Mostly Useless Econometrics? Assessing the Causal Effect of Econometric Theory

John Rust
Department of Economics
Georgetown University
USA
jr1393@georgetown.edu

now

the essence of knowledge

Contents

Mostly Useless Econometrics? Assessing the Causal Effect of Econometric Theory

John Rust

Department of Economics, Georgetown University, USA;
jr1393@georgetown.edu

ABSTRACT

Economics is highly invested in sophisticated mathematics and *empirical methodologies.* Yet the payoff to these investments in terms of uncontroverted *empirical knowledge* is much less clear. I argue that leading economics journals err by imposing an unrealistic burden of proof on empirical work: there is an obsession with establishing *causal relationships* that must be proven beyond the shadow of a doubt. It is far easier to publish theoretical econometrics, an increasingly arid subject that meets the burden of mathematical proof. But the overabundance of econometric theory has not paid off in terms of empirical knowledge, and may paradoxically hinder empirical work by obligating empirical researchers to employ the latest methods that are often difficult to understand and use and fail to address the problems that researchers actually confront. I argue that a change in the professional culture and incentives can help econometrics from losing its empirical relevance. Econometric theory needs to be more empirically motivated and problem-driven. Economics journals should lower the burden of proof for empirical work and raise the burden of proof for econometric theory. Specifically, there should be more room for descriptive empirical work in our journals. It should not be necessary to establish a causal mechanism or a non-parametrically identified structural model that provides an unambiguous explanation of empirical phenomena as a litmus test for publication. On the other hand, journals should increase the

burden on econometric theory by requiring more of them to show how the new methods they propose are likely to be *used* and be *useful* for generating new empirical knowledge.

1

Introduction

"As a mathematical discipline travels far from its empirical source, or still more, if it is a second and third generation only indirectly inspired from ideas coming from 'reality', it is beset with very grave dangers. It becomes more and more purely aestheticizing, more and more purely l'art pour l'art. This need not be bad, if the field is surrounded by correlated subjects, which still have closer empirical connections, or if the discipline is under the influence of men with an exceptionally well-developed taste. But there is a grave danger that the subject will develop along the line of least resistance, that the stream, so far from its source, will separate into a multitude of insignificant branches, and that the discipline will become a disorganized mass of details and complexities. In other words, at a great distance from its empirical source, or after much 'abstract' inbreeding, a mathematical subject is in danger of degeneration. At the inception the style is usually classical; when it shows signs of becoming baroque, then the danger signal is up. It would be easy to give examples, to trace specific evolutions into the baroque and the very high baroque, but this, again, would be too technical. In any event, whenever this stage is reached, the only remedy seems to me to be the rejuvenating return to the source: the re-injection of more or less directly empirical ideas."

von Neumann (1947)

This essay is a warning that von Neumann's danger signal is up for econometric theory, and his suggested remedy, "the re-injection of more or less directly empirical ideas", is overdue. In my opinion, econometric theory has run into seriously diminishing returns. It is increasingly abstract, technical, and difficult to understand: "baroque" is a good adjective for some of it. I am not anti-theory and realize that mathematical subjects are not easy. But too many econometric theory articles are poorly motivated, and the value of trying to implement the new estimators they propose is unclear.

Even though I teach econometrics to graduate students, I would have to admit that the subject can easily come across as a "disorganized mass of details and complexities" that does not really prepare them go out and discover new empirical knowledge. Instead I fear many of them regard it as an obstacle to be overcome: a vast catalog of *limit theorems* that archive the endless ways to use and reuse the Law of Large Numbers and Central Limit Theorem to prove even more arcane limit theorems that only tenuously pretend to have anything to do with the real challenges facing empirical researchers.

Though the title of this essay was motivated by *Mostly Harmless Econometrics* by Angrist and Pischke (2009), this is not a critique of their book. On the contrary, I would have titled their book "Mostly *Useful* Econometrics" because Angrist and Pischke are part of a group of (mostly labor-oriented) applied econometricians who *practice what they preach.* The enormous popularity of this book is due in part to the fact it is *not* written in the abstract style of econometric theory texts. Instead it provides compelling *empirical motivation* for a set of relatively easy to apply econometric methods that are actually used by applied researchers. As such, it is an essential part of the "tool kit" for doing good empirical work in economics.

On the other hand, our journals and many econometric textbooks devote too much space to estimation methods that very few of us actually use. Does anyone really use k-class estimators, or 3 stage least squares, or maximum score? Do we really need to know how to test for unit roots or cointegration in time series? Do we need to know how to non-parametrically estimate simultaneous equation systems with

non-separable errors or know about "identification at infinity" in order to do good empirical work?

Consider a typical econometric theory article in a typical issue of *Econometrica.* Rarely is there any specific empirical motivation as to why yet another estimation method is necessary, or a discussion of a *suggested* empirical application, much less an *actual* empirical application and a demonstration of how the new method changes a conclusion about a substantive empirical issue. Instead all too many are motivated by *previous econometric theory papers* and the main contribution is to generalize a previous model or proof (e.g. allow for non-additive errors instead of additive ones or re-prove a result using weaker assumptions).

Though I will avoid naming names, many econometric theorists are more like pure statisticians or mathematicians who do little empirical work themselves. Some of them have no apparent interest in economics: they call themselves "econometricians" mainly because salaries in economics are much higher than in statistics departments which have long been in decline and in some cases eliminated (such as at Princeton), or nearly eliminated (such as at Yale). Yet the celibate priesthood of econometric theorists yield great power over empirical researchers: it is difficult to publish empirical work in leading economics journals unless it is blessed by the high priests. So empirical papers published in *Econometrica* tend to be *illustrations of the latest methods* (justifying a demand for even more methods), rather than work that is focused on important economic problems or issues.

In comparison, the "hard sciences" such as physics or biology are less methodologically focused than economics, yet they appear to be much more productive of useful empirical knowledge. One only needs to look in the daily newspaper to see amazing new medical advances, huge advances in communications and computer technologies, and fundamental discoveries about the universe at both the smallest and largest scales. The hard sciences make progress because most of the research is *empirically motivated:* what causes cancer? why are atmospheric CO_2 levels rising? what does cosmic background radiation tell us about Big Bang asymmetries that kept matter from being completely annihilated by anti-matter?

Rather than endlessly debate *how to do science* people in the hard sciences just *do science.* They are much more focused on *data generation* particularly through the creation of sophisticated instruments and well designed experiments, and data gathering is often far more focused and theory-driven than in economics. For example, one of the last remaining fundamental particles predicted by the standard model of physics, the Higgs Boson (which in conjunction with the Higgs field gives particles their mass), was finally discovered in 2013 using the large hadron collider at CERN. Certainly a great deal of statistical sophistication is required to analyze the terabytes of data from billions of subatomic collisions to filter signal from noise, or to infer the presence of dark matter from doppler shifts in light from distant galaxies, or to infer small asymmetries 10^{-20} seconds after the big bang from cosmic background radiation. Even so, the average physicist is less statistically sophisticated and tooled-up than the average economist.

Economists, on the other hand, try to create an illusion of great scientific prowess by using advanced mathematics to study abstruse topics such as large square economies, unit roots, partial identification, moment inequalities and triangular models. Even topics that seem to have a useful goal, such as implementation of social choice rules, are analyzed at such a high level of abstraction that it is hard to see their practical value. Economists lionize ultra-mathematical theories, even if they have no clear real-world applicability or provide poor approximations to reality. This is certainly true of a lot of economic theory that assumes that people are rational expected utility maximizers, that firms maximize expected profits, that interactions between individuals and firms always occur in a state of Nash or competitive equilibrium, and that financial markets are complete and informationally efficient. Our math obsession has deluded us into thinking that these are good approximations to reality when there is lots of evidence that they aren't.[1]

---

[1]My critique differs from the "mathiness" critique of Romer (2015) who argues that some economists try to use mathematics to masquerade politically motivated viewpoints. I do not believe econometric theorists are "political" in this sense, and the mathematics they do is of very high quality. To the extent there is a masquerade, it is to foster the impression that all econometric theory must be useful because it ultimately enables economists to do better empirical work. Instead, my critique is

There is still time to change the orientation of theoretical econometrics to avoid the fate of pure economic theory, which has also suffered the consequences of being mostly useless. Economic theory had immense prestige in 1980s when I received my PhD, an era where empirical researchers (and even applied theorists) were viewed as distinctly second class citizens. However theoretical elitism turned out to be an unsustainable equilibrium, as the profession ignored von Neumann's danger signal and allowed economic theory to become increasingly abstract, baroque, and disconnected from reality. Hamermesh (2013) documents a significant decline in the amount of economic theory published by the top economics journals since the 1960s and speculates that "Economic theory may have become so abstruse that editors of the leading general journals, recognizing that very few of their readers could comprehend the theory, have cut back on publishing work of this type." (p. 169).

The problem of uselessness of economic theory became sufficiently severe to come to the attention of the popular press in a *New Yorker* article in 1996 titled "The Decline of Economics." It was motivated by the 1996 Nobel Prizes to James Mirrlees and William Vickrey. Vickrey had mentioned to a reporter at the *New York Times* that his famous paper on auctions (Vickrey, 1961) was "one of my digressions into abstract economics" and "At best, it's of minor significance in terms of human welfare." In response, Cassidy (1996) lamented that "Here is a world-renowned theorist confirming what many outsiders had long suspected — that a good deal of economic theory, even the kind that wins Nobel Prizes, simply doesn't matter much. That is a great pity, since economics is supposed to be a useful subject, and its intellectual founders stressed its practical importance." (p. 50).

Economic theory would be in a much better state today had profession followed the advice of Alvin Roth, who issued his own version of von Neumann's danger signal back in 1991: "if we do not take steps in the direction of adding a solid empirical base to game theory, but instead continue to rely on game theory primarily for conceptual insights

---

similar in many respects to the critique of McCloskey (2005) who concludes that the profession's obsession with "mathematical and statistical reasoning" "is a waste of time" and unless this changes "our understanding of the economic world will continue to stagnate."

(deep and satisfying as these may be), then it is likely that long before a hundred years game theory will have experienced sharply diminishing returns. In this respect, I think the next hundred years will likely bring about a change in the way theoretical and empirical work are related in economics generally, and that, if not, then the entire discipline of economics may also fail to realize its potential." (Roth, 1991a, p. 108). One only needs to look at current job statistics on *EconJobMarket.org* to see that Roth underestimated how rapidly economic theory would collapse: both the number of positions and the number of candidates who list economic theory as their primary field has dwindled to less than one fourth the corresponding numbers for econometrics.

Theoretical econometrics has a superficial appearance of usefulness since it is a field that supposed to provide us with the methods and tools for doing empirical work, but it is infected by the same theoretical elitism and detachment from reality that lead to the decline of economic theory. The current professional culture still offers far greater rewards for doing econometric theory and publishing new estimation methods than for doing applied work, especially when we recognize that data gathering and analysis is a far more laborious and less glamorous task than proving theorems. Similar to economic theory, econometric theory has become sort of an elite sport that can be played at elite upper tier departments that are rich enough to afford it (e.g. Harvard, Princeton, Yale, MIT, Stanford, Berkeley, Chicago, etc.). Among the trophies in this sport is the ability to publish in the elite upper tier journals, such as *Econometrica* which is devoted to publishing the most arid types of economic and econometric theory.

Economists should be more concerned about our collective influence and impact. One metric is citations, and compared to other sciences, economists have few citations. For example the Thomson Reuters Essential Science Indicators database from 2000 to 2010 ranks economics 17th out of its list of 21 sciences in terms of citations. Only engineering, other social sciences, computer science and mathematics had lower average citations over this period than economics. The science with the most citations according to the ESI index was molecular biology, followed by immunology. Within economics, the least cited subfields are economic

theory and (micro) econometric theory according to Ellison (2013). He finds that "The fields in which papers are estimated to have the fewest citations — political economy, history, micro theory, cross-section econometrics, and industrial organization — are all fields where we estimated that researchers have relatively low citation indexes." (p. 82).

The leading journals such as *Econometrica* have great influence on the type of work done in economics due to the hierarchical way the profession is self-organized. There is a hierarchy in the profession that is akin to a bee hive where a small number of theorists are the queen bees who have the influence to set the overall direction of research via their roles as editors at the top ranked journals. The sustainability of the bee hive depends on having cadres of worker bees who are willing to follow the directions they set in exchange for the chance to publish, get tenure, and try to make it slightly further up in the hierarchy. As von Neumann noted, this sort of academic hierarchy can be sustainable "if the field is surrounded by correlated subjects, which still have closer empirical connections, or if the discipline is under the influence of men with an exceptionally well-developed taste." Unfortunately too few leading theorists have a real interest in the real world application of their theories: this is a task left for the worker bees. The minority of theorists who do attempt to pay homage to reality often suffer from the hubris of confusing knowledge of their theories with knowledge of the real world.

For example, in his 2003 Presidential Address to the American Economic Association, Robert E. Lucas (2003) proclaimed that "central problem of depression prevention has been solved, for all practical purposes, and has in fact been solved for many decades." (p. 1). Oops! The 2008 financial crash and ensuing "Great Recession" revealed that economists actually knew far less about the world than they thought they knew. It suddenly became painfully clear that a large part of the profession was virtually clueless what was actually going on in the economy because few of the ugly real world complexities are captured in our oversimplified mathematical models.

Indeed, back in 2007 few economists knew what a "swap" was or how dangerously overlevered and interdependent most of Wall Street firms

were. This ignorance is not surprising: most real business cycle models assume away the financial sector as irrelevant. However leading experts in finance, such as Eugene Fama, are so wedded to idealized views such as the efficient markets hypothesis that they deny the possibility that a collapse of a credit/housing bubble could have lead to the 2008 Wall Street crash: "I don't know what a credit bubble means. I don't even know what a bubble means. These words have become popular. I don't think they have any meaning." (quoted from interview in Cassidy, 2010). In short, too many of us can be accused of being "egg heads" who are not in touch with the real world.

Of course some economists are/were in touch with reality, and several issued prescient warnings of stock market and housing bubbles and an impending financial meltdown well before the 2008 crash. However most of these economists, such as Robert Shiller or Noriel Roubini, were ignored or treated as "flakes" or "doctors of doom" — the "Chicken Littles" who always predict that the sky is falling. I believe Shiller and Roubini would confirm that it is hard to publish or be heard if you dare speak out against the conventional wisdom and orthodoxy in economics. While there is a nascent literature on behavioral economics that does try to learn and understand what people and firms actually do (as opposed to rationalizing behavior to force it to conform to prediction of existing theories) it is hard to publish empirical work that provides evidence of behavior that is inconsistent with orthodox theory unless it is accompanied by an elaborate non-expected utility theory that "explains" this behavior. Leading journals such as *Econometrica* are much more likely to publish *behavioral theories* but much less likely to publish *behavioral evidence.*

This is part of the burden of proof on empirical work that tends to keep economists focused on theory, but less aware of reality. The profession ought to be more receptive to self-evaluation, given that our sense of smugness and self-confidence has been shaken by dramatic embarrassments such as our failure to predict the crash of 2008 and the Great Recession, or to even agree on the best policies to deal with it after the fact. There is a surprising level of ignorance, or at least a glaring lack of professional agreement, on a host of other important questions

as well. For example there is huge disagreement about how strongly taxes affect labor supply. Despite intensive study for over three decades, there is still disagreement about the efficacy of job training programs on various outcomes such as unemployment duration and subsequent earnings and about how best to deal with crumbling institutions such as Social Security and Medicare. The profession does not agree about whether financial markets are inherently unstable and whether they need to be regulated, and if so, how. I believe that economists need to be considerably more humble and admit that there is very little that we can confidently say that we know and agree upon. Given this, perhaps it is time to have a discussion about whether this state of affairs is inherent in the subject (i.e. economics is a more difficult topic to study than physics or biology) or whether the profession's bias for deductive versus inductive inference is partly to blame for our lack of agreement and results.

