Journal of Economic Literature
Vol. XLV (December 2007), pp. 973–1000

Interesting Questions in *Freakonomics*

John DiNardo[*]

Freakonomics *is more about "entertainment" than it is a serious attempt at popularization. Consequently, rather than conduct a comprehensive fact check, I use the book as a springboard for a broader inquiry into social science research and take issue with the book's surprising premise that "Economics is a science with excellent tools for gaining answers but a serious shortage of interesting questions." Using examples from* Freakonomics, *I argue that some of the questions the book addresses are "uninteresting" because it is impossible to even imagine what a good answer would look like. I conclude with some thoughts about the role of economic theory in generating interesting questions and/or answers.*

1. *Introduction*

Freakonomics: A Rogue Economist Explores the Hidden Side of Everything (Harper Collins 2005) is certainly popular. Written jointly by the University of Chicago economist Steven Levitt and *New York Times* journalist and author Stephen Dubner (*Confessions of a Hero-Worshiper* and *Turbulent Souls: A Catholic Son's Return to his Jewish Family*), the book has appeared on best seller lists internationally and has occupied the *New York Times Best Sellers* list for more than a year.[1] Moreover, with the release of an Instructor's Manual, as well as a Student Guide written by S. Clayton Palmer and J. Lon Carlson,[2] *Freakonomics* may become part of the learning experience for many economics students.

However, *Freakonomics* is more about "entertainment" than it is a serious attempt at popularization. Consistent with its hagiographication of Levitt the book lapses into "truthiness"—telling versions of the research that comport better with what (presumably) the audience wishes were true; the book's nearly "photo negative" misdescription of the effect of the Romanian dictator's abortion

[*] DiNardo: University of Michigan.

[1] See the book's website, http://www.freakonomics.com, for information about the book. Since I began this review, Levitt and Dubner have produced a new "revised expanded" edition of *Freakonomics* that, among other things, corrects some minor mistakes and reorganizes the book in light of the fact some readers found the vignettes that preceded each chapter in the original edition "intrusive (and/or egomaniacal, and/or sycophantic)." (Levitt and Dubner 2006a). The bulk of the changes regard their current view that they were "hoodwinked" regarding their description of Stetson Kennedy (Levitt and Dubner 2006b). The book also contains a new section of material freely available from their blog that I will largely avoid discussing, except where it seems especially pertinent to the "main" text. My page references will be taken from the original edition.

[2] The manual is available by request from Harper Collins Academic www.harperacademic.com. The student guide is free and available at http://www.freakonomics.com/pdf/StudentFREAKONOMICS.pdf.

ban is a case in point. More generally, although some of the research discussed has been challenged by others, little of the substance of these debates is treated as central to the discussion; the controversies, when they appear, are often treated as a sideshow in the "blog" material. I provide the briefest sketch of some of the issues this raises in John DiNardo (2006a).

While a comprehensive fact check of the claims of the book might be of some value, in light of the book's apparent aims, it would seem beside the point.[3] Rather, my hope is that *Freakonomics* might provide a springboard for a discussion of issues that I think apply more broadly to social science research.

One of the more surprising claims in *Freakonomics* is that "Economics is a science with excellent tools for gaining answers but a serious shortage of interesting questions" (p. xi). I do not wish to dispute that there is a wealth of uninteresting research and, when I look for entertaining or interesting insights into "human behavior," I am more likely to turn to a good novel than the latest working paper in economics. However, this claim runs so contrary to my experience (and I suspect, to the experience of many economists and social scientists) that it seems worthwhile to explore.

There are many criteria for interesting questions that will be given short shrift, despite being among the most important: *who* is included in the discussion, for example, is often more important than the intellectual capacities of the debaters. The quest by Emperor Charles V of Spain who "set out to discover the truth by experiment" (Lewis Hanke 1935) whether American Indians had the "capacity" for liberty called forth a flurry of research and debate among the most serious Spanish intellectuals of the day. It would not have been made more "interesting" by a more thorough attention to matters of methodology.[4] One suspects that few "American Indians" doubted their capacity for liberty despite the absence of social science research demonstrating otherwise.

Instead I would like to focus on criteria that might be used to distinguish good social science from good literature. Even if one stipulates that a good story need only sound "believable" or "entertaining"—in social science I believe we should aim for a different standard.

One sensible criterion is that claims in the social sciences should distinguish themselves by the "severity" of the "tests" to which they are put.[5] Mayo (1996) cites the American philosopher Charles Sanders Peirce to provide a nice short account of what a "scientific" approach is and what is meant by a "severe test":[6]

> [After posing a question or theory], the next business in order is to commence deducing from it whatever experimental predictions are extremest and most unlikely . . . in order to subject them to the *test of experiment.*

[3] From Noam Scheiber (2007), "'There's no question I have written some ridiculous papers,' [Levitt] says. By way of explanation, [Levitt] draws an analogy to the fashion industry: haute couture versus prêt-à-porter. 'Sometimes you write papers and they're less about the actual result, more about your vision of how you think the profession should be. And so I think some of my most ridiculous papers actually fall in the high-fashion category.'"

[4] Hanke's useful book describes the "first social experiments in America" and makes for an informative yet harrowing read in part because it was intended as *defense* of the Emperor and because the aim of his book was to demonstrate that "the Emperor . . . was imbued with a spirit not unlike that of a modern sociologist" (Preface) The description of the wide-ranging experiments undertook by the Spanish ends with the observation that "probably the mountain of evidence piled up during almost thirty years of social experimentation was high enough to convince the [Spanish] government that nothing could be gained by further attempts to make the Indians live like Christian laborers in Castile" (p. 71).

[5] For an insightful and much more careful exposition of the notion of "severe" testing that motivates this discussion, see Deborah G. Mayo (1996).

[6] Perhaps an even pithier summary was provided by Lucien Lecam (1977) in his critique of Bayesian solutions to the problems of inference: "the only precept or theory which really seems relevant is the following: 'Do the best you can.' This may be taxing for the old noodle, but even the authority of Aristotle is not an acceptable substitute."

> The process of testing it will consist, not in examining the facts in order to see how well they accord with the hypothesis, but on the contrary in examining such of the probable consequences of the hypothesis as would be capable of direct verification, especially those consequences that would be very unlikely or surprising in case the hypothesis were not true.
>
> When the hypothesis has sustained a testing as severe as the present state of our knowledge . . . renders imperative, it will be admitted provisionally . . . subject of course to reconsideration.[7]

The context of Peirce's remarks is a discussion of the importance and usefulness of bringing statistical reasoning to bear on history, though clearly they apply more broadly. While accepting the notion that putting our questions to a severe test is a good idea, for most problems there is no simple formula for assessing severity. Nonetheless, it seems like such a sensible criterion that it might come as a surprise that much economics research is of the first sort mentioned by Peirce—evaluating how well the facts accord with a given economic hypothesis. Undergraduate economics textbooks are filled with stories, very few of which have been forced to bear mild, let alone severe scrutiny, but are "broadly consistent" with the data.

A convenient place to begin is the issue, raised several times in *Freakonomics* (and the student guide, which refers to it as a "basic economics concept"), of whether an alleged relationship is "cause" or "correlation." Indeed, *Freakonomics* invokes several different notions of causality and I begin by reviewing some of what it has to say on the subject. Stripped to its essence, my argument is that such a debate often seems beside the point: "cause" means many things. A more relevant question about a correlation is whether it provides a severe test of a hypothesis.

Next I turn to a description of the randomized controlled trial (RCT), *not* as an exemplar of what all, or even most, social science should be but rather as an exemplar of subjecting a hypothesis to a severe test.

A basic precondition to severe testing, of course, is to formulate questions that can be put to some kind of test. Unfortunately, many social science questions often fail to meet this precondition. I take a couple of examples from *Freakonomics* and argue that some of the questions it addresses are "uninteresting" because it is impossible to even imagine what a good answer would look like. Somewhat ironically, the issues in *Freakonomics* that have generated the most popular debate seem are the ones that seem to have no good answers.

I conclude with some thoughts about the role of economic theory in generating interesting questions and/or answers.

2. *Correlation is Causation?*

Causes make appearances in *Freakonomics* in many different and confusing ways.[8] In some places, *Freakonomics* seems to invoke causation as "explanation" or "motive":

> What might lead one person to cheat or steal while another didn't? How would one person's seemingly innocuous choice, good or bad, affect a great number of people down the line? In [Adam] Smith's era, cause and effect had begun to wildly accelerate; incentives were magnified tenfold. The gravity and shock of these changes were as overwhelming to the citizens of his time as the gravity and shock of modern life seem to us today (p. 15).

