unforetold." *Journal of Political Economy*, 117(4): 591–634.

Weinberg, Matthew C., and Daniel Hosken. 2009. "Using Mergers to Test a Model of Oligopoly." Unpublished paper, Cornell Department of Economics.

White, Halbert. 1980a. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, 48(4): 817–38.

White, Halbert. 1980b. "Using Least Squares to Approximate Unknown Regression Functions." *International Economic Review*, 21(1): 149–70.

Wolfers, Justin. 2008. "More on the Missing Macroeconomists." "Freakonomics" column of the *New York Times*, March 31. <http://freakonomics.blogs.nytimes.com/2008/03/31/more-on-the-missing-macroeconomists/>