A full discussion is beyond the scope of this essay because it deals with the interaction between econometric theory, economic theory, and empirical work. I have discussed the tensions between *economic theory* and empirical work elsewhere (Rust, 2014) and want to stress that I am not anti-theory, though I feel that the profession refuses to let go of favorite theories despite considerable evidence that they are at odds with reality. My focus in this essay is on *econometric theory.* I also stress I am not against econometric theory and certainly do not claim that all econometric theory is useless. However economics is "data poor" relative to other sciences, but not because economic data are inherently harder to collect. Rather, due to the professional culture and a methodological bias at the top departments, the rewards for data gathering and data analysis are low, whereas the payoff to econometric theory is much higher. This equilibrium is good for the elite departments, but bad for the profession as a whole.

The rest of this essay is organized as follows. Section 2 documents the theory bias at *Econometrica* and their efforts to correct this bias to stay relevant. Though I argue that the causal impact of econometric theory is small, the amount of econometric theory devoted to causality is huge — particularly the burgeoning literature on *treatment effects.* I

discuss this literature in Section 3, but argue that it has not resulted in the sort of *credibility revolution* in applied economics claimed by Angrist and Pischke (2010). In Section 4 I discuss a personal example that illustrates how even an incomplete understanding of causality can be *extremely useful.* Since my example is from medicine, I also discuss an economic example. In Section 5 I review a vast literature on the causal effect of training programs and conclude that the useful knowledge from decades of study by the leading econometricians has been disappointingly meager. To avoid writing a completely hopeless and depressing essay, I devote Section 6 to discussing several success stories where economics has produced useful knowledge. Unfortunately, these are not examples where econometric theory played much of a role. Section 7 concludes with some ideas about how things can be turned around to make econometric theory more useful to economics.

Although this essay questions the usefulessness of some ultra mathematical economic and econometric theory, I do not pretend to judge it from an superior vantage point. I readily admit that my own academic work has proved to be mostly useless. My views will no doubt cost me professionally, so why write this? While there are many prizes and ways economists self-congratulate themselves, there are fewer avenues for self-evaluation and few economists willing to take the professional risk to publicly voice their true concerns about the shortcomings of the discipline. I hope this essay will lead to a constructive discussion rather than be seen as an unhelpful rant on econometric theory and the state of the economics profession.

2

Unhealthy Theory Bias at *Econometrica*?

Econometrica, the flagship publication of the Econometric Society, is one of the pre-eminent journals that plays a strong leadership role in the economics profession. It has long been considered one of the top five "general interest journals" in economics. As such, the editorial policy of *Econometrica* has a big impact on professional incentives, and plays a "standard setting" role on the type of research done in economics. This view is confirmed by Card and DellaVigna (2013) who note that "Publications in the top journals have a powerful influence on the direction of research in economics, on the career paths of young researchers, and on the pay of academic economists." (p. 144). Having at least one publication in one of the top five general interest journals is virtually a requirement for a tenured appointment or promotion to tenure at the top ranked economics departments.

It is not a secret that there is a "theory bias" at *Econometrica:* after all, the Econometric Society defines itself as an "international society for the advancement of economic theory in its relation to statistics and mathematics." Table 2.1 presents data on submissions to *Econometrica* in recent years, provided courtesy of Daron Acemoglu, a previous editor. It provides a breakdown of submissions that were classified by

Table 2.1: Percent of *Econometrica* Submissions

Type	2005	2008	2009	2011	2012	2013
Empirical	6.9%	9.4%	9.9%	13.6%	14.7%	15.7%
Methodological	25.8	24.0	24.6	16.8	13.6	16.7

Econometrica staff into "empirical" "methodological" (i.e. econometric theory), and the remainder is "theory" (i.e. both applied and pure theory). We see that empirical submissions constitute only a small proportion of the submissions to *Econometrica* though there is an evident trend for an increasing share of empirical submissions that is matched by an offsetting decline in the fraction of methodological submissions. The remaining two thirds of all submissions are in economic theory. The submission data raises the question of whether the theory bias at *Econometrica* is partly our fault. Could it be that too few of us submit empirical papers to *Econometrica* out of a belief that an empirical paper will not be appreciated as much as an economic theory or econometric theory paper of comparable quality?

Table 2.2 presents data on "non-reject rates" — the fraction of submissions that are not rejected after initial submission to *Econometrica.* The non-reject rate is not a double negative synonym for acceptance rate. The reason is that some of the submissions that are not rejected may not be resubmitted, and those that are may ultimately be rejected after resubmission. We see that in 2005, empirical submissions had a much lower non-reject rate than methodological submissions, and lower than the overall non-reject rate for the journal as a whole. However by 2008 the situation had reversed and the non-reject rate for empirical submissions was significantly higher than for methodological submissions and for all submissions as a whole. By 2012 the situation had reverted and non-reject rates of empirical and methodological submissions were approximately equal and only slightly higher than for all submissions as a whole.

We can speculate the extent to which the increase in applied submissions to *Econometrica* in recent years is driven by a change in editorial policy to be more favorable to empirical work. We can speculate that

Table 2.2: Non-reject rates of *Econometrica* Submissions

Type	2005	2008	2009	2011	2012
Empirical	7.1%	18.5%	20.9%	15.9%	10.7%
Methodological	15.2	12.1	12.6	10.9	9.8
Overall	11.8%	10.0%	11.9%	9.4%	8.4%

there is an awareness among current and former *Econometrica* editors that its historical theory bias could be hurting the journal. For example in their analysis of trends of submissions and citations to the top five journals in economics, Card and DellaVigna (2013) find that "Starting in the 1990s, however, there is a discernible fall in the relative impact of *Econometrica* articles." (p. 153) and this fall in ranking is driven by "declining citations to recent papers in Econometrics and Theory" (p. 144). They find that there has been a surge in applied submissions to the other top five journals particularly at *Journal of Political Economy* and *Quarterly Journal of Economics.* These latter journals are publishing a much greater share of empirical papers than *Econometrica* in the rapidly growing fields of development, labor and industrial organization, "all applied fields that are substantially under-represented in *Econometrica* relative to the other top-five journals, particularly since 1990." (p. 158). Thus the surge of empirical submissions that *Econometrica* fails to publish (but which other top journals do publish) could be an important factor driving the falling impact of *Econometrica* relative to its competitors in the top five journal category.

The Econometric Society is no doubt aware that theory bias could be causing the decline in influence of *Econometrica* relative to other top five journals. I believe this has prompted a change in editorial policy at *Econometrica* to be more accommodating to empirical work and this is reflected in the trends in Table 2.2. However these changes may be coming too slowly to change perceptions of theory bias, and thus are failing to have a big impact on submission decisions in the profession as a whole.

For example, the Econometric Society introduced the *Frisch Medal* in 1978 as an award for "to encourage the creation of good applied

Table 2.3: Published articles in *Econometrica* March 2010 to March 2014

Number of papers	Applied Theory Papers	Empirical Papers
263	76	64

work and its submission to Econometrica. It is given every two years for an applied article (empirical or theoretical) published in Econometrica during the past five years." The papers that have been awarded the Frisch Medal have included some very important empirical contributions that are likely to have lasting influence on the profession. However despite the 19 highly publicized Frisch Medals that the Econometric Society has awarded since 1978, there appears to be a widespread perception that empirical work is not truly appreciated at *Econometrica* and it is still the case that only a minority of papers published in *Econometrica* are empirical.

I have a vantage point on this after having been asked to chair the 2014 Frisch Medal Committee that chose the paper by Cunha *et al.* (2010) as the winner. Since the Frisch Medal is awarded to any *applied article* (empirical or theoretical) the committee I chaired considered a very liberal definition of what "applied" is, so many of the submissions we considered were applied theory papers that most of us would not consider to be "empirical". Table 2.3 presents my own self-assessed classification of the set of 263 eligible applied papers published in *Econometrica* from March 2010 to March 2014 into those that are actually "empirical" and the remainder that I classify as "applied theory" (both econometric theory and economic theory). By my classification an empirical paper must contain an analysis of *actual data* and cannot simply be an econometric theory paper that introduces a new estimator that is applied only to artificial data such as a monte carlo simulation. Yet my definition of "empirical" is sufficiently broad to include papers such as Su and Judd (2012) that are actually *methodological illustrations* — they use real data instead of or in addition to monte carlo study to illustrate the properties of new estimators but otherwise it is clear that the empirical application itself is not the focus of the analysis.

The bottom line is that even by my excessively generous definition of what constitutes an "empirical" paper, less than one fourth of the papers published in *Econometrica* between March 2010 and March 2014 were empirical. It is my opinion that when readers see the preponderance of theory and the scarcity of empirical papers in typical issues of *Econometrica* they conclude that the journal is not interested in publishing empirical work and they vote with their feet by submitting empirical work to other journals.

Even though *Econometrica* publishes relatively few empirical papers, the ones it does publish are often very good. In particular, the paper that my committee chose as the winner of the most recent Frisch Medal, Cunha *et al.* (2010), meets my personal standard of "best empirical work" because it addresses a critically important economic question — what factors affect the evolution of cognitive skills and personality traits across the different stages of childhood? It also provides a significant methodological advance, but its primary contribution is to improve our understanding of a question we really do care about.

I was awarded the Frisch Medal in 1992 for my paper on replacement of bus engines (Rust, 1987). However to be perfectly honest, this paper does not meet my own personal standard of what I consider to be the best empirical work because it is essentially an illustration of a new estimation methodology, but using real data rather than with "fake data" such as a monte carlo simulation. Few readers of *Econometrica* really care about replacement of bus engines *per se* — the reason why *Econometrica* published it was because it introduced a new estimation method.

In retrospect, the editor of *Econometrica* in charge of my submission, Angus Deaton, did me a huge favor by rejecting my first attempt to introduce this new estimation method as a pure, abstract econometric theory paper (subsequently published elsewhere as Rust, 1988). Deaton was skeptical that the methodology I proposed was computationally feasible. This motivated me to go out and find data from some real world application to prove that the method was computationally feasible. To his credit, when I submitted the new applied paper, Deaton agreed that I had demonstarted that my method was computationally feasible

and might be useful, and he accepted it. The version Deaton accepted sacrificed econometric theory in favor of an empirical application but had a far greater impact and is one of my most cited papers. Readers could see how to generalize the method to other more interesting and important applications. *Econometrica* might improve its standing in the top five journal category if more of its editors emulated Deaton and required econometric theory submissions to contain some empirical applications instead of pure mathematics.

You might be skeptical at this point whether the theory bias at *Econometrica* is really doing much damage to the profession as a whole as opposed to simply damaging its own standing and relevance. The increased amount of empirical work being published by the other top journals is certainly a healthy development, but it is too early to tell whether this will fundamentally change the professional culture so that economics produces more useful knowledge. *Econometrica* does seem to be trying to change in response to this, including displaying a new found courage to publish empirical work that does not meet the standard criterion of "success" (e.g. Carrell *et al.*, 2013).[2] In the next section I discuss a second problem, namely that the prevailing agnosticism about "modeling assumptions" and obsession with establishing "causality" (something that is prevalent at all of the top journals), also contributes to the scarcity of useful empirical knowledge in economics.

---

[2] Where "success" means empirical work that confirms existing theories. I mention this paper since it is an example where an experiment revealed that an econometrically inspired theory of peer group effects that was used to "create optimally designed peer groups intended to improve academic achievement of the bottom one-third of incoming students by academic ability while not harming achievement of students at other points in the distribution" was actually completely misguided. Instead, their optimally assigned peer groups caused "negative and significant treatment effect for the students we intended to help." Why? "High and low ability students in the treatment squadrons appear to have segregated themselves into separate social networks, resulting in decreased beneficial social interactions among group members." I credit *Econometrica* for publishing empirical work that rejects invalid theories and *a priori* presumptions, helping to pave the way for better theories.

3

Econometric Theory and Causality

The question of causality, and the problem of distinguishing correlation from causality, has been one of the primary concerns of econometric theory from its inception, including the earliest work on the method of instrumental variables. According to Heckman and Pinto ([2014]) "Haavelmo's seminal 1943 and 1944 papers are the first rigorous treatment of causality. In them, he distinguished the definition of causal parameters from their identification. He showed that causal parameters are defined using hypothetical models that assign variation to some of the inputs determining outcomes while holding all other inputs fixed. He thus formalized and made operational Marshall's (1890) *ceteris paribus* analysis."

While it is certainly extremely important to try to distinguish correlation and causality, and methods such as instrumental variables are key contributions that distinguish econometrics from statistics, "proof of causality" should not the *sine qua non* for judging empirical work. Causality is an overly narrow concept that fails to capture other important reasons why we do empirical work. For example, a key reason we do empirical work is to provide *descriptive evidence* that can have value even if it does not fully reveal some underlying causal mechanism.

I believe that the economics profession has veered towards an obsessive focus on seeking overly simplistic causal explanations economic phenomena. Establishing or "proving" the existence of a causal relationship is extremely difficult even in the best case scenario where it really exists. But what do we do if a simple causal explanation doesn't exist? For example, while one can ask "what caused the Great Recession?" the way this question is posed presumes that there is a *single cause* that can be easily identified. What if most social phenomena are a result of far more complex, co-evolving dynamic interactions between large numbers of agents which makes it unlikely that one could identify a single simplistic causal explanation for any given outcome?

Too many journal editors seem excessively timid: reluctant to publish empirical work they fear could be controversial, or heaven forbid, overturned by future researchers. But controversy and disagreement are likely to surround the most important economic questions for the forseeable future, and it seems like a poor decision not to publish papers that try to make headway on difficult yet important questions. Even headway that results in an incomplete understanding of complex causal mechansims may be good enough to at least *partially control* our environment to reduce the likelihood of outcomes we consider undesirable. Adopting a carbon tax to mitigate the effects of global warming is one obvious example. In the case of financial instability and periodic financial crashes, it is unrealistic to expect we can identify a single cause and prevent them, but it would be a real success if we could understand them well enough to reduce their frequency or severity through effective policies that have widespread support.

Beyond instrumental variables, there are two main methodologies for establishing causality in econometrics: 1) structural models, and 2) the Rubin Causal Model (RCM) Rubin (2005). I have discussed some of the inherent limits of structural models elsewhere (Rust, 2014) and have only a bit more to say about them below. The Rubin causal model is based on a potential outcomes model that was originally proposed by Neyman (1923) for randomized experiments. In economics this has become known as the "treatment effects" literature, where the parameter of interest is to estimate the average treatment effect associated with

some policy intervention. Treatment effect models are unlike structural models (e.g. Heckman, 1979) that attempt to model individual *choices* and account for unobservables that affect *self-selection.* Instead even the terminology of these models reflects a perspective more appropriate for epidemiological studies with random assignment: individual choices are called "treatments" and rather than studying behavior and welfare the treatment effects literature focuses on the narrower antiseptic question of estimating the average effect of the treatment on a unidimensional *outcome.*

There is now a huge literature on alternative methods for estimating average treatment effects using non-experimental data when an "unconfoundedness" or conditional independence assumption holds. The basic idea is that if it is possible to condition on a sufficiently rich vector of covariates x, the remaining factors that cause some individuals to choose the treatment and others not are independent of x, and thus the outcome can be regarded as a virtual randomized experiment. Methods such as *matching* are used to aggregate over individuals with similar values of x in an attempt to estimate the average treatment effect. The treatment effects literature has ignored the complexity of individual choice and assumed away the effect of unobservables on decisions — an approach that Heckman, McFadden and others pioneered.