In another passage, the inability to reason about causation is described as an evolutionary by-product exploited by "experts":

[7] Peirce (1958) 7.182 (emphasis added) and 7.231 as cited in Mayo (1996).

[8] This is unfortunate as "causation" is a notoriously difficult topic even when treated by a serious philosopher and I can not do it minimal justice here. One place to start for a more ably argued introduction to these issues in economics is Julian Reiss and Nancy Cartwright (2004). See also Cartwright (2007).

> We have evolved with a tendency to link causality to things we can touch or feel, not to some distant or difficult phenomenon. We believe especially in near-term causes . . . a snake bites your friend, he screams with pain, and he dies. The snakebite, you conclude, must have killed him. Most of the time, such a reckoning is correct. But when it comes to cause and effect, there is often a trap in such open-and-shut thinking. We smirk now when we think of ancient cultures that embraced faulty causes: the warriors who believed, for instance, that it was their raping of a virgin that brought them victory on the battlefield. But we too embrace faulty causes, usually at the urging of an expert proclaiming a truth in which he has a vested interest (p. 140).

Confusion about correlation, when not being exploited by unsavory experts, is the product of soft-headed thinking:

> The evidence linking increased punishment with lower crime rates is very strong. Harsh prison terms have been shown to act as both deterrent (for the would-be criminal on the street) and prophylactic (for the would-be criminal who is already locked up). Logical as this may sound, some criminologists have fought the logic. A 1977 academic study called "On Behalf of a Moratorium on Prison Construction" noted that crime rates tend to be high when imprisonment rates are high, and concluded that crime would fall if imprisonment rates could only be lowered. (Fortunately, jailers did not suddenly turn loose their wards and sit back waiting for crime to fall.) . . . The "Moratorium" argument rests on a fundamental confusion of correlation and causality (p. 123).

While war, rape, and experts wielding dubious metaphysics may be as old as humankind, confusion about "correlation versus causation" is arguably quite recent. Even the idea of "probability" as we might understand it today emerged only in the seventeenth century (Ian Hacking 1975). At that time, there was a great deal of reluctance to introduce *any* notion of "chance" into laws of nature. Several years after Smith's *Wealth of Nations*, Laplace could still write "all events, even those which on account of their insignificance do not seem to follow the great laws of nature, are a result of it just as necessarily as the revolutions of the sun."

Karl Pearson (1930), proponent of eugenics and an important contributor to modern statistics and scientific philosophy (who did much to popularize the idea of correlation) argued that "correlation" superseded the notion of "causation":[9]

> Up to 1889 [when Galton published Natural Inheritance], men of science had thought only in terms of causation . . . . In [the] future, they were to admit another working category, correlation which was to replace not only in the minds of many of us the old category of causation, but deeply to influence our outlook on the universe. The conception of causation—unlimitedly profitably to the physicists—began to crumble to pieces. In no case was B simply and wholly caused by A, nor, indeed by C, D, E, and F as well! It was really possible to go on increasing the number of contributory causes until they might involve all the factors of the universe.

To put Pearson's views in context, he was reacting against a view held by many others that "stable" correlations—correlations that didn't change much over time, for example—were informative about causes or causal laws—an idea that is coterminous with the idea of correlation itself. One example, perhaps one of the earliest predecessors to *Freakonomics*, is André-Michel Guerry's (1883) *Essay on the Moral Statistics of France*.[10]

One of the most sensational of Guerry's findings was his refutation of the view that "ignorance is the principal cause of crime, and that to make men better and happier, it is sufficient to give them an education."

[9] Pearson was a complex figure who made contributions in many areas. His book, *The Grammar of Science* (1892), for example, was on a list of books read by the famous "Olympia Academy" reading group of Albert Einstein, Conrad Habicht, and Maurice Solovine in 1902.

[10] Guerry's work "appears to be the first to test 'armchair' assumptions about the relationship of certain variables to criminal behavior" (Sue Titus Reid 1985). Like *Freakonomics*, it was an international hit and was popular among "amateurs" (Hacking 1990).

According to Guerry, this view was based on the observation that *the departments where education is least widespread are those where the most crimes are committed* (Guerry 1883, page 87, emphasis in original). Guerry was able to refute conventional wisdom on the subject by merely demonstrating (with better data) that the correlation between education and crime at the department level was not negative, but positive. Moreover, Guerry apparently felt little need to consider the possibility of what we might call "confounders." Having established that the correlation was positive, he adduced further evidence that the conventional view was wrong by demonstrating the *stability* of the correlation over time: a law of sorts.

The impulse to embed statistical uncertainty in an otherwise "determinist" world view led to arguably one of the most bizarre intellectual strands in social science: the idea that statistical laws vitiated free will (Hacking 1990; Hacking 1983a). A wonderful illustration comes from Charles Dickens's *Hard Times* (1854). Mr. Gradgrind (who it may be recalled named two of his sons "Malthus" and "Adam Smith"!) was in part the satirical embodiment of statistical fatalism.[11] If the number and proportion of crimes displayed statistical regularities, could the criminals really have free will? When Mr. Gradgrind's son Tom is revealed to be a thief, Tom responds to his father's shock and dismay this way: "'I don't see why,' grumbled the son. 'So many people are employed in situations of trust; so many people, out of so many, will be dishonest.' I have heard you talk, a hundred times, of its being a law. How can *I* help laws? You have condemned others to such things, Father. Comfort yourself" (book three, chapter 7).

While we have long abandoned the view (I hope) that statistical laws have anything to say about free will, still with us is the idea that statistical distributions are "laws" that regulate human behavior on some macro scale.[12]

Notwithstanding the paucity of economic "laws," the idea that mere empirical regularities *might* embody "causes" can not always easily be dismissed. One might argue that Newton's law of gravitation was an example of an empirical regularity—correlation—that became a "cause." Surely we can talk about gravity causing my cup of coffee to fall off the table after I pushed it. However limited we might find such an account of gravity as a cause, the testable predictions from the law of gravitation can ultimately be put to rather severe tests in an awe-inspiring variety of contexts. Thus, one can sympathize with Leibniz's opposition to the concept of gravity—which he dismissed as an "occult" force—while maintaining it is a useful and powerful idea. Whether gravity is "real," a law of nature, or whether it is really a "cause" seems beside the point.

2.1 *Distant and Subtle Causes*

The word "cause," unfortunately, can mean many different things. Herbert Simon once observed that "in view of the generally unsavory epistemological status of the notion of causality, it is somewhat surprising to find the term in rather common use in scientific writing"(Simon 1953 as cited in Zellner 1984). Indeed one of the most confusing themes in *Freakonomics* is that "distant and subtle causes can have dramatic effects."

Their claim about "distant and subtle causes" is confusing in a couple of ways. First, it doesn't seem to speak to the type of "manipulationist" notions of causality that concern many in social science. Second, the claim evokes an echo of the Laplacean

[11] And economics as well. The colorful protagonist Sissy defines the basic principle of Political Economy as "To do unto others, as they would do unto me"!

[12] See, for example, Arthur DeVany (forthcoming) and the references cited within.

determinism I discussed above. While it is not precisely clear what notion of "cause" is being invoked, it seems to speak to some "causal" antecedent which sets off a long chain of events ultimately resulting in a specific event. It is a common narrative device in fiction—many a character's fate can be "traced back" to a single fateful act.

The search for the single (or small number of) causal antecedent(s) of an event is surprisingly common among economists: "what caused wage inequality to increase in the 1980s" or "what caused the Great Depression" or "what caused crime to fall in the 1990s" (a question taken up in *Freakonomics*) are three examples that come to mind. I won't deny that the search for answers to such questions can sometimes be informative. Nonetheless, except for the very simplest phenomenon, it is rarely clear what constitutes a good answer to such a question.[13]

Consider something as simple as "*the cause of death*." Enumeration of such causes dates back at least to 1592: "the occasion of keeping an accompt of burials arose first from the plague" (John Graunt 1676). Not surprisingly, the victims of the plague were *not* drawn randomly from rich and poor; neither was the focus on the cause of death politically inert. Anne M. Fagot (1980) reports on one Doctor Vacher who, seeking to understand the dramatic increase in deaths during the 1870 siege of Paris, went back to study an even earlier four-month siege of Paris in 1590. After studying the data, he was led to conclude that one of the "effects of insufficient food" was that the lethality of diseases such as typhoid was much greater. Nonetheless, "hunger" or "lack of food" was rarely cited as a "cause" of death, although he identified undernutrition as an "*underlying* potential cause."