Even though I prefer structural models as the vehicle to understand causality, I put the treatment effects literature in the "mostly useful econometrics" category because most of it is empirically motivated, and inspired by an attempt to answer practical questions. However I believe that structural econometric models are a more promising way to study questions related to causality and the analysis of counterfactual policy changes. Structural econometrics involves the estimation and testing of *economic models* that I find to be more interpretable and easier to understand than "statistical models" such as RCM.

We can rigorously define what causality means in a sufficiently complete economic model, and the model allows us to say much more about how counterfactual policy changes affect variables we cannot directly observe such as agents' welfare, and a richer array of outcome variables than can be analyzed in incomplete modeling frameworks such

as the RCM. As Heckman ([2005](#)) notes, "Models of causality advocated in statistics are incomplete because they do not specify the mechanisms of external variation that are central to the definition of causality, nor do they specify the sources of randomness producing outcomes and the relationship between outcomes and selection mechanisms. By not determining the causes of effects, or modeling the relationship between potential outcomes and assignment to treatment, statistical models of causality cannot be used to provide valid answers to the numerous counterfactual questions required for policy analysis." (p. 85-86).

While it is unambiguous how to define and analyze causality *in a sufficiently complete economic model,* there is fundamental disagreement about whether the causal effects we analyze in our models actually hold in reality. As Heckman and Pinto ([2014](#)) note "Haavelmo formalized Frisch's notion that 'causality is in the mind.' Causal effects are not empirical statements or descriptions of actual worlds, but descriptions of hypothetical worlds obtained by varying — hypothetically — the inputs determining outcomes. Causal relationships are often suggested by observed phenomena, but they are abstractions from it." While models are certainly abstractions of reality, it would be completely wrong to conclude that models are useless for understanding causality. The principle to keep in mind is that the causal effects we find in models that are *sufficiently good approximations to reality* are likely to hold in reality as well.

Of course the causal mechanisms in models and theories that are viewed as relatively good approximations to reality should always be regarded as *causal hypotheses* pending falsification by subsequent evidence or experiments. Popper ([2002](#)) notes that science progresses via the successive rejection of falsified theories and models, which are replaced by new theories or models that are consistent with all available evidence, including the evidence that falsified the leading previous theory.

Clearly it is hazardous to try to infer causality using falsified models or theories, or ones that provide very bad approximations to reality. As my previous discussion indicates, in light of the profession's cluelessness in understanding the 2008 financial crash and ensuing Great Recession, there is ample ground for skepticism about whether our economic

models bear much relation to reality. When we confuse bad models with reality, and fail to do adequate empirical tests to determine whether our models are any good, we risk the dramatic professional embarrassments I have discussed above. For example, it is not surprising that so many economists failed to see that a bursting of a housing or stock market bubble could crash an overleveraged financial sector and cause a severe recession: too many models ignore the possibility of bubbles or "irrational exuberance" (Shiller, 2015) or assume away the financial sector in the first place (Hayashi and Prescott, 2002).[3]

While some degree of skepticism about models (and shame about the poor state of our theories) is healthy and propels the profession to reject bad models and develop better ones, the mainstream of econometric theory in the past several decades has taken skepticism of economic models and agnosticism about economic assumptions to the *extreme.* This literature, inspired by work of leading figures such as Edward Leamer, Robert LaLonde, and Charles Manski, has led to a culture that regards modeling assumptions as arbitrary, unjustified and therefore bad. The new goal is to try to do empirical work in a way that is as *assumption-free as possible.* However in doing this, the literature ignores the steep tradeoff between modeling assumptions and useful empirical findings. As a result, empirical research that attempts to be assumption-fee is all too often content-free as well. While the saying "garbage in, garbage out" applies to research where incorrect assumptions distort

---

[3]There are promising new econometric methods for detecting asset bubbles such as Phillips *et al.* (2015) that "may serve as useful alert mechanisms to both market participants and regulators in real time." (p. 1), however it remains to be seen whether these new methods will actually be taken seriously and their warnings acted on by regulators of financial markets. The highly influential book "This Time is Different: Eight Centuries of Financial Folly" by Reinhart and Rogoff (2009) remind us that we are quite likely to ignore the lessons of the past and more likely to repeatedly make the same mistakes of failing to see the obvious warning signs of a financial or credit bubble that is likely to burst and cause extended period of real distress, including a recession. They reviewed the nascent "early crisis warning system" literature and concluded that "it can claim only modest success to date" though "there is tremendous scope to strengthen macro-prudential supervision by improving the reporting of current data and by investing in the development of long-dated time series (our basic approach here) so as to gain more perspective on the patterns and statistical regularities in the data." (p. 1241)

our empirical conclusions, we also have "nothing in, nothing out" — if we are unwilling to make any assumptions, then there is very little we can conclude that is of interest empirically (at least for deeper questions such as inferring causality) even with "big data."

Thus, there is relatively little work on *parametric estimation* in econometrics any longer. Instead over the last several decades it has been replaced by a proliferation of published papers on semi-parametric and non-parametric estimation. The message that is implicit in this change in focus of econometric theory is that parametric estimation depends on functional form assumptions that are arbitrary and unjustified, whereas non-parametric and semi-parametric estimators are viewed as closer to the assumption-free ideal. Many of the ideas underlying recent non-parametric and semiparametric estimation methods are imported from the statistics literature, including RCM.

However econometric theory ignores a crucial fact, namely that *models and theories* and more generally *assumptions* are key ingredients of human knowledge, but *models are by their nature simplified approximations to reality.* It is almost self-evident that a model or theory can never strictly be a full, correct description of the world, but at best only an approximation to the deeper complexity out there. Theoretical econometricians are obsessed with the concept of *consistency* which depends on the existence of a *true model* (or true "data generating mechanism") and the assumption that the world is perfectly described by this model.

The Quixotic search for statistical methods that can discover "truth" creates problems for applied researchers, since reality is always far more murky and less obvious than the econometric theorists are willing to admit. It would be much healthier for econometrics to adopt the perspective of *approximation theory* where models are properly regarded as approximations to reality. Instead of obsessing on proving consistency econometrics should focus on providing good metrics for judging how far a model is from the "truth" as well as methods and guidance to enable researchers to create better models that provide increasingly closer approximations to reality.

Another manifestation of the desire for an assumption-free path to empirical knowledge is the popularity of randomized experiments (also called randomized controlled trials or RCT) especially in development economics. It is frequently repeated that RCTs represent the "*gold standard* for evaluation of treatment effects" (see, e.g. *Wikipedia*). Some of the most extreme disciples of RCTs assert that it is even possible to dispense with the need for econometrics — we only need to know elementary statistics to do a simple comparison of mean outcomes for the control and treatment groups. It is not just econometrics that is threatened by such extremism: Tyler Cowen writes on his blog that "I fear that RCTs will . . . lead people to basically stop doing economics."

The extreme agnosticism in econometrics is a double-edged sword: not only has it lead people to reject the use of models because models involve assumptions, it has also lead them to reject *econometric methodology* that also depends on a host of assumptions. In cases where it is absolutely necessary to resort to econometric modeling (such as to analyze observational data that do not come from RCTs) the extreme agnosticists reject the use of economically inspired structural models and prefer instead the less behaviorally explicit statistical models such as the RCM treatment effects framework. As Heckman (1992a) noted, the agnostics who use econometric methods such as RCM that are *perceived* to be assumption-free are actually relying on a host of implicit assumptions that they are either unaware of or unwilling to acknowledge: "In many influential circles, ambiguity disguised as simplicity or robustness is a virtue. The less said about what is implicitly assumed about a statistical model generating data, the less many economists seem to think is being assumed. The new credo is to let sleeping dogs lie." (p. 882).

Given the prevailing skepticism in the profession about existing models and theories and the level of agnosticism/fear about making assumptions, it should not be surprising that structural estimation is not the predominant paradigm for inferring causality in economics. As Heckman (1992a) noted "There is a legitimate basis for the skepticism that greets structural econometrics." (p. 883), whereas Angrist and Pischke (2010) confidently proclaim that as a result of the widespread

use of the more agnostic RCM framework for inferring causality "Empirical microeconomics has experienced a credibility revolution, with a consequent increase in policy relevance and scientific impact." (p. 4).

Ultimately the proof is in the pudding: we can ask "what have we learned?" from all of the econometric methodology we have at our disposal under the several different paradigms for doing empirical work and uncovering causal structure in economics. While it is clearly beyond the scope of this essay to provide a convincing answer to such a sweeping question, the simple answer is that despite our vast arsenal of econometric methodology, we have learned far less about a host of important questions that we would have liked. I will simply note that there is substantial disagreement in the profession and in many areas of the literature about how much these econometric tools have contributed to knowledge.

Even RCTs, the so-called "gold standard" for scientific policy evaluation, have failed to usher in the Angrist-Pischke "credibility revolution" in the area where they have been most extensively applied — development economics. For example Basu (2014), Chief Economist of the World Bank, concludes that "many of the claims widely made on behalf of RCTs are exaggerated and invalid. It will be shown that RCTs do not give proof of any universal causality." (p. 456). Deaton (2010) concludes that "As with IV methods, RCT-based evaluation of projects, without guidance from an understanding of underlying mechanisms, is unlikely to lead to scientific progress in the understanding of economic development." (p. 424). Deaton notes that "Project evaluations, whether using randomized controlled trials or nonexperimental methods, are unlikely to disclose the secrets of development nor, unless they are guided by theory that is itself open to revision, are they likely to be the basis for a cumulative research program that might lead to a better understanding of development." (p.426).

Deaton identifies the fundamental catch-22 — the "nothing in, nothing out" problem with the prevailing extreme agnosticism about the use of models and theories in empirical work: "We are unlikely to learn much about the processes at work if we refuse to say *anything* . . ." (p. 430). So this is the quandary facing empirical researchers: on the one

hand, they are facing an increasingly high burden of proof to establish a *causal effect* as a condition for publication, yet at the same time the prevailing agnosticism with economic models penalizes and discourages them from making reasonable assumptions necessary to make causal inferences.

I have to confess that I am probably only contributing to the prevailing agnosticism by criticizing existing economic theories in this essay, and by pointing out some of the fundamental "limits to knowledge" that pertain to structural models in my review of the monograph of Ken Wolpin (2013), *The Limits to Inference Without Theory* (see Rust, 2014). However I believe it is important to be aware of the limitations of *all* approaches to learning and causal inference, and not oversell or overhype any of them. Instead of adopting a dogmatic approach that posits there is only one right way to do empirical work, it would be better for the profession if we could see different approaches as complements rather than as substitutes. In particular, the more agnostic approaches to empirical work can have an important role in testing and sensitivity analyses to determine whether the assumptions made in structural models are valid or are driving our empirical conclusions. We have much more to gain by working together than by being at war with each other.

While it is important to be aware of fundamental limits to inference, we generally make progress by finding ways to circumvent or overcome these limits. This typically requires making additional assumptions or by focusing on subclasses of problems with "special structure" that makes these problems tractable and enables us to move forward and learn. But we are living in a world of extremism, and there econometric priests with the power to prevent researchers from publishing empirical work that relies on *reasonable* assumptions. As a consequence the extreme agnostic approach to econometrics has severely crippled our ability to learn by preventing or discouraging researchers from making reasonable assumptions and simplifications in empirical work.

An example is the rapidly growing literature on *partial identification* inspired by the influential book by Manski (2003). The objective of this literature is to demonstrate that when we sufficiently weaken various assumptions in our econometric models, the key parameters of interest

in these models are no longer *point identified* and instead are only *set identified.* Regardless of how much data we can collect, if we are sufficiently restricted in the assumptions that we are allowed to make, then the only thing we can conclude is that the quantity or outcome or effect we are interested in estimating is a member of an identified set. The identified sets are often extremely large, and the partial identification literature is rapidly turning into an elaborate catalog of mathematical examples that illustrate the intuitively obvious principle, "nothing in, nothing out." While I believe Manski's work is well-intentioned, and is intended to make applied work more credible by delineating what we can learn from data that is not driven by *a priori* modeling assumptions, in practice the work on partial identification has contributed to a culture of helplessness that prevents us from drawing conclusions even in problems that can be solved relatively easily when reasonable assumptions are allowed.

For example Bontemps *et al.* (2012) study the identification of the parameters β in a standard linear regression model $y = x\beta + \epsilon$ where there are J instrumental variables z where $J \geq K$, and K is the dimension of β. The usual moment restriction is $E\{z'\epsilon\} = 0$, and under the usual rank and order conditions, these assumptions are sufficient to point identify β which can be estimated by the method of ordinary least squares (if x is "exogenous") or two stage least squares (if x is "endogenous"). Bontemps *et al.* (2012) consider whether β continues to be identified when the moment restriction is weakened to $E\{z'\epsilon\} = E\{z'u(z)\}$ where "$u(z)$ is an unknown bounded, scalar function" (p. 1129). Why is this particular alternative weak identification condition interesting or plausible? They note that one leading example where their weaker moment restriction holds is the linear model in situations where the dependent variable y is "censored by intervals" (p. 1130).

The published version of their paper does not include any empirical application, but a 2007 working paper version actually did contain an empirical illustration — a wage regression using the French Labor Force Survey where "respondents can either report their exact labor income or indicate the interval within which their income lies." A

subsample of approximately 23,000 respondents reported a continuous-valued dependent variable y (annual labor income), whereas a separate sample of 2000 individuals only reported which of 10 intervals their labor income fell in.

A natural but *ad hoc* way to deal with interval censored data is to use the midpoint of the interval y is reported to fall in instead of y itself. Another more sophisticated approach is to estimate β by maximum likelihood under a parametric distributional on ϵ such as $\epsilon \sim N(0, \sigma^2)$, i.e. that the unobservable component ϵ is normally distributed with mean 0 and variance σ^2. This results in an "ordered probit" model in the case where we assume $x = z$, i.e. where there is no "endogeneity" or omitted variable bias problem in the wage regression $y = x\beta + \epsilon$. However the ordered probit approach requires a *distributional assumption* so the most super-sophisticated way to estimate β is to make no distributional assumption at all. When we do this, β is no longer point identified but Bontemps *et al.* (2012) show how to estimate the identified set of β.

Bontemps *et al.* (2012) compared the point estimates and 95% confidence intervals for β from the sample where the continuous dependent variable y was observed to the estimates from the sample where only the intervals y fell in were observed. They estimated β by OLS in both cases, but for the interval censored sample they used the midpoint of the interval y was reported to fall in. Their regression point estimates and 95% confidence intervals are very similar for these two samples illustrating that the common sense though *ad hoc* approach of using the mid points of the intervals as a proxy for y did not dramatically affect inferences of β such as the effects of variables such as age or education on earnings. However when they use the agnostic partial identification approach, the identified sets for β were *huge:* "The results of the procedure described in this text is thus striking. The length of the confidence interval increases by a factor of 10 in case of the coefficient of age and slightly less for the coefficient of education. As returns to education are much more precisely estimated, they are still significantly positive in a large range however, from 4.5% to 12%. Any significance of the age coefficient is utterly lost."

This is how the agnosticism of the partial identification approach to econometrics leads to a culture of empirical helplessness: econometric problems that are easy to solve when reasonable assumptions are employed become orders of magnitude more difficult and often completely impossible to solve when we are prohibited from making reasonable assumptions. The answers to even relatively simple questions suddenly become highly uncertain and ambiguous: we can learn very little when we adopt a worst-case mentality and assume almost nothing. If the size of the identified set is already too large to learn much even in the "easiest" problems, it seems unlikely that partial identification will prove useful as a tool to generate new knowledge about the truly difficult questions in economics.

One such question is to measure the causal effect of taxes on labor supply. We do not need the agnosticism of partial identification to know that there huge professional disagreement on this question. For example in "An Open Letter to Professors Heckman and Prescott" Ljungqvist and Sargent (2014) note that "From microeconomic evidence, you Jim inferred that the aggregate labor supply elasticity is low and that therefore high tax rates on labor do little to distort the allocation of resources. From macroeconomic evidence over business cycles, you Ed inferred the opposite." However to resolve these differences, Ljungqvist and Sargent focus on identifying the assumptions and models that Heckman and Prescott can agree on, rather than abandoning models and assumptions entirely.

While Ljungqvist and Sargent acknowledge that "adopting a common framework for modeling individuals' life time work-consumption-savings choices" will "not automatically lead the two of you to agree on the size of the aggregate labor supply elasticity" it seems to me that their approach of attempting to find assumptions and models that we can agree on is a constructive way to make progress. In contrast, consider the agnostic approach to this question taken by Manski (2014) who attempts to minimize the use of models and assumptions. Manski notes that "Empirical studies of labor supply have imposed strong preference assumptions that lack foundation" so his paper "examines anew the problem of inference on income-leisure preferences and considers the implications for evaluation of tax policy." (p. 145).

But what do we learn from Manski's exercise in partial identification? "Based on the analysis in Sections 2 and 3, I conclude that we lack the knowledge of preferences necessary to credibly rank policies. Considering the classical static model, Section 2 showed that basic revealed-preference analysis has little power to predict labor supply under proposed policies. Importantly, it does not predict whether increasing tax rates reduces or increases work effort." (p. 166). While it is another great illustration of the principle of "nothing in, nothing out" it is not clear how this style of research will advance the state of knowledge on this question.

In fact, Manski's analysis represents a step backward in terms of economic modeling, since he analyzes the effect of taxation using a *static* utility framework, whereas whereas labor supply is inherently a dynamic decision. There is widespread agreement in the profession that we need to use dynamic models to capture the causal effect of taxes on labor supply and Manski acknowledges this: "A potentially serious analytical issue is that interpretation of longitudinal data in the manner of Sections 2 and 3 rests on acceptance of the classical static model. Suppose, to the contrary, that one views labor supply in dynamic terms, as in MaCurdy (1985) and elsewhere. Then repeated observations of time allocation provide data on a single sequential choice path rather than data on choice from multiple independent choice sets. Interpretation of the data from a dynamic perspective requires assumptions about the information that persons possess when making choices, the way they form expectations for relevant future events, and the criteria they use to make decisions under uncertainty." (p. 170).

I fully agree with Manski that to understand the causal effect of taxes on labor supply, we need a *dynamic perspective* and interpreting data using this perspective *requires assumptions and models.* But how do we move forward using econometric methods that seek to avoid assumptions wherever possible? It is important to understand how variations in assumptions affect our conclusions, but Sargent and Ljungqvist's open letter to Heckman and Prescott already show us in a much more concrete and relevant way how specific differences in models and assumptions can lead to dramatically different conclusions. They are at the forefront

of the debate, using the relevant models and discussing constructive ways to help move the theory and analysis of this important question forward.

Although Manski's analysis does not advance the frontier of economic theory on this question, he does make constructive suggestions about how to move forward: "One possibility, difficult to achieve, would be for governments to promote exogenous variation in choice sets by decentralizing tax policy or by performing experiments that randomize persons into alternative tax schedules." (p. 149).

Economists used to have more influence in convincing our governments to collect better data. For example Widerquist (2005) notes that "The United States government conducted four negative income tax (NIT) experiments between 1968 and 1980." (p. 95). While they were costly to run, he concludes "we learned a lot from the experiments, not just about the basic income guarantee, but many of the results, such as the income and substitution elasticities of labor supply, that can be useful in other areas." These days it seems that most economists find it easier to invest in methodological sophistication and lament the lack of good data. The professional rewards to devoting your career to data collection are small relative to the rewards to inventing new estimation methods, but the reward structure seems distorted relative to the true social and scientific payoffs from investing in better data rather than in more and more methodology.

While I believe that econometricians such as Leamer or Manski do have a serious interest in improving the state of empirical work, their attempts to make econometrics more assumption-free have contributed to the prevailing agnosticism/skepticism that does not seem to be helping researchers to produce better empirical work. Leamer (1982) questioned the validity of traditional classical econometric inference because it does not properly reflect the results of what he called "specification searches." Manski (2013) questioned the credibility of policy analyses that "rest on critical unsupported assumptions or on leaps of logic." (pp. 2). While these critiques were motivated by valid concerns, they have had unintended effect of demonizing activities that are critical to the scientific process.

In particular, specification searching can be viewed as an informal model selection procedure by which researchers discard models that do not fit the data well, as they search for others models that fit the data better. That is, *specification searches reflect the process by which empirical economists revise their theories/models to better approximate the data they observe.* It seems crazy to demonize this activity because existing econometric theory has great difficulty formalizing how this search process affects our inferences. Similarly, it seems crazy to criticize researchers for the act of inventing models and making reasonable assumptions and other "leaps of logic" because some of these assumptions could be wrong. What Leamer, Manski and their disciples ignore is that while yes, some models and assumptions are bad and can cause us to misunderstand casual relationships and make erroneous policy decisions, science advances through the process of rejecting bad models and searching for new, better ones. To develop the newer, better models researchers need to 1) make assumptions, and 2) engage in specification searches. Discouraging these activities as examples of bad empirical practice actually has opposite effect: it greatly impedes our ability to make scientific progress.

Econometricians would do well to study how the human brain learns and makes causal inferences, because our brains are remarkably good at it. As Griffiths and Tenenbaum (2009) note "People can infer causal relationships from samples too small for any statistical test to produce significant results ... and solve problems like inferring hidden causal structure ... that still pose a major challenge for statisticians and computer scientists." (p. 662). What is increasingly clear is that a key reason why humans are as intelligent as we are is because our brains readily engage in the two key activities that Leamer and Manski question as bad scientific practice: 1) our brains rely on mental models and are constantly making a huge number of assumptions, and 2) our brains are doing the equivalent of "specification searching" to create better mental models when the brain finds sufficient evidence that an existing mental model is problematic. Griffiths and Tenenbaum (2009) conclude their study by noting "We have argued that human causal induction — the inference to causal structure from data — is the result of a statistical

inference comparing hypotheses generated by a causal theory. This approach explains how people are able to infer causal relationships from small samples and identify complex causal structures." (p. 708).

It is certainly true that humans often have poor mental models and sometimes these models are never actually abandoned when the consequences of our internal "model misspecification" or "incorrect assumptions" are not sufficiently serious to adversely affect our day to day life. For example psychologists have identified numerous optical illusions that are clear examples of incorrect assumptions our brain makes all the time. But when there are potentially serious consequences to bad assumptions or a poor mental model, our brains are motivated to expend energy to discard the bad assumptions and create better mental models. This is how humans learn and experiment and via this process we update our theories in light of new data we collect. Our brains do this automatically, and we are often not even consciously aware of it.

For example decades of experiments with infants suggest that humans are born with substantial amount of *a priori* theories and assumptions, such as basic understanding of laws of physics, and ability to do three dimensional visual spatial processing and reasoning, etc. Psychologists refer to this prior knowledge as *core knowledge,* and have shown how humans use experimentation to help them develop better mental models in cases where their core knowledge conflicts with observations. A recent study by Stahl and Feigenson (2015) reported experiments that "tested learning after violations of expectations drawn from core knowledge of object behavior — knowledge that is available from early in life, is universal across human cultures, and is present in other species. . . . our experiments reveal that when infants see an object defy their expectations, they learn about that object better, explore that object more, and test relevant hypotheses for that object's behavior." (p. 94).

Econometric theorists like to convey an impression that statistics and econometrics is a completely "solved problem" and that all of the relevant pieces of a full mathematical model of inductive inference and learning are in place. But as Heckman (1992b) observes, statistics and econometrics are very far from constituting an adequate theory or guide to empirical scientific discovery and are an inadequate and incomplete

theory of how individual scientists and the scientific community at large should optimally learn from data. "Since we do not fully understand the process of discovery or the social nature of the knowledge achieved from this process (agreement in the scientific community) and the role of persuasion in forming consensus, it is not surprising that mathematically precise models of discovery are not available." (p. 883).

It is hard to deny that the human brain is a powerful engine for discovery and learning, and together with oral and written communication that enable us to share these discoveries in our cultures and societies, humanity has made absolutely stunning progress in understanding the causal structure of the world and the larger universe. At its root, all of this knowledge has come from our complex mental models and continuous experimentation. Thus, rather than rejecting this fact and denying the importance of models, assumptions, and specification searching, econometricians could hugely benefit by refocusing econometric theory in a more empirically motivated direction, perhaps something more like *artificial intelligence* or its modern incarnation *machine learning* to see how econometric theory can assist rather than hinder empirical researchers in their attempts to learn about the world.

4

Examples of Useful Causal Inference

In this section I discuss several examples to show how empirical knowledge of causality can be *very* useful. The first example is from medicine, and is a personal example where I credit medical knowledge of causal mechanisms underlying *atrial fibrillation* (A-fib) as potentially saving my life, or at least to help me avoid major brain damage or paralysis due to a stroke. The second example shows how better knowledge of the causal effect of installation of pedestrian traffic countdown timers could save lives of pedestrians, even though they might have the unintended effect of increasing the number of collisions between vehicles.

In January 2012 I had an annual physical exam that included an EKG (electro cardiogram) which measures electrical signals from the heart. The EKG revealed that I had atrial fibrillation, a condition where the upper chamber (atrium) of the heart does not beat normally. Figure 4.1 shows EKG readings for someone with A-fib (top) versus a normal heart rhythm (bottom). A-fib is generally detected in EKGs via the absence of clear *P-waves* (indicated by the purple arrow in the bottom panel of Figure 4.1) which is the electrical pulse that starts the contraction of the atrium prior to the large pulse (called QRS complex) that triggers the contraction of the ventricles (lower chambers of the heart).
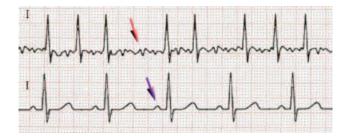
Figure 4.1: Normal vs A-Fib EKGs

According to *Wikipedia* A-fib occurs when "normal regular electrical impulses generated by the sinoatrial node in the right atrium of the heart are overwhelmed by disorganized electrical impulses usually originating in the roots of the pulmonary veins. This leads to irregular conduction of ventricular impulses that generate the heartbeat." A-fib symptoms are often mild (as mine were) so there are millions of individuals who have A-fib but do not know it. A 2013 article in the *American Journal of Cardiology* estimates that 5.2 million people in the U.S. suffered from A-fib in 2010, and it projected this number will increase to 12.1 million cases by 2030. The danger associated with A-fib is that when the atrium is not contracting normally, blood tends to pool in the chamber and clot. When a clot or part of a clot becomes dislodged, it can travel straight to the brain, where it can cause a stroke that can lead to blindness, loss of mental capacity, paralysis, or death. The risk of stroke is 5 times higher for people with A-fib.

When I heard these facts from my cardiologist, of course it came as a shock. The obvious questions immediately came to mind: 1) what caused my A-fib? and 2) how could I get rid of it? Unfortunately, here is where things started to get murky. My cardiologist told me that there are lots of factors that are *associated* with A-fib but no clear causes. Known associations include high blood pressure, heart disease, lung diseases, excessive alcohol consumption, hyperthyroidism, dual chamber pacemakers, and family history of A-fib (30% of diagnosed A-fib patients had a parent who had A-fib).[4]

---

[4] My mother died in 2003 of congestive heart failure. Though I did not know it at the time, my Aunt suffered from A-fib and died of a massive stroke in November, 2013.

After running a series of medical tests including an echocardiogram (an ultrasound of the heart), my cardiologist told me I had no sign of heart disease, I had normal blood pressure and none of the other usual factors associated with A-fib applied to me. So my cardiologist was stumped: he could not offer me any clear explanation of what lead to my A-fib. I asked him if stress could cause it, and he said it was unlikely. So without a clearly identifiable cause, I asked him what my treatment options were. One is the standard response, "live with it" which does not seem too satisfactory. Living with A-fib means taking blood thinners such as Warfarin for the rest of your life to try to reduce the chance of blood clotting in the atrium. I already suffer from a bleeding condition, *von Willibrandt's disease,* so my blood is already "thin" and the prospect of making it even thinner on Warfarin was not too appealing.

Surely there must be other treatments? My cardiologist said yes, there are two main treatments. One is called *cardiac ablation* which can sometimes be done via catheter but can require open heart surgery. The cardiologist essentially destroys or "ablates" some of the *pacemaker cells* that trigger the heartbeat. Sometimes these cells start misfiring and destroying some of them can return the patient to normal heart rhythm. But sometimes ablation does even more damage and the patient has to wear an external pacemaker for their rest of his/her life. This hardly seemed like a satisfactory treatment either.

Was there anything less invasive? My cardiologist told me yes, there is a more moderate treatment called *cardioversion.* This requires hospitalization and sedation because the doctor applied paddles to your chest and subjects your heart to a large electrical shock. This shock convulses the heart and typically returns it to normal rhythm, sort of like rebooting a computer. However the downside is that the mean time to reversion to A-fib after cardioversion is typically two or three months. That is, cardioversion is a treatment that generally has only *temporary effects* and is thus not a permanent cure.

But surely there must be some *cause* that had been overlooked. I asked my cardiologist if he was aware of any other new scientific findings or novel treatments for A-fib. After looking at me blankly for a moment

he asked, "Do you ever have dreams where you feel you are drowning?" The question seemed strange and out of the blue, but come to think of it, yes, I replied that actually I had such dreams rather frequently. My cardiologist said, "OK, I would like to schedule you for an overnight diagnostic visit at the hospital's sleep clinic."

My night at the sleep clinic confirmed that I suffer from sleep apnea. This information was relayed back to my cardiologist, who told me "Here is what I think, John. You have sleep apnea, which means you periodically stop breathing at night when you are sleeping. These periods are long enough to seriously deplete your blood oxygen levels. When this happens your dreams of drowning are a way of waking you up so you don't suffocate. But the frequent low blood oxygen levels could be stressing your heart. It could be a cause of your A-fib. I recommend you get a CPAP machine right away."

The acronym CPAP stands for "continuous positive air pressure" and it is a machine that blows pressurized air into your lungs through a mask. It is not a sexy thing to wear to bed, but if my cardiologist recommended it, I was going to give it a try. While the mask is uncomfortable, it did enable me to breath through my nose which stopped me from snoring at night, something my wife immediately appreciated. But what does sleep apnea have to do with A-fib and would the use of the CPAP machine cure my A-fib?