This arbitrariness, of course, persists. In the United States, for example, most people die of more than a single cause of death; yet even on the death certificate, where up to twenty causes of death can be reported, the distinction between "underlying causes" and other types of cause remains! (Center For Health Statistics 1998).[14] Despite its arbitrariness, such information can be useful. Indeed, if there is any clear doctrine on how to attribute the cause of death, perhaps it is the requirement that the classification scheme is somehow minimally "useful" (Fagot 1980). No amount of diligent record keeping, however, will be able to create a "complete" description of "why" some people die—debate on "why" Jesus died continues! (W. D. Edwards, W. J. Gabel, and F. E. Hosmer 1986; Anonymous 1986; C. G. Gosling 1987; B. Brenner 2005; H. ur Rehman 2005; W. R. Saliba 2006).

2.2 *Cause as Explanations*

Surely "crime" or other social science issues are at least as complicated as "death." Yet it is surprising how much social science research seems dedicated to telling *simple* stories. This suggests another related notion that might be called "cause as explanation." While such stories appear to have great appeal, I must confess I don't understand why.

A well known reductio ad absurdum of this type of reasoning concerns the famous Dr.

[13] For a more optimistic view about the types of questions that are "susceptible to empirical investigation," see Zellner (1984).

[14] The U.S. Implementation of the International Classification of Diseases includes this instruction: "A cause of death is the morbid condition or disease process, abnormality, injury, or poisoning leading directly or indirectly to death. The underlying cause of death is the disease or injury which initiated the train of morbid events leading directly or indirectly to death or the circumstances of the accident or violence which produced the fatal injury. A death often results from the combined effect of two or more conditions. These conditions may be completely unrelated, arising independently of each other or they may be causally related to each other, that is, one cause may lead to another which in turn leads to a third cause, etc." (National Center for Health Statistics 2006, p. 6).

Pangloss in *Candide*.[15] At one point Candide is reunited with his former teacher Dr. Pangloss, who has been reduced to a beggar with his nose half-eaten off, covered in scabs. Surprised by this (and a lot of other) misfortune, Candide "inquired into the cause and effect, as well as into the sufficing reason that had reduced Pangloss to so miserable a condition." We learn that Dr. Pangloss had "tasted the pleasures of Paradise" with Pacquette, a pretty servant girl who had, as it turns out, been infected with a disease, the impressive genealogy of which Dr. Pangloss is able to trace back to a Countess, a Jesuit, a novitiate (among others), and ultimately Christopher Columbus. Candide asks *why* did Dr. Pangloss suffer such a horrific fate? What *caused* his degradation? For Dr. Pangloss, causal questions were straightforward: things could not be otherwise than they are, all things are created for *some* end, and thus all things are created for the best. In this case, Dr. Pangloss concludes his suffering was "a thing unavoidable, a necessary ingredient in the best of worlds" for had this disease not come to pass "we should have had neither chocolate nor cochineal."[16]

The humor in *Candide* comes from the creativity with which one can generate a "theoretically justified" explanation of "why" for *any* set of facts. One obvious problem with Dr. Pangloss' explanations is the impossibility of putting such views to a severe test of any sort.[17]

Much economics as "explanation" it seems to me, resembles Dr. Pangloss's explanations. With enough cleverness one can dream up a mathematical model of utility–maximizing individuals to explain *anything*. It is not always clear what purpose such explanations serve. In a throw-away line in *Freakonomics*, for example, the authors attribute the putative fact that "the typical prostitute earns more than the typical architect" (p. 106) to a "delicate balance" of "four meaningful factors."[18] I don't mean to deny that such factors often play some role—certainly, for example, an intervention to make a job more unpleasant may act to reduce the number of people willing and able to do that job. But even if we stipulate to these "four meaningful factors," that is only the *beginning* of an explanation at *best*.[19]

[15] Voltaire (1796) describes Pangloss this way: "[He] was a professor of metaphysico–theologo–comsolo–nigology. He could prove, to admiration, that there is no effect without a cause; and, that in this the best of all possible worlds, the baron's castle was the most magnificent of all castles, any lady the best of all possible baronesses. It is demonstrable, said he, that things cannot be otherwise than as they are: for all things having been created for some end, they must necessarily be created for the best end. Observe, that the nose is formed for spectacles, and therefore we wear spectacles. The legs are visibly designed for stockings, and therefore we come to wear stockings" (p. 4).

[16] See chapter 4, page 14, of Voltaire (1796). The translator of this version of Voltaire's story attributes this style of reasoning to the "maxims of Leibniz" and as put into the mouth of Dr. Pangloss is a "most Capital and pointed stroke of Satire." Cochineal is apparently a red dye made from ground up insects.

[17] John Maynard Keynes (1921, p. 297) argues that "the discussion of [Aristotelian] final causes and of the argument from design has suffered from its supposed connection with theology. But the logical problem is plain and can be determined upon formal and abstract considerations." He illustrates the case with the evidentiary value of observing of some unusual event and ascribing its cause to some "supposed conscious agent." With a simple application of Bayes' rule he does conclude that "no conclusion, therefore, which is worth having, can be based on the argument from design alone; like induction, this type of argument can only strengthen the probability of conclusions, for which there is something to be said on other grounds."

[18] "When there are a lot of people willing and able to do a job, that job doesn't generally pay well . . . the others are the specialized skills a job requires, the unpleasantness of a job, and the demand for services that the job fulfills" (p. 105).

[19] As to the truth of the claim about prostitute wages, it is too imprecise to verify or deny; moreover Dubner and Levitt provide no reference. In a previous version of this essay, I concluded that it would be a major project to verify such a claim. Putting aside the almost insuperable problems of defining prostitution and measurement of hours worked, a comparison of data from a probability sample of Los Angeles prostitutes, Lee A. Lillard (1998) revealed that measured in 2004 dollars, the mean income for "Street Prostitutes" in Los Angeles was $36,325 in 1989. In May 2004, data from Occupational Employment Statistics for "Architects, Except Landscape and Naval" suggested an annual income from work of $66,230 (assuming 2,080 hours of work per year).

3. *The Randomized Controlled Trial as* One *Type of Severe Test*

Some writers have sought to *define* a cause as something that arises from the predictable consequence of an intervention that can be evaluated by something approximating randomized design. As the foregoing has made clear, this definition is too limited given the many different notions of the word "cause." Rather than "no causation without manipulation" (Paul W. Holland 1986) it might be more truthful to say that discussions in social science about causes are more *intelligible* when they involve an intervention of some sort; moreover, a focus on such "policy evaluation" questions often leads to more interesting questions, and importantly often leads to situations when we may able to subject our views to some kind of test. In a helpful discussion, Reiss and Cartwright (2004) suggest the slogan "disambiguate before you evaluate."

My purpose in discussing a RCT is that it is useful to review a framework where *what* question is being asked, and the ground rules under which we might find an answer credible is arguably more transparent than is usual. As is common practice, I will describe questions answered in such a framework as "causal," although they are often causal in a very limited sense.[20] Indeed, the origins of the RCT lie in the attempt to put some of the "squishiest" beliefs to a severe test—some of the earliest examples arose in the study of telepathy (Hacking 1988).[21]

In *Freakonomics*, regression analysis is described as the tool of someone who can't conduct a RCT:

> In a perfect world, an economist could run a controlled experiment just like a physicist or a biologist does: setting up two samples, randomly manipulating one of them, and measuring the effect. But an economist rarely has the luxury of such pure experimentation. (That's why the school-choice lottery in Chicago was such a happy accident.) What an economist typically has is a data set with a great many variables, none of them randomly generated, some related and others not. From this jumble, he must determine which factors are correlated [sic] and which are not (p. 162).[22]

Putting aside whether this is a description of good practice, the view that regression is a (sometimes inadequate) substitute for a randomized controlled trial is not universally held by economists. More surprisingly, perhaps, is that as a philosophical matter "it is hard to think of a more controversial subject than that of randomization (Patrick Suppes 1982, p. 464)[23] Convincing Bayesian rationales for randomization, for example, are evidently difficult to

[20] For those who prefer a definition of "cause," one that seems to capture some of the ideas I have in mind is due to James J. Heckman (2005): "Two ingredients are central to any definition [of causality]: (a) a set of possible outcomes (counterfactuals) generated by a function of a set of 'factors' or 'determinants' and (b) a manipulation where one (or more) of the 'factors' or 'determinants' is changed. An effect is realized as a change in the argument of a stable function that produces the same change in the outcome for a class of interventions that change the "factors" by the same amount. The outcomes are compared at different levels of the factors or generating variables. Holding all factors save one at a constant level, the change in the outcome associated with manipulation of the varied factor is called a causal effect of the manipulated factor" (p. 1). For a discussion of the limitations of any single definition of causality relevant for economists, see Cartwright (2007) and Reiss and Cartwright (2004). For a thoughtful and well-reasoned discussion of views about causation that do not have a central role for "manipulation," see Zellner (1984). Zellner prefers to work with a definition proposed by Feigl "the clarified (purified) concept of causation is defined in terms of predictability according to a set of laws." By doing so, he appears to be able to consider many sorts of questions—albeit subject to a logical (Bayesian) calculus—which could not be put to a severe test in my way of viewing of the issue.