Gottlieb (2014) discusses very recent research that suggests a strong causal effect between sleep apnea and A-fib, though the exact causal mechanism is not completely understood: "A strong association between obstructive sleep apnea (OSA) and atrial fibrillation has been consistently observed in both epidemiological and clinical cohorts, and multiple studies demonstrate that OSA is associated with an increased risk of atrial fibrillation recurrence following chemical or electrical cardioversion or pulmonary vein isolation by catheter ablation." (p. 1). There are multiple hypotheses about the causal pathways: one hypothesis is that OSA causes enlargement of the atrium which interferes with electrical signals, the P-waves that initiate the contraction of the atrium: "As left atrial diameter is a known predictor of recurrence of atrial fibrillation, cardiac structural changes are plausibly responsible for the elevated rates of recurrence observed in patients with OSA." (p 1).

Having read some of the epidemiological literature on this topic, I am struck by the balance of common sense and statistical expertise; epidemiologists are aware of problems that can confound their analysis, but they make reasonable assumptions and employ pragmatic work-arounds rather than throw up their hands in helplessness. Gottlieb (2014) provides a clear discussion of how confounding problems such incomplete ability to observe OSA, imperfect compliance with the CPAP treatment, lack of RCTs, and lack images of heart structure before and after CPAP treatment make it difficult to identify the precise causal mechanism linking OSA to A-fib. He concludes that "Notwithstanding these limitations, the consistency of the observations in these clinical cohorts contributes to a body of evidence that strongly implicates OSA as a cause, and not merely a correlate, of atrial fibrillation." (p. 1). Thus, despite the challenges and lack of perfect data and RCTs (a typical situation we can relate to as economists), the epidemiological studies have contributed a great deal to our understanding, and this knowledge filtered down to my cardiologist even though at the time (in 2012) it was still recent unpublished research.

How did this knowledge benefit me? My cardiologist had me take Warfarin for about 6 months and told me to use the CPAP machine religiously. Then in August 2012 he performed a cardioversion on me. My heart resumed normal rhythm immediately after the cardioversion procedure. Why did the cardiologist wait for six months to do the cardioversion? Because during that six month period, the CPAP enabled structural changes to occur in my heart, which cardiologists refer to as *remodeling.* As Gottlieb (2014) noted, "In a sample of patients with severe OSA, normal left ventricular ejection fraction, and no history of atrial fibrillation studied with cardiac magnetic resonance imaging before and 6 and 12 months after initiation of CPAP, it was shown that CPAP treatment was associated with a marked decrease in left ventricular mass index and both left and right atrial volume indices." Thus, by the time my cardiologist performed cardioversion on me, my heart had returned to normal size due to the CPAP treatment. With a normally sized atrium, there was a much lower chance that I would revert to A-fib after the cardioversion. In fact, I have been out of A-

fib for over three years now, whereas had I not complied with CPAP treatment, my cardiologist predicted I would have reverted to A-fib in a matter of a few months.

Of course, it is not surprising why I consider this to be an example of "useful science" because the scientific knowledge about sleep apnea and A-fib, imperfect though it may be, clearly benefited my own life in a very direct way. This knowledge was certainly partly due to the sophisticated statistical analysis that the epidemiologists performed, but the vast majority of the credit is due to the technology for measuring and observing the heart: the EKGs, the echocardiograms and magnetic resonance images (MRI), and the technology that lead to the CPAP machines that turned out to be the indirect means of curing my A-fib by treating the sleep apnea condition that appears to be the ultimate cause.

I also give credit to the *Journal of the American Heart Association* for publishing research that was highly suggestive of a causal link between sleep apnea and A-fib but still short of a causal "proof". For example, I have colleagues who have (untreated) sleep apnea but who are not aware of having A-fib. So sleep apnea is not a necessary condition for A-fib, and other causal factors (or pre-existing conditions) might also have to be present for apnea to be a sufficient condition for A-fib. Given the risk of legal liability stemming from medical misinformation, one could imagine that editors of this journal might be exceedingly cautious about publishing results that are not based on the "gold standard" of a RCT. However this is not how science progresses; it is important that researchers can publish findings that may be less than definitive based on the data that they are able to get now, even though the data we have now are not always ideal. Initial suggestive findings could actually be right, and they can pave the wave for subsequent studies that can confirm or disconfirm them.

Can the sort of "useful science" be done in economics? Certainly. Further it is not impossible to find examples that, while perhaps not as personally compelling as the example above, contribute new knowledge that can improve our health, safety, or economic well-being. I would like to discuss another example that I find compelling, which achieves

credibility from a cleverly chosen *research design* and *good data* even though the econometric methodology it employs (simple regression analysis) is quite standard.

Kapoor and Magesan (2014) studied the impact of pedestrian countdown timers that were installed in Toronto, Canada starting in 2006. The timers are designed to improve safety of pedestrians by showing how much time is left before the stoplight in the other direction turns green. This information helps pedestrians determine if they have time to make it across the intersection safely. This study resulted in the following surprising finding, "Although they reduce the number of pedestrians struck by automobiles, countdowns increased the number of collisions between automobiles." and they estimate that "the installation of countdown signals resulted in approximately 21.5 more collisions citywide per month, a more than 5 percent increase over the average without countdown signals." (p. 111).

This is an example of empirical work that cleverly exploits a *natural experiment* and this is an example where the findings are as convincing as those that come from well designed RCTs. Kapoor and Magesan (2014) exploited the fact that countdown timers were installed in Toronto in a staggered fashion over time. The intersections at which they were installed were dictated by cost rather than safety considerations. This enabled them to use a regression matching strategy to "compare nearby intersections with and without a countdown at the same time." (p. 94).

While I am convinced by their analysis, they failed to answer the natural question their finding raises, "why do the countdown timers increase automobile collisions?" Could it be that automobile drivers are using the countdown timers to predict how much time *they* have to get across the intersection before their light turns red, and their sideways glances to try to look at the timers distracts them? As with most scientific findings, new questions are raised as others are answered, and the new questions will probably spur more research to refine our understanding of the problem. But it is important to recognize that there is credible, well-motivated empirical research being published in economics journals that provides new knowledge that is useful to ordinary people. This is an example of useful econometrics and it is an

application of the techniques promoted by Angrist and Pischke (2009). Though I think it is too strong to say that their methods have ushered a credibility revolution in economics, they have proven to be very useful to applied researchers and the popularity of these methods has been a positive development, helping to get economics back on a useful track.

5

Measuring the Causal Effect of Training Programs

In this section I consider the extent to which econometric theory has contributed to useful knowledge on another important topic — evaluating the causal effect of job training programs, also known as *active labor market programs* in Europe. Whether various types of job training can reduce unemployment spells, or provide skills that enable workers to find better matches and have higher subsequent earnings and lower chance of unemployment in the future are clearly vitally important economic questions. Unfortunately, even after decades of research our knowledge of which types of job training actually work is meager.

Heckman *et al.* (1999) note that most OECD countries devote a significant share of GDP to active labor market programs, and "Few US government programs have received such intensive scrutiny, and been subject to so many different types of evaluation methodologies, as has governmentally-supplied job training." (p. 1867). Further, few topics have attracted so much research on *evaluation methodology* as active labor market programs. The enormous investment by the leading econometricians on this topic alone is almost singularly responsible for the massive econometric literature on treatment effects.

Most of the methodological work was spawned by the influential paper by LaLonde (1986) who showed that the treatment effect predicted by the Heckman (1979) selection model differed significantly from the treatment effect in a RCT — the National Supported Work Demonstration. Angrist and Pischke (2010) describe LaLonde's work as a "landmark" that lead other researchers such as "Ashenfelter (1987) to conclude that randomized trials are the way to go." (p. 5).

But was this sweeping conclusion warranted? Heckman *et al.* (1987) and Heckman and Hotz (1989) showed that parametric structural econometric models of self-selection into the training program were able to accurately predict the treatment effect from the National Supported Work Demonstration, contrary to the claims of LaLonde (1986). They showed that LaLonde's failure to do adequate specification testing lead to his erroneous conclusion that parametric models were unable to predict the experimental treatment effects "A simple model-selection strategy based on easily implemented specification tests eliminates non-experimental evaluation models that do not produce estimated program impacts close to the experimental results: the models that are not rejected produce impacts close to the experimental results, at least in the case of women on AFDC." (Heckman and Hotz, 1989, p. 874).

Heckman *et al.* (1999) noted that the estimated effects of the training program were at least as sensitive to the type of *data* used as they were to the *method used to analyze the data.* For example the estimated treatment effects are sensitive to how the outcome, earnings, is measured and are less than half as large when administrative data on earnings were used (from Social Security earnings records) compared to self-reported earnings data. Yet, the profession seems to have used LaLonde (1986) as an excuse to focus on *methodology* when in retrospect we might have learned far more by investing in better *data.* Heckman *et al.* (1999) conclude that "much of the bias reported by LaLonde (1986) in his influential study of the effectiveness of econometric estimators arises from the second source - the inadequacy of the data." (p. 1998).

Thus, in my opinion the main effect of LaLonde (1986) was to set off a "wild goose chase" that resulted in a huge overinvestment in econometric methodology at the expense of investment in better

data and a focus on concrete empirical findings. Econometric theorists published hundreds of econometric theory papers that attempted to devise estimators under "weak assumptions" to avoid the supposed problems with parametric models that LaLonde's study had uncovered. In retrospect, most of the problems Lalonde found were not due to a lack of good econometric methods, but rather were a result of *bad data* and *inappropriate construction of comparison groups.* "A major conclusion of the analysis of Heckman et al. (1998b) is that a substantial portion of the bias and sensitivity reported by LaLonde is due to his failure to compare comparable people and to weight them appropriately." (p. 2007).

The (mis)allocation of professional attention to methodology rather than focusing on collecting better data and identifying promising job training programs came at a significant cost. Heckman won the Nobel Prize for "methods for handling selective samples in a statistically satisfactory way." The Nobel Committee judged that the methods he introduced would help lead to useful knowledge such as to "evaluate the effect of public labor market programs and educational programs, and to estimate the effect of length of unemployment on the probability of getting a job." However the development of the literature on estimating the effect of job training hugely discounted Heckman's contributions by jumping to the conclusion that the normality assumption Heckman (1979) made was responsible for the misleading conclusions that LaLonde (1986) obtained.

The misallocation of research effort is reflected in the Heckman *et al.* (1999) survey chapter itself: their 233 page tome devoted only 40 pages to a survey of empirical results. Their own conclusion was that "Too much emphasis has been placed on formulating alternative econometric methods for correcting for selection bias and too little given to the quality of the underlying data. Although it is expensive, obtaining better data is the only way to solve the evaluation problem in a convincing way. However, better data are not synonymous with social experiments." (p. 1867).

Fortunately, after four decades of research we have at least achieved more of a consensus on econometric methodology, but only on weaker

conclusion that there is no single "best method" for estimating treatment effects. As Heckman *et al.* (1999) acknowledge "there is no universally 'correct' experimental or non-experimental estimator that applies in all contexts. The overwhelming reliance on IV, fixed effects or difference-in-differences and matching estimators in recent research lacks theoretical and empirical justification." Instead, a more sensible recommendation emerged: "Evaluators should use economic theory, the available data and prior information to guide the choice of non-experimental estimators, carefully state the conditions under which counterfactual states are generated, and defend their plausibility" (p. 2085).

The other important conclusion that emerged from four decades of research also seems to be a matter of common sense, namely that progress in this area is at least as dependent on having better data than on the types of econometric methods that are used. Heckman *et al.* (1999) conclude that "no econometric or statistical cure-all fixes the problem of fundamentally bad data. ... The solution to the evaluation problem lies in both the method and the data. The literature on evaluating job training programs has focused largely on methods and not issues of data, taking a passive approach to data collection." (p. 2082-2083).

But what about the actual *empirical conclusions* — the knowledge and insights gained after four decades of intensive study and massive investments? Unfortunately this is the baby that got drowned in the bathwater. The findings that are of practical use to policymakers are disappointingly meager: "Previous evaluations of policies in OECD countries indicate that these programs usually have at best a modest impact on participants' labor market prospects." (p. 1866). "The evidence we summarize also suggests that it is unlikely that even a substantial increase in government-funded training services will significantly improve the skills in the work force." (p. 2080).

While it is useful to know that many of the active labor programs don't work and are ineffective (especially in the context of cost-benefit evaluations), it would do a lot to improve our reputations as "dismal scientists" if we could arrive at less helpless conclusions, and identify specific types of training programs that *do work.* Unfortunately the evidence here is much more mixed and inconclusive. For example Card

et al. (2010) conclude that "Job search assistance programmes yield relatively favourable programme impacts" (p. F452), and "Classroom and on-the-job training programmes appear to be particularly likely to yield more favourable medium-term than short-term impact estimates." (p. F475).

However a U.S. government report, "What Works in Job Training: A Synthesis of the Evidence" (U.S. Department Labor, 2014), concludes that job search training provides only "soft skills" that "speeds up job placement, although it has no long-term employment impact." (p. 14). This report concludes that effective training programs involve more costly and lengthy basic education to provide long lasting "hard skills" for occupations that are in demand: "A post-secondary education, particularly a degree or industry-recognized credential related to jobs in demand, is the most important determinant of differences in workers' lifetime earnings and incomes." (p. 1).

In my opinion, one of the reasons that there has been so little progress in identifying specific types of training programs that actually work, is that the profession has bought in to the overly simplistic focus on estimating the "average treatment effect" of a training program as if it were a single universal parameter. The survey by Heckman *et al.* (1999) identifies the dozens of distinct types of training that fall under the general rubric of "active labor market policies." We see the tendency to oversimplify complex problems in many other areas of economics as well. For example there is a huge literature on estimating the "return to education" as if it were a single time-invariant universal parameter.

It is much more appropriate to think of these as *variables* that differ hugely over time, across individuals, and across schools (or training programs). It makes no sense to try to estimate a variable by pretending that it is a time-invariant parameter. For example it makes little sense to estimate the return on investment of a specific company, such as IBM. We should realize that the return to education, like the return on investment at IBM, is a stochastic process that is continually evolving in complex ways. For this reason alone, we should not expect the literature on training programs to produce any clear, uniform conclusion about the effectiveness of job training. The question is simply ill-posed.

In my opinion, the best current thinking about training programs recognizes the need to tailor them to the specific needs of different individuals in different locations: the most effective training for an 18 year old fatherless gang member in New York City is likely very different than a 54 year old unemployed aerospace engineer in Seattle. The literature seems to ignore the common sense reality that in a rapidly changing economy the types of training that are effective will be changing rapidly too.

Thus, estimating the effect of a training program means chasing a constantly evolving target. It seems unlikely that we will ever have enough observations to arrive at definitive conclusions about the effectiveness of specific *historical* training programs conducted in specific locations with specific groups of trainees. The problem of *external validity* of results will plague this literature. The reality is that we never have enough data and will always have residual uncertainty about nearly any economic issue we study. Nevertheless people and policymakers have to make their best decisions based on the limited information at hand.

To their credit, Heckman *et al.* (1999) recognized "there is substantial heterogeneity in the impacts of these programs. For some groups these programs appear to generate significant benefits both to the participants and to society." (p. 1868). Unfortunately, the focus of most of the econometric literature was not on the pragmatic issue of trying to identify the types of programs that do generate significant benefits to participants and society, but rather to use job training as a methodological testing/proving ground to settle scores on whether experimental or non-experimental econometric methods constitute the "best" way to estimate illusory average treatment effects.

While there is a nagging concern that active labor market programs could have macroeconomic effects, this has also been an understudied area. To a first approximation, sufficiently small scale RCTs will not have any macroeconomic impacts, since the impact of training received by an individual assigned to the treatment group should have no effect on outcomes to an individual who is randomly assigned to the control group — an assumption known as the "Stable Unit Treatment Value Assumption" Rubin (1980) that underlies the validity of RCTs.