[21] It is also unsurprising that Peirce was one of the earliest to conduct a high quality RCT (Stephen M. Stigler 1978). Even economists played a role: Francis Ysidro Edgeworth (1885, 1887) wrote up two excellent analyses of the results of a trial involving randomization in the *Journal of Psychical Research*.

[22] It is not clear what is intended. Even in a "jumble" of data, determing what variables are *correlated* is straightforward.

[23] For a sample of this debate, see Suppes (1982), David A. Harville (1975), Zeno G. Swijtink (1982), and the illuminating debate in Leonard J. Savage (1962) especially pages 62–103.

generate[24] and this difficulty has been the source of criticism of Bayesian methods for their failure to recognize a "distinction between 'experiences' and 'experiments'" (Lecam 1977, p. 137).[25]

Rather than hold up the RCT as a paradigm for all research, I review it here because it represents *a single case* in which we *sometimes* have *some* hope of evaluating (limited, context dependent) causal claims, and because what constitutes a severe test is somewhat clearer.

Second, the RCT is a useful framework to discuss the "intelligibility" of putatively causal questions. That is, if one is discussing a "causal" question, whether or not one is discussing an RCT, the RCT often provides a useful template to evaluate whether the causal question is answerable. It allows us to try to answer the question "what do you *mean* by a causal effect?" as well as the related question "how *credible* is your inference about the 'cause'?"

A natural by-product of considering the RCT is that the limitations of a research design to answer "interesting" questions (and what might provide evidence for and against the validity of the design) is easier to understand. Ironically, I suspect that some of the disenchantment with RCTs relates to the relatively transparent notion of "cause"—in particular the possibility that the putative cause under examination is "implementation-specific," which I discuss below.

3.1 *Randomized Controlled Trials*

In an RCT, a single potential cause is randomly "assigned" to a treatment group and an (inert) placebo is assigned to the control group.

Let y_i be an outcome which can be measured for all individuals, and let $T_i = 1$ signify that person *i* has been assigned to treatment and $T_i = 0$ otherwise. Suppose the following characterizes the true state of the world[26]:

$$y_i = \alpha + \beta T_i + f(X_i) + \varepsilon_i, \tag{1}$$

where α and β are constants, $f(.)$ is some unknown function of all the observable characteristics that affect y_i before being assigned to the treatment or control, and ε_i is all the other unmeasurable influences. Even at this level of generality, it takes a considerable leap of faith to think that this simple (partially) linear representation can yield anything but the most limited understanding of the effect of T even when some understanding is possible.

A fundamental problem we face is that, for an individual *i*, we can only observe the person in one of the two states—treatment or control. Another related problem is that we don't observe everything that affects the outcome y. For any individual then, we can never be certain that some unobserved determinant of the outcome y is changing at the same time we are assigning the person to treatment or control.

The key to this design is that by coin toss, nature, or some other contrivance that generates "random numbers," persons are

[24] There are many different flavors of Bayesian arguments against randomization. One argument, not necessarily the best, will be familiar to economists. From Scott M. Berry and Joseph B. Kadane (1997): "Suppose a decision maker has two decisions available d_1 and d_2. These two decisions have current (perhaps posterior to certain data collection) expected utilities $U(d_1)$ and $U(d_2)$ respectively. Then a randomized decision, taking d_1 with probability λ and d_2 with probability $1 - \lambda$, would have expected utility $\lambda U(d_1) + (1 - \lambda)U(d_2)$. If the randomization is nontrivial, i.e., if $0 < \lambda < 1$, then randomization could be optimal only when $U(d_1) = U(d_2)$, and even then a nonrandomized decision would be as good." Another rationale is that it helps "simplify" the appropriate likelihood (Donald B. Rubin 1978).

[25] Although salient to this discussion, limitations of scope do not permit an extensive discussion of these issues. For a useful discussion and a defense/reformulation of classical statistical inference, see Mayo (1996).

[26] Another way to proceed, which is often helpful, is to establish a notation for counterfactuals. Let $Y_i(1)$ be the outcome when the person is assigned to the treatment and let $Y_i(0)$ be that same person's outcome when they are assigned to the control. The treatment effect for person *i* is then $\tau_i \equiv Y_i(1) - Y_i(0)$. It is generally impossible to observe t_i since the individual is in one state or the other. We could then talk about trying to define $E[\tau_i]$ (for some population) as the object of interest. See Holland (1986) for an exposition along these lines. See Heckman (2005) for a critique of that approach and related points.

next assigned to either treatment or control in a way that is independent of their characteristics. If this assignment is conducted on a random sample of individuals from a particular population, then the mean outcome for individuals in the treatment group—$\bar{y}_{T=1}$—is a good estimate of the average outcome of individuals from this population under the treatment—$\alpha + \beta + E[f(X_i)]$. By similar logic, $\bar{y}_{T=0}$ is a good estimate of the average outcome for the control group—$\alpha + E[f(X_i)]$(provided, of course, that there is in fact some stable relationship between the cause and the outcome.)[27] The difference between these two means is likewise a "good" estimate of the average treatment effect for this group.[28]

The assertion that the estimate so formed is a "good" one is fortunately not one that has to be taken solely on faith: it can be tested. While not "assumption free," our confidence in estimates generated this way *does not* rely on us having complete knowledge of the data generation process given by equation (1). Specifically, it is reasonable to hope that we can get a good answer without having to hope that somehow we can "control" for all possible confounds.

In a typical RCT, in fact, any of the variables in X are generally not used for any purpose but to test the design. Under random assignment, *any* X should be the same on average for the two groups. This is, of course, a consequence of random assignment that is routinely tested in every RCT. If the groups look very different on average, this is generally considered evidence against the design, and one reason to have less confidence in the results. A related implication is that in an RCT, the answer should be insensitive to the addition of additional controls.[29]

It is the fact that the important *X's* are the same on average that gives us some reason to believe that the same is true for the ε. Even in this simple case, we can never be sure that this is true. At best, the answers from identical experiments have the "tendency" to be correct.

Several attractive features of a well designed RCT that are usually too obvious to deserve mention become more important when one turns to the sorts of "approximations" we are often faced with in social science:

(1) *Prespecified research design.* In an RCT, the researcher specifies *in advance* to the extent possible the conditions that have to be satisfied, and what will be concluded under every possible result of the experiment. (This is articulated with the usual degree of tentativeness associated with any technique involving sampling.)[30]

[27] Already this aspect of the RCT highlights its weakness for a lot of social science questions. Many social scientists are interested "why" someone does what they do or why things turned out as they did. In the RCT, however, the credibility of the answer hinges on the fact that part of human *choice* has been handed over (implicitly or explicitly) to a (hypothetical) chance set up. This is also a source of the considerable ethical problems that are frequently involved in RCTs.

[28] Even in this short description we have swept several very important issues under the rug that can arise even in a simple medical example. For instance, we are assuming that "general equilibrium" effects are unimportant so that one isn't concerned that the controls are affected by the treatment also. These and related concerns become even more important when we raise our ambitions to seek to extrapolate the results of the experiment to other possibly different contexts. There is a long tradition in economics of seeking answers to these more difficult questions that dates back at least to the Cowles Commission (see Heckman (2000) and Heckman (2005) for useful discussions). I focus on "simpler" less ambitious questions. (Heckman and Edward Vytlacil 2005).