However if an effective job training program were to be identified and implemented at a large scale, we can no longer ignore the macroeconomic, general equilibrium effects that such a program can have. In fact, RCTs have been conducted on a sufficiently large scale that they can capture these general equilibrium effects, but the findings are not encouraging: "After eight months, eligible, unemployed youths who were assigned to the program were significantly more likely to have found a stable job than those who were not. But these gains are transitory, and they appear to have come partly at the expense of eligible workers who did not benefit from the program, particularly in labor markets where they compete mainly with other educated workers, and in weak labor markets. Overall, the program seems to have had very little net benefits." (Crépon *et al.*, 2013, p. 532).

But is this conclusion warranted in general? Will any implementation of a training program on a sufficiently large scale result in "winners" (who receive the training) and "losers" (who don't) but result in little if any net benefits? It is far from clear that all types of training are akin to a zero sum game for society as a whole versus programs that increase overall human capital and productivity that are akin to a positive sum game. Finding training programs that result in positive net social benefits may require more use of theory and models rather than exclusive use of large scale RCTs which can be prohibitively costly.

Although our theories and models may be crude, the additional benefit from conducting large scale experiments with some theoretical guidance is the potential to design better experiments that can also help us to improve our models and theories. As Heckman *et al.* (1999) noted: "In bypassing the need to specify economic models, many recent social experiments produce evidence that is not informative about them. They generate choice-based, endogenously stratified samples that are difficult to use in addressing any other economic question apart from the narrow question of determining the impact of treatment on the treated for one program with one set of participation and eligibility rules." (p. 1936).

Thus, my reading of the literature leads me to conclude that LaLonde (1986) and the stampede of RCTs and the RCM treatment effects methodology that resulted from his influential article has not resulted in

a "credibility revolution" for measuring the causal impact of job training programs. Instead these are just additional methods in our toolkit, but ones that are not necessarily any more reliable or trustworthy than the myriad of other methods in the toolkit, including the original, structural, econometric models of Heckman that LaLonde's study criticized.

While the training literature overall can be commended for focusing on a practical and important issue, I fault it for losing touch with the motivating question and transforming the research into a long, and ultimately not so productive methodological goose chase. At the end, common sense should have told us that there is no magic methodological bullet that can provide a definitive answer to the question that can enable us to avoid investing in better data and thinking harder about what is really going on in a training program. We would have made more progress by focusing on how to generate better data and develop better models of how individuals respond to training. These investments would likely have lead to a better understanding of what types of training and pedagogical methods are most helpful to different individuals in different situations.

Before embarking on an unfocused effort to collect newer, better data, it would be advisable to reanalyze the huge amount of data we already have using better models of how people choose training programs and how much effort they devote to these programs including the decision to drop out, resulting in attrition that is typically ignored in many evaluations. Not all individuals who are offered a free or subsidized training opportunity choose to take it, and the information on why they choose not to enroll in the program can have considerable value for inferring the effect of the program.

The literature has widely ignored the importance of measuring and modeling individuals' *expectations* of the benefits to enrolling in a training program, going back to the original work of Heckman (1979). There is much more work on modeling expectations in the literature on college choice, but strangely, this perspective is not typically adopted in the literature on treatment effects of job training programs. It seems clear that the literature can advance by adopting a more structural approach, one that models individuals' expectations and their endogenous choice of

effort, and that allows us to estimate a *distribution of treatment effects* for a range of outcome variables (earnings, unemployment duration, employment duration, job quality, etc) rather than the narrow focus on a single average treatment effect parameter that most of the literature has focused on so far.

Chan and Hamilton (2006) provide an excellent example of how much can be learned by modeling subject behavior in a RCT instead of using the RCT as an excuse to avoid any economic modeling. This particular study focuses on a RCT called ACTG 175, "a landmark randomized double-blind clinical trial designed to evaluate the effectiveness of four alternative therapies for HIV-infected individuals. . . . A notable feature of ACTG 175 was that roughly half of the subjects dropped out (i.e., did not return for followup) by the end of the second year of the trial." (p. 1001–1002).

Chan and Hamilton (2006) note that "Some have questioned the inferences obtained from many randomized experiments, in part because of substantial attrition and noncompliance that plagues many experiments." Instead of thinking more deeply about what the attrition or noncompliance might tell us about subjects' beliefs about the treatment, most existing studies "generally view attrition as primarily a statistical issue that should be addressed using statistical or econometric methods accounting for sample selection." (p. 998). Rather than constituting an annoying problem that is either ignored or dealt with in a superficial manner, the attrition may actually constitute valuable information about subjects' evaluation of the treatment: "as discussed by Heckman and Smith (1998), attrition may be an important indication of how subjects themselves are evaluating the experiment and may reveal information concerning subject preferences toward treatments and outcomes in the experiment." (p. 998).

The result of their analysis is an evaluation framework "based on subject utility rather than solely on the publicly observed outcomes that have typically been the focus of the literature. The standard approach does not capture a variety of features often observed in randomized experiments, such as attrition among subjects receiving the

more effective (in terms of publicly observed outcomes) treatment and changing dropout patterns across treatment arms over the course of the experiment." (p. 1031).

Using their dynamic choice framework, Chan and Hamilton (2006) arrived at very different conclusions about the effectiveness of alternative treatments for AIDS. Previous studies "tout the benefits of combination therapies for the treatment of AIDS, rather than the use of AZT alone, because of the superior impact of combination therapy on patient CD4 counts in the experiment." (p. 1032). However using their utility based model, Chan and Hamilton (2006) found that the combination therapies resulted in negative side-effects for a significant number of patients that lead them to stop taking them. Instead "for a significant fraction of subjects (generally about 18–20 percent), AZT alone yields higher utility than the other treatments.". Further, they found there was substantial subject learning that drove a U-shaped pattern of attrition in the experiment "early dropout is primarily driven by side effects, whereas later attrition reflects declining CD4 counts for many subjects."

The key insight from their approach of *modeling subject behavior in a RCT* is that "Overall, an important implication of our findings, not recognized using the standard evaluation approach, is that patient welfare may be enhanced by offering a menu of therapies, since no single treatment is preferred by a majority of patients." These same insights could help policymakers design more effective job training programs that enables them to select a program from a menu of programs that would be most beneficial given their particular situation, interests, and comparative advantages.

I believe there is a huge potential from building better economic models to analyze subject behavior in experiments. Chan and Hamilton (2006) was not the first to recognize this potential. Other landmark studies include El-Gamal and Grether (1995) who analyzed a laboratory choice experiment and showed that not all human subjects use Bayes rule to make decisions in a simple problem of inference where subjects were asked to choose which of two bingo cages a random sample of colored balls (with replacement) was drawn from. They found substantial subject heterogeneity and though not all subjects used Bayes rule, it was

used by the largest share of students in the experiment relative to other "behavioral" decision rules. Of course, their conclusions required some *a priori* assumptions including the assumption that subjects restricted their choice of decision rules to a particular family of *cutoff rules* that includes Bayes Rule as a special case.

Another important study I would like to mention is Todd and Wolpin (2006) who "demonstrated a potential synergy between social experimentation and observational methods that can be exploited to overcome the limitations of each approach for use in policy analysis." (p. 1408). They used an important RCT called PROGRESA which provided payments to poor families in Mexico for keeping their children enrolled in primary school. Using data from the PROGRESA experiment, they "estimated a behavioral model of family decisions about fertility and schooling without using post-program data on treated households" and "validated the model by comparing its predictions about program impacts to those estimated directly from the experiment. The model produced reasonable forecasts of the effect of the program on school attendance rates of children." Once a sufficiently good model is validated in this manner, it becomes credible to use the model to evaluate a range of other counterfactual policies that were *not* considered in the design of original RCT. Thus, "given a specific policy objective, the model can contribute to the design of an optimal program." (p. 1408).

The point of these examples is that we have far more to gain by integrating a more structural economic model building approach into the analysis of RCTs. Of course doing this requires us to *think more deeply about the problem at hand.* Though this is more difficult, I would assert that by trying to develop better behavioral models and estimating/testing them using both experimental and non-experimental data we are more likely to advance the state of knowledge. Further, having better models often guides us to collect better, more well targeted data to help address the aspects of the problem we are most uncertain about. This is what is done in physics where there is a close complementarity between theorizing and data gathering. It stands to reason that emulating this approach could help us advance the state of useful knowledge in economics.

6

Many Paths to Useful Applied Economic Research

In this section I point out several other examples of useful economic research and emphasize that there is no single path to knowledge, no single methodology that is likely to lead to a "credibility revolution" in economics. I agree with Angrist and Pischke (2010) that credibility is critically important, since without it policymakers are unlikely to pay attention or act on the results of economic research, making it hard to show that it is useful. Where I disagree with Angrist and Pischke (2010) is whether a specific econometric methodology has lead to a "credibility revolution" in economic research.

Instead, credibility comes from demonstrating the value of the knowledge rather than the particular methodology used to obtain this knowledge. In this section I illustrate how knowledge and credibility can be achieved in variety of ways and methods. Perhaps the defining characteristic of researchers who have achieved substantial credibility and whose research has proved to be extremely useful is simply their *attitude* — they are focused on producing useful knowledge, they have a genuine concern and interest in the topics they are studying, and they are willing to do what it takes to improve our understanding using the most appropriate methods. That is, the most useful research seems to

be *topic driven* rather than *method driven.* While I don't deny that most important contributions are made possible because they build on an existing huge body of scientific knowledge (which include methods), whether new knowledge will have practical applicability depends more on the attitude and the motivation of the researcher than on the particular methods they used to discover and communicate this knowledge.

I believe an excellent example of this is Alvin Roth, who was awarded the Nobel Prize in 2012 for his research on market design, particularly for what he calls "practical market design." In addition to his theoretical contributions, the Nobel Prize committee noted that "Roth has also developed systems for matching doctors with hospitals, school pupils with schools, and organ donors with patients." Roth's practical orientation, or what I would refer to as his *attitude* toward research, is clear from his writings and the type of work he has done over his career. For example while most game theorists are content to write articles that are confined to the world of abstract mathematics, Roth (1991a) wrote that "the real test of our success will be not merely how well we understand the general principles that govern economic interactions, but how well we can bring this knowledge to bear on practical questions of microeconomic engineering" (p. 113).

It is the "engineering mentality" that distinguishes Roth's contributions. This leads to a more active approach to knowledge than the more passive approach taken in much of economic and econometric theory: it is not enough to understand how the world currently works, the goal of the market design theory is to develop new institutions and mechanisms that create a better world. But unlike most theorists, who are content to write theories that assert that it is *possible* to create a better world, Roth has distinguished himself by showing how to apply these theories to actually create these better worlds.

Roth's most important contributions were motivated by an understanding of the world, specifically the observation of "common market failures we have seen in recent work on more senior medical labour markets, and also on allocation procedures that do not use prices, for school choice in New York City and Boston, and for the allocation of live-donor kidneys for transplantation." (Roth, 2008, p. 286). However

the value of this knowledge was manifested through design of better institutions, and Roth has sufficient credibility to convince policymakers to implement better mechanisms and thereby demonstrate that the theories that lead to these better mechanisms are relevant and useful.

Methodologically, Roth has contributed useful empirical knowledge though laboratory experiments, field experiments and "natural experiments" rather than using the latest econometric estimators. For example Roth (1991) studied the use of matching algorithms adopted by medical schools in the UK and compared them to a different matching system used in the U.S. to allocate interns to residencies. He showed how economic theory correctly predicted that the matching algorithms adopted by the UK medical schools were *unstable* and the theory predicted that these systems would ultimately "unravel" — which they did. On the other hand Roth showed that the matching algorithm used by U.S. medical schools is stable, and unlike the UK, the system in the U.S. did not unravel and continues to operate today.

But Roth's most impressive contributions came from his success in designing better matching mechanisms that were implemented in practice that noticeably improved outcomes in a variety of real world settings. This is where theory plays a critical role since blind experimentation or following gut instincts about what a better mechanism might be is not feasible given the costs and obstacles to making changes in the *status quo.*

One example is kidney exchanges. Roth (2008) describes a dysfunctional state of kidney transplants in the U.S. with over 70,000 needing them in 2006, but a "grave shortage of transplantable kidneys." Most kidneys come from cadavers, and there is no formal "market" in transplantable kidneys" due to ethical concerns. Roth's approach to this is pragmatic: "how to increase the number of transplants subject to existing constraints, including those that forbid monetary incentives." (p. 291). Via *theoretical analysis* "Roth *et. al.* (2004) showed that in principle a substantial increase in the number of transplants could be anticipated from an appropriately designed clearinghouse that assembled a database of compatible patients."

But rather than stop there (which is where most economists stop, assuming that someone from the "real world" will actually read and implement their ideas), Roth "sent copies of that paper to many kidney surgeons and one of them, Frank Delmonico (the medical director of the New England Organ Bank), came to lunch to pursue the conversation. Out of that conversation, which grew to include many others (and led to modifications of our original proposals), came the New England Program for Kidney Exchange, which unites the fourteen kidney transplant centres in New England to allow incompatible patient donor pairs from anywhere in the region to find exchanges with other such pairs."

Roth actively participated in the setup and operation of this exchange program, resulting in significant improvements: "By building a database of incompatible patient-donor pairs and their relevant medical data, it became possible to arrange more transplants, using a clearing-house to maximise the number (or some quality or priority-adjusted number) of transplants subject to various constraints." (p. 292). The kidney exchanges have been extended to include kidney transplants from live donors and now over 10% of all such transplants are arranged via kidney exchanges.

A recent article in the *New York Times Magazine* (Wollan, 2015) noted that currently over 100,000 people are on the waiting list for a kidney, and in 2014 "4,720 people died waiting" but sophisticated computer matching algorithms such as Anderson *et al.* (2015) are able to "find the best medical matches — thus increasing the odds of a successful transplant — by decoupling donors from their intended recipients. . . . The software ranks those possible pairings based on hundreds of different immunnological, genetic and demographic citeria, while also aiming to create longer chains of harder-to-match people which will ultimately result in more transplants." (p. 59). This is not only useful research, it is research that is *saving lives.*

Another, perhaps less dramatic but still extremely important example of Roth's useful research contributions is his work on school matching. On December 5, 2014 the *New York Times* published an article "How Game Theory Helped Improve New York Citys High School Application Process" that described Roth's work to improve the system

that the New York City Public School System used to match students to schools. "In the late 1990s, for instance, tens of thousands of children were shunted off to schools that had nothing going for them, it seemed, beyond empty desks. The process was so byzantine it appeared nothing short of a Nobel Prize-worthy algorithm could fix it. Which is essentially what happened. About a decade ago, three economists Atila Abdulkadiroglu (Duke), Parag Pathak (M.I.T.) and Alvin E. Roth (Stanford), all experts in game theory and market design — were invited to attack the sorting problem together. Their solution was a model of mathematical efficiency and elegance, and it helped earn Professor Roth a Nobel Memorial Prize in Economic Science in 2012."

The article concludes that "Before the redesign, the application process was a mess. Or, as an economist might say, it was an example of a congested market." But after Roth *et. al.*'s new matching system was implemented, "In 2004, the first year that students were sorted in this way, the number who went unmatched plummeted, from 31,000 in 2003 to about 3,000 — still a lot of disappointed teenagers. That year, and every year since, the algorithm has assigned roughly half of all students to their first choice schools; another third or so have been assigned to their second or third choices."