[29] This is, one is tempted to speculate, the source of the intuition, that many appear to have, that somehow if a result "survives" the inclusion of a long list of covariates, it is a more trustworthy estimate.[30]

[30] I don't mean to advocate a simple-minded caricature of the Fisher or Neyman–Pearson significance testing approach. Long-standing criticisms of insisting on prespecification is that they are rarely strictly applied (with good reason). See Mayo (1996) for a discussion of the debate about "predesignation" and a helpful reformulation of Neyman–Pearson "error statistics." In her framework, violations of predesignation are licensed when they don't make the test of the hypothesis less "severe." Surprisingly, some Bayesians argue for the irrelevance of predesignation on the grounds that the "mental state" of the person collecting the data should have no relevance for the evidential import of the data. See Mayo and Michael Kruse (2002) and the references therein for a useful discussion.

If we are assessing the efficacy of a drug, for instance, it is pointless to decide in advance that the drug "works" and then massage the data, sample, specification, etc. until we "reach" that conclusion. Doing so would seem to vitiate using the RCT (or regression more generally) as a method for anything but confirming our previously held beliefs.[31] Indeed, historically and etymologically the notion of an "experiment" is intimately related to the effort to put one's views to the test (DiNardo 2006b). Clearly, *after the fact* research design is less "severe."

(2) *"Transparent" research design.* In the classical RCT, as one example, it is transparent what constitutes evidence against the design (for example, if the predetermined characteristics of the treatment and control are very different) and what comparison or regression coefficient constitutes evidence in favor of, or against, the claim.

Another set of assumptions—again usually too obvious to be discussed in the case of the RCT—deal with whether a question or set of questions are "well posed" or whether the answer suggested by RCT addresses the "intended" question.

(1) We can identify a "treatment" or "policy." At one level, since we are dealing with human beings, one often has to carefully distinguish between "assignment to treatment" and the "treatment." You can assign someone to take a specific medicine but it isn't always reasonable to assume that the person has taken the medicine. Even if we can ignore such distinctions it may be difficult to identify what our treatment *is*. Even the most routine, minor medical manipulation often comes bundled with other things. Many years ago it would have been a sound inference based on much unfortunate experience that the causal effect of a spinal tap (lumbar puncture) would be a serious headache afterward. Is this effect caused by the substance used to sterilize the needle? The type of needle? The size of the needle? Despite the fact that lumbar punctures have been performed for more than one hundred years (A. Sakula 1991), these questions continue to be subject of debate despite *many* randomized controlled trials (Carmel Armon and Randolph W. Evans 2005).

(2) The effect of a treatment is always *relative* to the control. The state of being assigned to the control is the "counterfactual" against which the treatment is evaluated. An effect is a comparison of outcomes in different possible states.

(3) The treatment involves an "intervention" and/or is "manipulable." In the RCT, this is so basic it hardly deserves mention; it is, however, a subject of some debate among economists.[32] In the limited way I wish to use the word "cause," it is not meaningful to question the effect of "being black" on one's propensity for crime. Only in a fantasy world does it make sense to consider the fate of John DiNardo as a "black man." If a misguided social scientist had been able to secretly reach back into the womb to manipulate John DiNardo's DNA to make him "black" (something that would have no doubt come as a surprise to his Italian parents) would it even be meaningful to describe the person generated from that process as the "black John DiNardo" to which the "white John DiNardo" could be compared? The issue is not "Is such a manipulation possible?" but "Were such a manipulation

[31] For an illustration of evolving definitions of the "appropriate" specification *after* having seen the results, and the consequences of failing to adopt a prespecified research design, see the discussion of Finis R. Welch (1974), Frederic B. Siskind (1977), Welch (1976), and Welch (1977) in chapter 6 of David Edward Card and Alan B. Krueger (1995). Although the extent of this research style is unknown, I suspect that the example is unusual only because it is documented.

[32] See Clive Granger (1986) for example.

conceivable, would it answer the question we are asking?" If the answer to that question is "no," I would describe the question as ill-posed or unintelligible even if it is the answer to a different well-posed question. I have no wish to overstate this issue: some of debate may be of no greater moment than questions of terminology. For example, I think it is possible to talk in a limited way about the effect of changing a person's *perception* of the race of, say, a job applicant because it is perhaps meaningful to think about manipulating a person's perception of race.[33]

(4) Related to this last issue is the hope that "how" the treatment is assigned is irrelevant to the effect (β) on the outcome. If the effect of the putative cause is implementation specific, it is often more helpful to abandon the effort to find the effect of the putative cause and "settle" for the effect of the "*implemented* cause." For example, if the effect of aspirin on headache differs when it is given to a patient by a nurse than when it is given to a patient by a doctor, the most we may be able to do is describe the causal effect of "nurse administered aspirin" or "doctor administered aspirin." In the limit, of course, if only the method of administration matters we might even wish to conclude that aspirin *qua* aspirin doesn't cause anything to do with headache. At a very minimum, if such were the case, a debate about the causal effect of aspirin would be, at a minimum, unintelligible.

(5) I would add, although this is not properly thought of as a "requirement," the most interesting studies involve manipulations that correspond to real policies. In these cases, even if we learn little about the "structure" of a true model, we have perhaps learned something about the consequences of one possible action we have taken.

I do not mean to suggest by the foregoing that a RCT is always or usually the "best" evidence. Quite to the contrary, I don't even think that a singular focus on "well-posed" questions would be a good idea.[34]

I would go even further and suggest that in many areas under study by economists, the focus on "treatments" can be, perhaps unintentionally, narrow. As David Thacher (2001) observes, "Reducing crime is clearly one important goal for the police. But it must compete with other goals like equity, due process, just desserts, and parsimony." Rather I argue that if a putatively causal question can not be posed as some sort of

[33] Robert Moffitt (2005), for example, explains that "[The argument in Holland (1986) that race can not be a cause because it can not be manipulated results from] . . . a mistaken application of the experimental analogy, and the more basic counterfactual analogy is the superior and more general one. It does make conceptual sense to imagine that, at any point in the lifetime of (say) an African-American, having experienced everything she has experienced up to that time, her skin color were changed to white (this is sometimes called a gedanken, or thought, experiment). Although it is a well-defined question, it may nevertheless be unanswerable, and it may not even be the main question of interest. For example, would the individual in question move to a different neighborhood, live in a different family, and go to a different school? If not, the question is not very interesting" (p. 105). While a distinction between comparisons one could make and those that are possible is important (I wish to think of manipulable quite broadly), I find such discussion confusing. If I were to wake up tomorrow and discover that my skin color had changed dramatically, one possible reaction might be a visit to the Centers for Disease Control to learn if I had acquired an obscure disease! *Whether or not* I moved to a different neighborhood, or divorced my wife, if that response were typical of other white folks who woke up one day to find themselves "black," I would nonetheless hesitate to say that the "causal effect of being black" (or white) is an increase in the probability that one makes a visit to the CDC [as above], though it could be so described. Again, absent some discussion of a class of counterfactual states *and* hypothetical manipulations, for me it is hard to know what to make of such causes, even when they can be defined.

[34] In this regard, the philosopher Hacking has done a great deal to show that useful work can be done in areas that vary quite widely in how well posed the questions are. For a study of statistical questions, see Hacking (1965), the role of experimentation in natural science (Hacking 1983b), multiple personality disorder (Hacking 1995) and the "social construction of reality" (Hacking 2000), for example.

"approximation" to a question satisfying the above desiderata, the burden of explaining what is meant in plain language should be borne by the author. Too frequently, however, it is not.[35]

4. *Just Because We Can Manipulate It Doesn't Mean We Can Learn About It*

One of the serious problems with a focus on the RCT is the misleading view that we can always learn about causes from manipulations. Cartwright (2007a, 2007b) makes the point with greater generality than I can here. Rather I would like to focus on one class of problems with direct relevance for much of the research described in Freakonomics. My argument is simply that although we can learn about the effect of an intervention in a well-designed study, we aren't guaranteed to learn about *the* putative cause in question, because the cause under consideration may be inextricably implementation specific. Consider the "causal effect" of obesity on all-cause mortality. The literature hardly seems to doubt that it is possible to measure such an effect, though there may be problems—perhaps body mass index (BMI) is an inappropriate measure, for example.

Nonetheless, I would argue that it is unlikely that anyone will devise a severe test of the proposition that obesity causes an increase in all-cause mortality. Simply put, the effect of obesity (or of ideal weight) is inextricably implementation specific. That is, it is not helpful to think about the "effect" of obesity for the same reason it is not helpful to debate the "causal effect" of race on income (Granger 1986, p. 967).

Many of us suspect, for example, that encouraging obese individuals to "starve themselves" for short periods of time might help one lose weight but wouldn't necessarily promote longevity (although it might, who knows?). Similarly, we might expect weight loss that results from increased physical activity to be more protective than weight loss that results from increased life stress.