So what was Roth's path to knowledge and influence in economics? It was not econometrics, but rather it was "primarily game theory, backed by experiments and close observation of institutions that underlies the work we do" (personal communication). Thus, Roth's work is an excellent example of what can be obtained from more basic tools: an excellent command of economic theory, observation, and experimentation. By following this path to knowledge, Roth (1991a) forecasted that "a hundred years from now, game theory will have become the backbone of a kind of micro-economic engineering that will have roughly the relation to the economic theory and laboratory experimentation of the time that chemical engineering has to chemical theory and bench chemistry."

There are many other examples of useful economic research, and some of the important contributions *do* depend relatively sophisticated econometric and computational methodologies. One example is Misra and Nair (2011) who who estimated a dynamic structural model of the

sales effort of a sample of contact lens salespeople. They showed that the company had adopted a suboptimal compensation plan consisting of salary, quota, and bonus that inefficiently motivated its sales force. Their structural model revealed that the company's combination of a sales quota and maximum commission ceiling introduced a particular inefficiency; namely, the most productive salespeople would slack off once they had reached the commission ceiling. Using the estimated structural model, they used economic theory to design an improved incentive plan that reduced the sales quota and removed the commission ceiling. The company actually implemented their recommended alternative compensation scheme. "Agent behavior and output under the new compensation plan is found to change as predicted. The new plan resulted in a 9 percent improvement in overall revenues, which translates to about $12 million incremental revenues annually, indicating the success of the field-implementation. The results bear out the face validity of dynamic agency theory for real-world compensation design. More generally, our results fit into a growing literature that illustrates that dynamic programming-based solutions, when combined with structural empirical specifications of behavior, can help significantly improve marketing decision-making, and firms' profitability." (p. 211-212).

I wish to discuss a final example, that illustrates that research does not have to be inspired by economic theory, use the latest econometric techniques, and may not (yet) have inspired real-world policy changes to nevertheless be very useful. I would classify the work of Hoxby and Turner (2013) in this category. Based on previous work the authors recognized that high achieving but low income high school students are substantially less likely to apply to selective colleges based on an incorrect belief that these colleges are out of their price range. The reality is that generous financial aid programs at the most selective colleges makes them price-competitive or even cheaper than the less selective colleges (including community colleges) that these students are most likely to apply to.

Hoxby and Turner (2013) "use a randomized controlled trial to evaluate interventions that provide students with semicustomized information on the application process and colleges' net costs. The interventions

also provide students with no-paperwork application fee waivers. The ECO Comprehensive (ECO-C) Intervention costs about \$6 per student, and we find that it causes high-achieving, low-income students to apply and be admitted to more colleges, especially those with high graduation rates and generous instructional resources." (p. 1). Even though this type of intervention has not actually been adopted by selective colleges to help recruit better students, and though the intervention may have been motivated more by intuition and previous empirical studies than a formal economic theory, the study is nevertheless an example of useful economic research, and one that was achieved through standard research methods. The payoff to this study came from the degree of care and effort that the authors expended to carry out this large scale experiment, rather from using fancy theories or econometric methods.

7

Towards Mostly Useful Econometrics

In 1986, I spent a semester as a visiting assistant professor at the University of Minnesota. While I was unpacking some of the books I had brought with me to the shelves in my office, I noticed a man standing in the office doorway, glaring at me. He was macho but slightly balding, wearing blue jeans, sneakers, and a rugby shirt: I thought he might be a local soccer coach. I tried to introduce myself, but he refused to shake my hand, and kept his arms crossed over his chest while he continued to glare at me, disapprovingly. I had just put my Theil ([1971](#)) *Principles of Econometrics* on the shelf (the book I used it to learn econometrics as a PhD student at MIT), when he suddenly grabbed it and rifled it, holding the two covers and fanning the pages back and forth violently until I feared that it would split down the spine. When I asked him what he was trying to do, he threw the book in my trash can and said, "*you read this crap?*" and abruptly turned and walked away.

Later, I learned the identity of that man: Edward C. Prescott. I also learned this incident was not a prank: Prescott really does hate econometrics and thinks that most of it is crap. No doubt he would rejoice in the complete demise of econometric theory. Unlike Prescott, I don't hate econometrics and I actually think that a lot of it is extremely

useful. However it is important to consider why leading members of the profession such as Prescott have such a negative attitude towards econometric theory.

This essay suggests a possible reason: the antipathy that Prescott and others have towards econometrics (and one that I share) may be due to the methodological/intellectual elitism and the lack of concern that many econometric theorists appear to have about economics and the practical application of the methods they propose. While I do think that some of the "calibration" methods that Prescott champions for confronting models to data are inferior in many respects to closely related "method of moments estimators" that have been extensively studied by econometricians, Prescott received a Nobel prize, not for his methodological work on calibration, but rather for his seminal contributions to *economic knowledge.* Prescott focused on important economic questions rather than spending his time on mathematical techicalities about how best to estimate various unknown parameters in his models. This focus enabled him to make important *practical* contributions to knowledge in economics.

Why should anyone care about my personal concerns about the direction theoretical econometrics is going? Presumably researchers are well informed and can choose for themselves the best path to follow, and make their own judgements about whether there is a higher professional payoff to doing econometric theory relative to trying to collect better data and doing empirical work. Regardless of whether I have much personal credibility in the economics profession I feel it is incumbent on someone to point out that the culture in economics has lead to distorted incentives by rewarding pure mathematical theory skills much more highly than data gathering and practically oriented empirical skills. As a result the profession has become increasingly detached from practical realities and is increasingly in danger of being regarded as irrelevant.

Regardless of what I think, the financial pressures facing higher education may ultimately force the economics profession to downsize or streamline and produce research that is more clearly useful, especially from the perspective of legislators allocating scarce government funding. For example, a December 2014 report by the American Economic

Association noted that "Last year economic research continued to be threatened with cutbacks in funding for research grants and economic data. . . . The House Science & Technology Committee passed legislation that would reallocate almost half of the 2014 budget for the Social, Behavioral, and Economic (SBE) Directorate, the source of almost all funding for economics at NSF, to other science and engineering Directorates."

Economists would like to believe that economics is closer to a "hard science" than one of the "soft sciences" such as psychology, sociology, or political science, so that the usefulness of economics is more self-evident, making it more immune to budget cuts. For example, Fourcade *et al.* (2015) note that "There is an implicit pecking order among the social sciences, and it seems to be dominated by economics. For starters, economists see *themselves* at or near the top of the disciplinary hierarchy." (p. 1). However, the hard line conservatives who are currently determining the allocation of research funding in the U.S. have failed to recognize "the superiority of economists" and treat economics as another one of the other soft social science that fail to produce research that is critical to the future of U.S. technology and competitiveness. For example, Lamar Smith, the Republican Chair of the House Space, Science and Technology stated that "Several recent studies conducted by NSF's SBE funding have been of questionable value, and something our nation can ill-afford. These SBE funds are better spent on higher priority scientific endeavors that have demonstrated return on investment for the American taxpayer."

In addition to cuts in research funding, many universities in the U.S., particularly public universities, are facing very tough financial circumstances in the aftermath of the Great Recession: reductions in state funding of public universities of 25% or more have been common in many states over the past five years. This could lead to situation where econometric theorists (as well as economic theorists) become a luxury that many universities can no longer afford. Most universities have *many* statistically trained faculty, but they are spread throughout various departments and are valued for their connection to discipline-specific applications: biometricians, sociometricians, econometricians,

etc. However due to declining enrollments and financial pressures, there is an increasing pressure to close down the pure statistics departments if they do not have sufficiently close connections to particular applied domains. As I noted in the introduction Princeton University closed down its statistics department and Yale University nearly eliminated its statistics department in the 1990s. Unless theoretical econometrics finds a way to have a clearer value-added to the applied fields in economics, the same sort of pressures are likely to lead to a big reduction in slots for pure econometric theorists in coming years.

Finally, there is a problem that I would refer to as *methodology fatigue* within the economics profession itself. I have already discussed this when I mentioned the increasing number of economists in a number of fields to reject econometrics altogether and simply rely on RCTs as their preferred path to knowledge. Many econometric theorists are oblivious to the computational demands their estimators impose on empirical researchers, and (having read too many econometric theory papers) it is also evident that they are not concerned with communicating the essential ideas and benefits/costs of a new estimator in a fashion non-specialists can easily understand. Most econometric theory papers are highly technical and appear to require a PhD in mathematics or statistics to fully understand all the assumptions and results, much less the proofs of results that are crammed into dense 20 page appendices. Leading economists already seem to have enough difficulty making valid causal inferences due to failure correctly implement standard econometric techniques such as weighted least squares, fixed effects estimators, or even to capture all relevant data from an Excel spreadsheet![5]

---

[5]McCrary (2002) showed that Levitt (1997) improperly inverted weights in a weighted least squares estimation that lead Levitt to draw an erroneous conclusion that there is an electoral cycle in the hiring of police. Foote and Goetz (2006) showed that Donohue and Levitt (2001) did not include state fixed effects as they had claimed, which nullified their conclusion that abortion "causes" a significant reduction in crime. Finally, Herndon *et al.* (2014) showed that an influential paper by Reinhart and Rogoff (2010) suffered from "coding errors, selective exclusion of available data, and unconventional weighting of summary statistics" in the Excel spreadsheets used for their analysis, obliterating Reinhart and Rogoff's finding that GDP growth significantly declines once a country's external debt to GDP exceeds 90%.

Though I am not arguing that we need to "dumb down" econometrics because some leading researchers have misused elementary econometric methods, we do need to recognize that all of us have a finite mental capacity and the more of our mental space we devote to thinking about arcane methodological details, the less space remains for doing the actual economics and analyzing empirical results. Thus, sometimes "less is more" and the profession might benefit from some house cleaning to help younger researchers focus on the econometric methods they really need to know, rather than burden them with learning a mass of other estimation methods that have proved far less useful in empirical work. Given the low citation counts of many econometric theory papers, some of this seems to be happening automatically.

Thus it seems it is time for a wake up call to the economics profession. It will be interesting to see if it can respond more proactively to preserve its "superior" position and perks (including the much higher salaries that economists earn relative to other social sciences) before they are simply taken away from us. I argue that the single most critical task is for economists to do a better job of showing that we can produce useful knowledge. While it would be great if economists could tackle and solve some of the hardest problems in a heroic effort to show we are capable of "big science," in my own case I choose to follow a more humble path and try to do "little science" and analyze simpler problems where the value-added from the science and models can be convincingly demonstrated. One of the paths I find most exciting is to combine structural model building and testing and laboratory and field experimentation, following the strategy of El-Gamal and Grether (1995), Chan and Hamilton (2006), Todd and Wolpin (2006) and Misra and Nair (2011) discussed in Section 5.

While I have argued that sufficiently complete economic models provide a theoretical laboratory that enable us to conduct "thought experiments" to rigorously study questions of causality and to undertake counterfactual policy experiments (or do mechanism design), we can never be completely certain whether our models are sufficiently good approximations to reality to assert that the causal effect that exists in our model, or the counterfactual policy prediction from the model,

or the optimal mechanism implied by the model will actually apply or work well in reality. Yet in Section 5 I showed that structural models are capable of providing good predictions of the treatment effect in a RCT, and in Section 6 I showed that economic theories are sufficiently good to enable economists to design new mechanisms or institutions that do result in measurable practical improvements over the *status quo*. While there are some modest successes we can point to, to avoid the hubris that I discussed in the introduction, it is better to be more humble and keep in mind the view of Popper (2002) and treat all models as working hypotheses that could always be rejected after observing new data or conducting additional experiments, but retaining a faith that we can make great progress despite this ever present uncertainty.

The insights from neuroscience about how the human brain works, along with the example provided by physics where there is a close integration between theorizing and experimentation, make me confident that there is a promising future for doing useful science by trying to do a better job of integrating economic theory into empirical work, in much the fashion Alvin Roth suggested in the quote in Section 6, and from the example he has set in his own research. The role of good leadership to help move the profession in a better direction cannot be underestimated. At the same time, another important message from this essay is that there is no single superior path to knowledge in economics, and some of the most productive and useful paths involved almost no econometrics at all. Thus, it is ultimately a waste of time to try to prove that one path is better than another. I would like to trust in a *laissez-faire* approach where each researcher chooses the path they feel will be most productive for them. Researchers are intelligent and informed, and will choose the methods that provide the biggest payoff. If the professional incentives are well structured, this should result in researchers choosing the paths that generate more useful knowledge.

Unfortunately professional incentives are not well structured, and have resulted in a sort of "market failure" that has lead researchers in economics (guided by their own instincts) to do excessively mathematical and theoretical types of research that has failed to generate as much useful knowledge as we would like. Like all researchers, I do the best I

can and follow the path that makes the most sense to me. Beyond this, one can only suggest some changes in the profession to try to change attitudes and incentives, though it is less clear what concrete actions can be taken to make these changes come about. But for what it is worth, I list several ideas below.

1. **Increase the return to data gathering and dissemination,** such as treating the creation of new data sets and data dictionaries and online databases and other tools to make data more accessible to other researchers as equivalent to additional publications that count towards tenure.

2. **Increase the return for practical applications of economic ideas,** such as giving "super cites" (i.e. citations that count more than several publications in top ranked journals) to practical applications of economic ideas such as measured by award of patents, creation of successful software, startups of successful non-profit or for-profit ventures, or public service that leads to implementation of academic ideas in public policy (being appointed as Chief Economist of the FCC and implementing a new policy that improves the quality of communications infrastructure, etc.).

3. **Improve the incentives to demonstrate empirical applicability and relevance of new econometric estimators,** such as journals requiring that new econometric estimators be applied to an actual data set and the new estimator is shown to make a difference in conclusions of practical interest as a precondition for publication. This will tend to reduce the amount of journal space devoted to econometric theory that is not clearly useful.

4. **Increase the incentives for productive collaborations between theorists and applied researchers,** emulating the way research is done in physics. This can be done similarly to suggestion 3 above by having journals impose a higher standard on pure theory papers, by giving preference to publishing pure theory papers that are accompanied by an empirical application.

5. **Reduce unnecessary roadblocks to publication of empirical work,** by not subjecting empirical work to an unrealistic burden of proof such as requiring authors to establish a causal mechanism beyond the shadow of a doubt, or obliging empirical researchers to demonstrate that the observations can be uniquely "explained" by a theoretical model, accompanied by a proof that this model is non-parametrically identified.

6. **Improve our ability to evaluate competing estimation methodologies,** by structuring competitions to estimate quantities of interest in artificial environments where the true data generating mechanism is known to the organizer of the competition but not to the teams participating in the competition, similar in spirit to the annual *Econometric Games* held in Amsterdam.

7. **Improve the empirical relevance of theoretical econometrics,** by encouraging collaborations between econometricians and neurosciensts to better understand how humans and highly skilled economic researchers make causal inferences, form economic models, test and reject models or theories, and to understand lessons this can have for econometric theory and machine learning. This can be done by leading journals by lowering the barriers to publication of work that cross disciplinary boundaries, and raising barriers to publication of work that is more narrowly in the mode of traditional econometric theory that primarily draws on ideas from pure statistics.

8. **Help mathematically gifted econometricians to find more productive uses of their skills,** such as applying their econometric skills to more applied areas where the marginal return is higher, e.g. computational economics.