The experience in the United States with the drugs fenfluramine and dexfenfluramine (Redux) is a case in point. Despite good evidence that the causal effect of taking Redux was weight loss, the drugs were pulled from the market because a "side effect" of the medication was an increase in potentially serious heart problems (Food and Administration 1997) .

Indeed, it would appear that the presumption that obesity is a cause of ill health made it virtually impossible to debate whether *nonobesity* was the *cause* of the increased heart problems. Rather, the consensus seems to be that the heart problems were not caused by *nonobesity*, but rather by Redux's "side effects."[36] I don't want to argue that "ideal weight" is bad for one's health, only that this example highlights the fact that the effect of weight loss or weight gain is inextricably implementation-specific. If one accepts, this logic, much of the research claiming that being nonobese (and nonunderweight) is causally related to better health demonstrates no such thing. Indeed, this literature is filled with "anomalous" results.[37] Moreover, "theory" seems to provide little help: even the weakest research

[35] This point is not in any way unique to me. For different, but not unrelated, views of these issues with relevance to social science, see Holland (1986), David A. Freedman (1999), Jude Pearl (1997), Heckman (2005) and William R. Shadish, Thomas D. Cook, and Donald T. Campbell (2002), to name just a few.

[36] I am merely stipulating to the existence of a distinction between "effects" and "side effects"; frequently the distinction seems to be based on marketing rather than scientific concerns.

[37] In an excellent but unfortunately unpublished study, Jerome Timothy Gronniger (2003) finds that if one includes more careful OLS type controls for income the putative effect of obesity is actually protective for many income groups. In arguing against viewing obesity as a "causal" factor in all-cause mortality, he also observes that the salient policy question "is not what obesity does to people, but what removing obesity would do to people." A heavily abridged version of the article appears as Gronniger (2006).

design often comes with an impressive theory. M. Cournot et al. (2007), for example, finds an "association" between obesity and lower "cognitive functioning" (verified by a simple cross-sectional design regressing measures of cognitive functioning on a small number of covariates and BMI) and posits one possible "theoretical" reason for why the link might be "causal": "direct action of adiposity on neuronal tissue through neurochemical mediators produced by the adipocyte" (fat cell).

My point is simple: when each way of "assigning" obesity that we can imagine would be expected to produce a different effect on all-cause mortality or other outcomes, it is not at all clear that it is helpful to debate the "effect of obesity." It seems more intelligible (and more policy relevant) to discuss the effect of Redux or exercise than it is to talk about the "effect" of obesity.

4.1 *How Much Do Parents Matter?*

Though some of the "interesting" questions in economics might admit of a meaningful causal (or other) interpretation, one often hopes for more explanation than is provided in several of the examples in *Freakonomics*. Indeed, the obesity example above is arguably a bit clearer than the question they pursue in two chapters—"how much do parents really matter?"

Let me begin by stating that there is much I agree with in the chapters:

(1) The advice of "parenting experts" should be met with deep skepticism at best.

(2) The research in Julie Berry Cullen, Brian A. Jacob, and Levitt (2003) justifies a longer discussion than the two pages the book provides. It is qualitatively several notches above most of the research done on school choice, evaluates an actual (not a hypothetical) policy, and is a marvel of clarity and honest reporting of results. They exploit a randomized lottery that determines whether some children get to "choose" the public school they attend. Perhaps surprisingly, those who win the lottery perform no better on the usual measures of "performance" (and sometimes worse) than lottery losers.

(3) Even though I can't come up with a simple "experiment" to test the hypothesis that "honesty may be more important to good parenting than spanking is to bad parenting" (p. 171), I think honesty is a good strategy (even if it didn't have a causal effect on a child's test scores; the salient issues have to do with ethical behavior.)

In the setup to this discussion, Levitt and Dubner begin with a summary of previous work: "A long line of studies, including research into twins who were separated at birth, had already concluded that genes alone are responsible for perhaps 50 percent of a child's personality and abilities" (p. 154). As any student of regression knows, this statement doesn't even make sense unless the world is of the simplest sort imagined by regression's eugenicist forefathers.[38] Obviously as careful as Cullen, Jacob, and Levitt (2003) is, it is completely silent on this unanswerable question.

Much of the chapter, a discussion of Roland G. Fryer and Levitt (2004b) (pp. 163–76), is a long hike in a forest of confusion. Surprisingly, the authors use it to deliver a short tutorial about regression analysis ("knowing what you now know about regression analysis, conventional wisdom, and the art of parenting") and they spend a great deal of time discussing what is essentially a pair of "kitchen sink regressions"

[38] Suppose the world was as complicated as Behavior $= G + E - G \cdot E$ where G is some index summarizing "genes" and E is some index summarizing "environment." In this simple example, the fraction of variation in behavior induced by differences in genes isn't separable from the environment—indeed, the effect of genes is a *function* of the environment. In some environments, introducing differences in genes would introduce little change in behavior, and in some environments it would change behavior a lot. For a useful discussion that addresses this and other related points, see Heckman (1995).

(regressions with enormous numbers of covariates) from appendix A-2 of Fryer and Levitt (2004b) using data from the Early Childhood Longitudinal Study. In their presentation, they invite the reader to consider several things that are positively correlated with a child's test scores (presumably after conditioning on a huge laundry list of [unmentioned] variables):

> the child has highly educated parents, the child's parents have high socio-economic status, the child's birth mother was thirty or older at the time of her first child's birth, the child had low birth weight, the child's parents speak English in the house, the child is adopted, the child's parents are involved in the PTA, the child has many books in his home.

As well as things that "aren't correlated"[39]:

> the child's family is intact, the child's parents recently moved into a better neighborhood, the child's mother didn't work between birth and kindergarten, the child attended Head Start, the child's parents regularly take him to museums, the child is regularly spanked, the child frequently watches television, the child's parents regularly read to him every day.

At some points, they seem to suggest that the results of this analysis speak to nothing causal: "the ECLS data don't say that books in the house [or any of the variables in their analysis] *cause* high test scores; it says only that the two are correlated." Elsewhere they seem to suggest the opposite:

> Now a researcher is able to tease some insights from this very complicated set of data. He can line up all the children who share many characteristics—all the circuit boards that have their switches flipped in the same direction—and then pinpoint the single characteristic they *don't* share. This is how he isolates the true impact of that single switch—and, eventually, of every switch—becomes manifest (p. 162).

I would maintain that, even allowing for the simplification of the argument for a general audience, this is a bad description of what makes for credible research—nothing is being severely tested.

For example, whatever one thinks of Head Start, accepting Dubner and Levitt's observation that "according to the [kitchen sink regression using] ECLS data, Head Start does nothing for a child's future test scores" seems unwise at best. The research design can not credibly support that inference. To make this clear, consider other inferences (though not discussed in *Freakonomics*) from the same regressions. Why not, for example, observe that participation in WIC (Women, Infants, and Children) significantly lowers test scores?[40] Perhaps such assistance actively harms children. I would argue that the good reason for avoiding *that* inference works just as well as a rationale for avoiding the inference they *do* make about Head Start: there is no reason to believe that (conditional on the other nonrandomly assigned regressors) that a coefficient in a kitchen sink regression reliably informs us about causation in any sense.

Again, even kitchen sink regressions have their place: one can sometimes make a case for inclusion of scores of covariates in some very selected contexts. However, an algorithm which allows the researcher to decide which coefficients represent "causal" effects and which ones are regression artifacts *after* one has seen the regression output is unlikely to result in much progress in understanding. It is the very antithesis of a severe test.

4.2 *Can Regression Help Distinguish "Cause" from "Consequence"?*

Chapter 6, "Perfect Parenting, Part II; or: Would a Roshanda by Any Other Name Smell as Sweet?" begins this way:

Levitt thinks he is onto something with a new paper about black names. He wanted to

[39] I think they mean "so imprecisely estimated that a null hypothesis of no correlation can not be rejected using standard procedures."

[40] From Appendix A-2, when the dependent variable is math scores the coefficient on WIC is −0.120 with a standard error (0.020). When the dependent variable is reading scores, the coefficient on WIC is –0.104 with a standard error (0.021).

know if someone with a distinctly black name suffers an economic penalty. His answer—contrary to other recent research—is no. But now he has a bigger question: Is black culture a cause of racial inequality or is it a consequence? For an economist, even for Levitt, this is new turf—"quantifying culture" he calls it. As a task, he finds it thorny, messy, perhaps impossible, and deeply tantalizing (p. 177).

As with eugenics, the history of social science suggests that scholarly research into race that makes extensive use of correlations should be taken with a large grain of salt.[41] When talking about race, it is my view that being *clear* about what is meant is even more important.

As someone who is frequently called upon as an econometric "script doctor" to "fix the econometrics" of some existing paper which is putatively about "causation," I have found it useful to begin with two seemingly simple questions:

(1) What is y, the outcome, you wish to explain?

(2) What are your key x variables and what potential "causes" or "interventions" are you interested in?

As a practical matter, the inability to provide a simple reply to the question is a good predictor of my inability to understand the empirical work.

The above quote from *Freakonomics* is in a chapter which, inter alia, discusses research from Fryer and Levitt (2004a) and (far more briefly) Marianne Bertrand and Sendhil Mullainathan (2004). In Fryer and Levitt (2004a), much of the evidence on whether "black names" are cause or consequence comes from two types of regressions involving their measure of "black culture" the "Black Name Index" (BNI).[42]

It is not clear whether the BNI is an x or a y: superficially, it would appear that they run the regressions "both ways": in one set of regressions, BNI is an independent variable, in a second set, it plays the role of a dependent variable. As is well appreciated, this is a problem even when it occurs in different literatures (John Kennan 1989).

Further inspection suggests that this is not strictly the case: in the first set of regressions (see table 2 "Determinants of Name Choices Among Blacks," of Fryer and Levitt 2004a) the dependent variable is the BNI of a given child, and the explanatory variables are a number of things, many of which are presumably correlated with outcomes (mother's age at time of birth, father's age at time of birth, months of prenatal care, percentage of Black babies in zip code, per capita income in the birth place, parental education, etc.). In another set (table 3, "The Relationship Between Names and Life Outcomes"), BNI becomes an explanatory variable and the dependent variables are outcomes such as "percent Black in residential zip code as an adult," years of education (the woman herself), the woman's age at first birth, etc.

Fryer and Levitt (2004a) are forthright in admitting that their evidence is consistent with a number of very plausible (but very different) alternatives that are consistent with their regressions but not necessarily with their conclusion: "With respect to this

[41] The most notorious example perhaps is the controversy over the 1840 census that involved the putative negative correlation between the number of "insane and idiotic colored persons" living in a state and the proportion that were slaves. The data, which are still available today from the ICPSR show that incidence of insanity was far, far lower in the South, and the implication for the debate on slavery was clear (Gerald N. Grob 1978). (A far different version of "acting white" is mentioned several times in *Freakonomics*.)

[42] I am stipulating, of course, that Levitt and Fryer's measure of "distinctiveness" of a "Black" name (BNI)—crudely put a function of the relative frequency with which a specific name is chosen for black children and the relative frequency with which the same name is chosen for white children—provides a measure of whatever "culture" is. A lot of nonobvious measurement issues arise. A few moments reflection, for instance, makes clear that the level of "black culture" is, by definition, a function of "white" culture although one doubts this research design would have found much appeal as a study of "white culture." Second, a white man named Maurice Ravel might be measured as have more black culture than a black man named Paul Robeson Jr. regardless of their actual "culture" if Maurice was relatively more popular among blacks than Paul.

particular aspect of distinctive Black culture, we conclude that carrying a black name is primarily a consequence rather than a cause of poverty and segregation."

I have no wish to dispute their conclusion; rather, I wish to suggest that there is no configuration of the data of which I am aware which would credibly support the view held by Fryer and Levitt *and not support very different alternatives.* In short, this is because it is very difficult to know what is being asked and what would constitute an answer. Put differently, there is at least one ill-posed question floating about. Is it possible to talk meaningfully about "manipulating" culture? (And if one could, would one want to?)[43] Might reasonable people agree on some variable or policy that served exclusively to manipulate black culture and affected economic outcomes only through its effect on "culture?" It is not even clear that "culture" and "economic outcomes" or "racial inequalities" are distinct entities. Indeed, as the word is often understood, culture often includes the distribution of "economic outcomes." For instance, one might remark: "the fact that Bill Gates earns several times more in a year than the sum earned by all Chicago Public School teachers is a distressing fact about U.S. culture."

Further muddling the issue is the way Levitt and Dubner discuss studies such as Bertrand and Mullainathan (2004):

> So how does it matter if you have a very white name or a very black name?. . . In a typical audit study, a researcher would send two identical (and fake) résumés, one with a traditionally minority-sounding name, to potential employers. The "white" résumés have always gleaned more job interviews . . . . The implication is that black-sounding names carry an economic penalty. Such studies are tantalizing but severely limited, for they can't explain *why* [someone with a black sounding name like] DeShawn didn't get the call (p. 186).

Even if one agrees to stipulate that a limitation of such studies is their inability to explain "why" (although the concern for "why" is not pressed very hard elsewhere in *Freakonomics* regarding the motives of Sumo wrestlers or school teachers, for example), Bertrand and Mullainathan (2004) clearly explain that they are *not* interested in the lifetime "economic cost" of a black sounding name—which is not obviously an interesting or well-posed question. Rather they are interested in "experimentally manipulat[ing] [an employer's] perception of race." In contrast to the thought experiment of manipulating a person's "culture" or "black name," Bertrand and Mullainathan seem to ask a well-posed question: it is much easier to conceive of a salient experiment manipulating "perceptions" than a salient experiment manipulating the naming decisions of parents. One can argue that the causal effect of manipulating perceptions of race is "uninteresting" on a number of grounds, not the least of which is that the manipulation itself doesn't suggest an intervention we might wish to undertake as a society. On the other hand, in contrast with some experiments in "experimental economics" their study is embedded rather more deeply in "real life" than experiments that occur in a lab. Nonetheless, the question seems well-posed and may be answerable with regression, even if one wants to argue that it is uninteresting on other grounds.[44]

Second, although Dubner and Levitt are correct to argue that studies involving résumé randomization are unlikely to provide convincing evidence on "*why* DeShawn gets fewer callbacks," it is not clear what a satisfactory explanation of "why" would look like.

[43] The paper seems to suggest that they have the usual "manipulationist" version of cause in mind. For example, there is a brief mention of the fact that there are no obvious instrumental variables which would be of no moment *unless* they conceived of a potential manipulation.

[44] The fact that employers call back "Jamals" much less frequently than "Johns" may not be based solely on self-conscious racial hatred, but might reflect "only" "statistical discrimination" (i.e., employers are merely acting as sophisticated econometricians, extracting all the useful information not provided by a résumé about the likely productivity of workers based on their first names, and then choosing based exclusively on "merit") or some other mechanism (although this may be of little comfort to Jamal or John). See Thacher (2002) for a thoughtful discussion of the issues involved in "profiling."

It is even harder to understand how the type of regressions performed in Fryer and Levitt (2004a) would, in principle, be relevant to this discussion. (Again, they might be, but the link is not obvious to me.) Perhaps like Dr. Pangloss, we could trace Jamal's bad luck with employers to necessity: it is necessary for this to be the case, for us to be able to live in this the best of all possible worlds.

More generally, reasoning backward from a single effect (not calling back Jamal) to a "cause" (why employers don't call Jamal) in social science is generally fraught with peril; people are complicated enough that there is rarely a single answer to the question "why"—often there are many interacting "reasons." Absent some fairly articulated model of how the world works, it seems difficult even to know what would constitute a good answer. A severe test of the claim seems even more unlikely. Moreover, it often seems that putative explanations of "why" some complex human interaction occurs are frequently used as a device to *end* a debate just at the point when the issue begins to get interesting. If *X* is *the* reason *Y* occurs, why look further? Many readers might be familiar with this aspect of some answers to "why" questions: one thinks of a parent who tries to end a long conversation with a child who, in response to a parent's increasingly complicated responses, keeps asking "Why?" Again it is not that a satisfactory answer to such question is not desirable: it just seems like way too much to hope from a small set of OLS regressions.

Finally, in asking a regression to distinguish "black culture" as a *cause* from black culture as a *consequence* of economic conditions, we are very far from the types of questions I discussed in section 3. But there is no clear discussion in *Freakonomics* of what question is being asked nor the "ground rules" that we might use to determine when the question has been answered satisfactorily. It is possible that the question is well posed, but at a minimum, it is not very obvious. After reading *Freakonomics* and the original source material, I haven't gained any understanding of the issues involved or even how to think about what are the answerable questions.