The idea of creating econometric games (with high stakes, to encourage participation by leading econometricians) to evaluate competing estimation methodologies seems to be particularly important to me. So much professional energy is wasted in battles that are carried out in competing published articles, each arguing that their method is "best"

in particular circumstances. It seems to me that a much more effective test is to "put your money where your mouth is" and use your preferred method to find the "answer" in a situation where the correct answer (e.g. the true causal effect or causal mechanism) is known *a priori* because the data used to infer the causal effect are *artificial data generated from a model where the correct causal effect is known in advance.* I use this type of competition to teach my graduate students: I provide them with data generated from a sufficiently complicated model that it appears to be "realistic" and ask them to use whatever econometric method they like to try to answer a well defined, practical question (e.g. estimate an average treatment effect or use their prefered model to make a counterfactual prediction).

I conclude by emphasizing again that I am not out to demonize theoretical econometrics, or to create another counterproductive dogma: namely, that *all research must have practical value.* We know that some of the most important research contributions initially started out seeming to be interesting but "useless ideas" but later turned out to have very important practical applications.[6] Inevitably there will be some "useless" research published in our journals, just as our DNA is populated by long stretches of apparently useless "junk DNA." But the "survival of the fittest" principle of evolutionary theory does teach us that research that has practical value is more likely to be rewarded by the "system" by directing more resources and energy towards helping to produce more research that is useful to society.

Thus, if economics is to continue to survive and prosper, it is important for at least a few of the "elders" in the profession to focus on finding sufficient practical applications of the science, while protecting the youngest and most mathematically gifted scientists by enabling them to follow their interests without the pressure of having to justify their research in terms of its immediate practical applicability. In this way we can establish an academic "food chain" that enables mathe-

---

[6]An example that comes to mind are *buckey balls* and artificially synthesized spherical carbon molecule that resembles the geodesic domes created by Buckminster Fuller. First synthesized in 1985 with no apparent practical use, these molecules have subsequently found to be very useful in the creation of high temperature superconductors.

matically gifted researchers to focus on their comparative advantage, with the knowledge that there are enough "worker bees" further down in the food chain who can eventually use some of their ideas in a productive fashion. But it is critical for all parts of this food chain to be rewarded appropriately. Currently, the rewards to doing econometric theory are disproportinate in terms of its value to the overall food chain in economics.

It is clear to me that with even modest changes in professional incentives and rewards, that there are many gifted young econometric theorists who would be willing to slightly modify their research agenda and use their skills and talents to work on topics that they know are in demand and will have practical payoffs but still involve the use of high level mathematics. In particular, econometric theorists should not be forced to do empirical work if they dislike it or conclude it is too far away from their comparative advantage. But some of them should consider whether their theory skills could be put to more productive use in related areas rather than to continue to slog away doing econometric theory if they agree with me that this is an area that is running into sharply diminishing returns.

I see plenty of demand for the services of skilled econometric theorists in computational economics, for example. Many of the algorithms used by applied researchers to solve structural models can be recast as "estimators" that converge to "true values" as certain parameters (e.g. the number of grid points or quadrature points in an approximation or numerical integration algorithm, or the number of basis functions in a sieve approximation scheme) tend to infinity. With small adjustments, the arguments theoretical econometricians use to prove the consistency of parametric and non-parametric estimators could be adapted to establish the convergence (or failure of convergence) of many of the algorithms that are commonly used in computational economics to solve models, but which lack adequate theoretical justification. An example is the "stochastic algorithm" of Pakes and McGuire (2001) whose convergence is demonstrated by stochastic simulation of a specific example, but for which a general convergence proof (similar to a proof of consistency of an econometric estimator) is lacking.

Another important uncharted area is to establish the asymptotic properties of structural estimators that account for both sampling error and the numerical approximation error involved in using a variety of algorithms to solve agents' dynamic programming problems that are embedded as subproblems of the overall estimation problem (see, e.g. Kristensen and Schjerning, 2015).

Thus, I believe that it is possible for the economics profession to refocus itself in a way that does not involve radical, wrenching departures from our preferred research agendas, but which can ultimately enable econometric theory and economics as a whole to produce more useful economic knowledge. Mostly what is necessary is a change in professional culture, attitudes and the incentives for doing theory versus applied research. While I do not see any particular path or methodology that has or is likely to lead to a "credibility revolution" in economics in the near future, I feel that with a gradual change of course, the future outlook for economics will be brighter than it currently seems with all the depressing funding cutbacks and financial pressures that universities are facing. In my own experience credibility is not something that is easy to acquire. It is like a stock of capital that is built up slowly, and at substantial cost. And similar to capital, credibility can easily be squandered. The economics profession has squandered some of its credibility in the events leading up to the Great Recession and its aftermath. Whether economists can regain this lost credibility will depend a lot on how successful we are in coming decades in producing more useful knowledge that can be recognized and appreciated by ordinary citizens.

Acknowledgements

References

Anderson, R., I. Ashlagi, D. Gamarnik, and A. E. Roth. 2015. "Finding long chains in kidney exchange using the traveling salesman problem". *Proceedings of the National Academy of Sciences.* 112(3): 663–668.

Angrist, J. and J.-S. Pischke. 2009. *Mostly Harmless Econometrics An Empiricist's Companion.* Princeton, Princeton University Press.

Angrist, J. D. and J.-S. Pischke. 2010. "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics". *Journal of Economic Perspectives.* 24(2): 3–30.

Ashenfelter, O. 1987. "The case for evaluating training programs with randomized trials". *Economics of Education Review.* 6(4): 333–338.

Basu, K. 2014. "Randomisation, causality and the role of reasoned intuition". *Oxford Development Studies.* 42-4.

Bontemps, C., T. Magnac, and E. Maurin. 2012. "Set identified linear models". *Econometrica.* 80(3): 1129–1155.

Card, D. and S. DellaVigna. 2013. "Nine facts about top journals in economics". *Journal of Economic Literature.* 51(1): 144–161.

Card, D., J. Kluve, and A. Weber. 2010. "Active labour market policy evaluations: A meta-analysis". *Economic Journal.* 120(548): F452–F457.

Carrell, S. E., B. I. Sacerdote, and J. E. West. 2013. "From natural variation to optimal policy? The importance of endogenous Peer group formation". *Econometrica.* 81(3): 855–882.

Cassidy, J. 1996. "The decline of economics". *New Yorker.* December(2).

Cassidy, J. 2010. "Interview with Eugene Fama". *New Yorker.* January(13).

Chan, T. Y. and B. H. Hamilton. 2006. "Learning, private information, and the economic evaluation of randomized experiments". *Journal of Political Economy.* 114(6): 997–1040.

Crépon, B., E. Duflo, M. Gurgand, R. Rathelot, and P. Zamora. 2013. "Do labor market policies have displacement effects? Evidence from a clustered randomized experiment". *Quarterly Journal of Economics.* 128(2): 531–580.

Cunha, F., J. J. Heckman, and S. M. Schennach. 2010. "Estimating the technology of cognitive and noncognitive skill formation". *Econometrica.* 78(3): 883–931.

Deaton, A. 2010. "Instruments, randomization, and learning about development". *Journal of Economic Literature.* 48: 424–455.

Donohue, J. J. I. and S. D. Levitt. 2001. "The impact of legalized abortion on crime". *Quarterly Journal of Economics.* 116(2): 379–420.

Ellison, G. 2013. "How does the market use citation data? The Hirsch index in economics". *American Economic Journal: Applied Economics.* 5: 63–90.

Foote, C. L. and C. F. Goetz. 2006. "The impact of legalized abortion on crime: Comment". *Quarterly Journal of Economics.* 123(1): 407–423.

Fourcade, M., E. Ollion, and Y. Algan. 2015. "The superiority of economists". *Journal of Economic Perspectives.* 29(1): 89–114.

El-Gamal, M. A. and D. M. Grether. 1995. "Are people Bayesian? Uncovering behavioral strategies". *Journal of the American Statistical Association.* 90(432): 1137–1145.

Gottlieb, D. J. 2014. "Sleep apnea and the risk of atrial fibrillation recurrence: Structural or functional effects?" *Journal of the American Heart Association.* 113(654): 1–3.

Griffiths, T. L. and J. B. Tenenbaum. 2009. "Theory-based causal induction". *Pscyhological Review.* 116(4): 661–716.

Hamermesh, D. S. 2013. "Six decades of top economics publishing: Who and how?" *Journal of Economic Literature.* 51(1): 162–172.

Hayashi, F. and E. C. Prescott. 2002. "The 1990s in Japan: A lost decade". *Review of Economic Dynamics.* 5: 206–235.

Heckman, J. J. 1979. "Sample selection bias as specification error". *Econometrica.* 47(1): 153–161.

Heckman, J. J. 1992a. "Haavelmo and the birth of modern econometrics: A review of the history of econometric ideas by Mary Morgan". *Journal of Economic Literature.* 30(2): 876–886.

Heckman, J. J. 1992b. "Randomization and social policy evaluation". In: *Evaluating Welfare and Training Programs.* Ed. by C. F. Manski and I. Garfinkel. Harvard University Press. 201–230.

Heckman, J. J. 2005. "The scientific model of causality". *Sociological Methodology.* 35: 1–97.

Heckman, J. J. and V. J. Hotz. 1989. "Choosing among alternative non-experimental methods for estimating the impact of social programs: The case of manpower training". *Journal ofthe American Statistical Association.* 84(408): 862–874.

Heckman, J. J., V. J. Hotz, and M. Dabos. 1987. "Do we need experimental data to evaluate the impact of manpower training on earnings?" *Evaluation Review.* 11: 395–427.

Heckman, J. J., R. J. LaLonde, and J. A. Smith. 1999. "The economics and econometrics of active labor market programs". In: *Handbook of Labor Economics, Volume 3.* Ed. by O. Ashenfelter and D. Card. New York: Elsevier. 1866–2097.

Heckman, J. J. and R. Pinto. 2014. "Causal inference after Haavelmo". *Econometric Theory.* 31(1): 115–151.

Heckman, J. J. and J. Smith. 1998. "Evaluating the welfare state". In: *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium.* Ed. by S. Strom. Cambridge University Press.

Herndon, T., M. Ash, and R. Pollin. 2014. "Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff". *Cambridge Journal of Economics*. 38(2): 257–279.

Hoxby, C. and S. Turner. 2013. "Expanding college opportunities for high-achieving, low income students". *SIEPR Discussion Paper*. 12(014).

Kapoor, S. and A. Magesan. 2014. "Paging inspector Sands: The costs of public information". *American Economic Journal: Economic Policy*. 6(1): 92–113.

Kristensen, D. and B. Schjerning. 2015. "Implementation and estimation of discrete Markov decision models by Sieve approximations". *manuscript, University of Copenhagen*.

Labor, U. D. of. 2014. *What Works in Job Training: A Synthesis of the Evidence*. U.S. Department of Labor.

LaLonde, R. J. 1986. "Evaluating the econometric evaluations of training programs with experimental data". *American Economic Review*. 76(4): 604–620.

Leamer, E. E. 1982. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: John Wiley & Sons.

Levitt, S. D. 1997. "Using electoral cycles in police hiring to estimate the effect of police on crime". *American Economic Review*. 87(3): 270–290.

Ljungqvist, L. and T. J. Sargent. 2014. "An open letter to professors Heckman and Prescott". *manuscript, New York University*.

MaCurdy, T. 1985. "Interpreting empirical models of labor supply in an intertemporal framework with uncertainty". In: *Longitudinal Analysis of Labor Market Data*. Ed. by J. Heckman and B. Singer. Cambridge University Press. 148–170.

Manski, C. F. 2003. *Partial Identification of Probability Distributions*. Springer Verlag.

Manski, C. F. 2013. *Public Policy in an Uncertain World: Analysis and Decisions*. Harvard University Press.

Manski, C. F. 2014. "Identification of income-leisure preferences and evaluation of income tax policy". *Quantitative Economics*. 5: 145–174.

McCloskey, D. 2005. "The trouble with mathematics and statistics in economics". *History of Economic Ideas.* XIII(3): 85–102.

McCrary, J. 2002. "Do electoral cycles in police hiring really help us estimate the effect of police on crime? Comment". *American Economic Review.* 92(4): 1236–1243.

Misra, S. and H. Nair. 2011. "A structural model of sales-force compensation dynamics: Estimation and field implementation". *Quantitative Marketing and Economics.* 9: 211–257.

Neyman, J. 1923. "Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes." *Masters Thesis.*

Pakes, A. and P. McGuire. 2001. "Stochastic algorithms, symmetric Markov perfect equilbrium, and the 'Curse' of dimensionality". *Econometrica.* 69(5): 1261–1281.

Phillips, P. C. B., S.-P. Shi, and J. Yu. 2015. "Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P 500". *International Economic Review*: forthcoming.

Popper, K. R. 2002. *The Logic of Scientific Discovery.* Psychology Press.

Reinhart, C. M. and K. S. Rogoff. 2009. *This Time is Different Eight Centuries of Financial Folly.* Princton University Press.

Reinhart, C. M. and K. S. Rogoff. 2010. "Growth in a time of debt". *American Economic Review.* 100(2): 573–578.

Robert E. Lucas, J. 2003. "Macroeconomic priorities". *American Economic Review.* 93(1): 1–14.

Romer, P. M. 2015. "Mathiness in the theory of economic growth". *American Economic Review: Papers and Proceedings.* 105(5): 89–93.

Roth, A. E. 1991. "A natural experiment in the organization of entry level labor markets: regional markets for new physicians and surgeons in the UK". *American Economic Review.* 81: 415–440.

Roth, A. E. 1991a. "Game theory as part of empirical economics". *Economic Journal.* 101: 107–114.

Roth, A. E. 2008. "What have we learned from market design?" *Economic Journal.* 118: 285–310.

Rubin, D. 1980. "Discussion of randomization analysis of experimental data in the Fisher randomization test'". *Journal of the American Statistical Association.* 75(591).

Rubin, D. 2005. "Causal inference using potential outcomes". *Journal of the American Statistical Association.* 100(469): 322–331.

Rust, J. 1987. "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher". *Econometrica.* 55(3): 999–1033.

Rust, J. 1988. "Maximum likelihood estimation of discrete control processes". *SIAM Journal on Control and Optimization.* 26(5): 1006–1024.

Rust, J. 2014. "The limits to inference *with* theory: A review of Wolpin (2013)". *Journal of Economic Literature.* 52(3): 820–850.

Shiller, R. J. 2015. *Irrational Exuberance.* Princeton University Press.

Stahl, A. E. and L. Feigenson. 2015. "Observing the unexpected enhances infants' learning and exploration". *Science.* 348(6230): 91–94.

Su, C.-L. and K. L. Judd. 2012. "Constrained optimization approaches to estimation of structural models". *Econometrica.* 80(5): 2213–2230.

Theil, H. 1971. *Principles of Econometrics.* New York: John Wiley & Sons.

Todd, P. E. and K. I. Wolpin. 2006. "Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility". *American Economic Review.* 96(5): 1384–1417.

Vickrey, W. 1961. "Counterspeculation, auctions, and competitive sealed tenders". *Journal of Finance.* 166(1): 8–37.

von Neumann, J. 1947. "The mathematician". In: *Works of the Mind, Volume 1.* Ed. by R. Heywood. Chicago: University of Chicago Press. 180–196.

Widerquist, K. 2005. "A retrospective on the negative income tax experiments: Looking back at the most innovative field studies in social policy". In: *The Ethics and Economics of the Basic Income Guarantee.* Ed. by W. Lewis and Pressman. Aldershot: Ashgate. 95–106.

Wollan, M. 2015. "Pay it forward: Sophisticated software has enabled a 'market' for organ donation, without any money changing hands". *New York Times Magazine.* May(3).

Wolpin, K. 2013. *The Limits of Interference without Theory.* Cambridge: MIT Press.