4.3 *Why a Transparent Research Design Helps—Abortion as a "Cause"*

For me the most confusing section of *Freakonomics* is the discussion of "Why do drug dealers live with their moms?" and "Where have all the criminals gone?" Between them, the chapters contain references to scores of articles of varying degrees of scholarship. Much of the former chapter discusses Levitt's work with sociologist Sudhir Alladi Venkatesh who collected a large amount of detailed data on one Chicago gang. For those surprised as to why gang members don't frequently live in the nicest homes in town, it will be a useful corrective. (For an earlier discussion that covers similar ground, see Peter Reuter, Robert MacCoun, and Patrick Murphy 1990.) The discussion also includes the conclusions of some very careful work by Douglas V. Almond, Kenneth Y. Chay, and Michael Greenstone (2003) that document the key role that hospital integration in Mississippi played in improving the appalling infant mortality rate of black children—before integration, these infants were often left to die of very preventable causes such as diarrhea and pneumonia.

Much of the chapter on "where have all the criminals gone?" deals with Romania's abortion ban, which I have discussed elsewhere (DiNardo 2006a). This chapter also includes the controversial material on whether "abortion lowers crime rates."

As a purely personal matter, given the long, deep, and ugly relationship between statistical analysis and eugenics, what might emerge from this debate seems too meager to justify the effort on this subject. I don't find the question "interesting."[45] Merely

[45] Eugenics, often popular among "progressive" members of the elite, was a leading motive for the development of regression. Sir Francis Galton, who gave us the word "regression," was an ardent eugenicist. For example, what is now the "Galton Laboratory, Department of Human Genetics and Biometry" at University College London, was originally named the "Galton Laboratory of National Eugenics."

participating in the discussion one runs the risk of coarsening the debate on how we treat the poor—the usual the target of eugenic policies.[46] Caveats aside, here goes.

In their original article, John J. Donohue and Levitt (2001) cite two possible "theories" about the consequences of abortion legalization. Neither of them fit well into the framework described in section 3.[47] Donohue and Levitt (2001) discuss two possible mechanisms at length.

Donohue and Levitt (2001) first argue that "The simplest way in which legalized abortion reduces crime is through smaller cohort sizes" (p. 386).[48] While possibly "simple," it is amazingly difficult to articulate clearly in a regression framework where the unit of observation is the individual. At its core this hypothesis appears to include the implicit assertion that among other things, my mother's decision not to abort the fetal John DiNardo caused some other children's propensity to commit crime to increase. (Although it should be said, it clearly raised mine!) Such effects are difficult to identify, even in the easiest cases (Charles F. Manski 1993).

A far more subtle mechanism is distinct from the first, although it could certainly

[46] Indeed, the debate has grown coarser. Consider this partial transcript and discussion by Levitt of remarks by William Bennett. (*For clarity, in what follows, text and transcript material from the blog is in italics.*) Bennett, a former government official, after appearing to dismiss the "abortion–crime" hypothesis in *Freakonomics*, remarked in a talk show that:

BENNETT: Well, I don't think it is either, I don't think it is either, because, first of all, there is just too much that you don't know. But I do know that it's true that if you wanted to reduce crime, you could—if that were your sole purpose, you could abort every black baby in this country, and your crime rate would go down. That would be an impossible, ridiculous, and morally reprehensible thing to do, but your crime rate would go down.

Everyone agrees with Bennett that "it would be a morally reprehensible thing to do." On the other hand, his premise that "you could abort every black baby in this country and the crime rate would go down" is unsupportable at best, racist at worst. Levitt's thoughts on the subject (as well as a transcript of the relevant portion of Bennetts's remarks) are available at the website http://freakonomics.blogs.nytimes.com/2005/09/30/bill-bennett-and-freakonomics/. For what it's worth, Levitt's remarks are a mixture of what strike me as reasonable assertions and others that are confusing at best, wrong at worst. For example, consider Levitt's points 6 and 7:

6) If we lived in a world in which the government chose who gets to reproduce, then Bennett would be correct in saying that "you could abort every black baby in this country, and your crime rate would go down." Of course, it would also be true that if we aborted every white, Asian, male, Republican, and Democratic baby in that world, crime would also fall. Immediately after he made the statement about blacks, he followed it up by saying, "That would be an impossible, ridiculous, and morally reprehensible thing to do, but your crime rate would go down." He made a factual statement (if you prohibit any group from reproducing, then the crime rate will go down), and then he noted that just because a statement is true, it doesn't mean that it is desirable or moral. That is, of course, an incredibly important distinction and one that we make over and over in Freakonomics.

7) There is one thing I would take Bennett to task for: first saying that he doesn't believe our abortion–crime hypothesis but then revealing that he does believe it with his comments about black babies. You can't have it both ways.

As far as I can tell, Levitt's statement about lowering the level of crime by abort¬ing Native American, Republican, . . . fetuses is a non sequitor at best. Bennett is clearly talking about the rate of crime. I can only make sense of the statement by construing it to mean that ridding the planet of human life would eliminate crime (at least that caused by humans). As to the rest of the explanation, Levitt gives no reason to believe that *"if we lived in a world in which the government chose who gets to reproduce, then Bennett would be correct in saying that 'you could abort every black baby in this country, and your crime rate would go down.'"*

Contrary to Levitt's claim, I do not think it necessary to believe that the termination of black fetuses would lower the crime rate even if the "causal effect of abortion legalization" in the United States had been a reduction in crime. As I explain below, even if one stipulates that crime reduction was a causal effect of abortion legalization in the United States this would tell us nothing about the causal consequences of aborting black (or any) fetuses.

[47] One could conceive of cases where abortion might be thought of (for better or worse) as a treatment: that is generally true when the subject of interest was child-bearing women (not their fetuses). The question of what happened to the welfare of women who are given the choice of having an abortion relative to those that have been denied such choice, is well posed. One merely would seek to compare a group of women given the opportunity to have an abortion to those who did not. Even putting aside the serious ethical questions, this is much easier said than done (and indeed is the subject of much of the pre–Donohue and Levitt (2001) work by economists on the consequences of abortion legalization).

[48] I have not been able to figure out what role this hypothesis plays in the empirical work. See Christopher L. Foote and Christopher F. Goetz (forthcoming) for an attempt to make sense of this; they come to different conclusions than Donohue and Levitt (2001).

interact with it. "Far more interesting from our perspective is the possibility that abortion has a disproportionate effect on the births of those who are most at risk of engaging in criminal behavior" (Donohue and Levitt 2001, p. 386).

Even if we could agree that the effect of abortion legalization is independent of other aspects of the society (access to birth control, women's rights in other spheres, etc.), for anyone who has given the problem of "missing data" some thought, it is difficult to be sanguine about the possibility of inferring much about the criminal propensities of those who are never born. Even in the context of a medical RCT, the analogous problem of attrition is often distressingly difficult to cope with. Moreover, the problem is so difficult that in the RCT one often abandons hope of modeling nonresponse or sample selection and seeks merely to bound the difference between the treated and control groups (Joel L. Horowitz and Manski 1998).

Moreover, as Donohue and Levitt (2001) observe, there are many mechanisms besides abortion either to stop the "criminogenic" fetus from being born or to prevent the child from becoming a "criminal" once born: "Equivalent reductions in crime could in principle be obtained through alternatives for abortion, such as more effective birth control, or providing better environments for those children at greatest risk for future crime" (p. 415).

Ironically, this observation points to a lot of (unasked) questions which are interesting and might conceivably be put to a severe test. The focus in *Freakonomics* unfortunately is elsewhere:

> How, then, can we tell if the abortion–crime link is a case of causality rather than simply correlation? One way to test the effect of abortion on crime would be to measure crime data in the five states where abortion was made legal before the Supreme Court extended abortion rights to the rest of the country . . . And indeed, those early-legalizing states saw crime begin to fall earlier than the other forty-five states and the District of Columbia. Between 1988 and 1994, violent crime in the early-legalizing states fell 13 percent compared to the other states; between 1994 and 1997, their murder rates fell 23 percent more than those of the other states (p. 140).

Of the identification strategies employed in this literature, this is the most transparent. To understand what is going on, assume that pre-Roe legalization provided a Brandeisian natural experiment. Instead of the individual being the unit of observation, think of each state as a sort of identical petri dish to which a drop of abortion legalization is being added. Fifteen to twenty five years later, the petri dishes will be checked again to see how much per capita crime is occurring. If legalization had been an actual experiment (perhaps run by a dictator), we might have expected half the states to be legalizers and the other half to never legalize (assume that items in the petri dishes can't jump into other petri dishes.) That of course did not happen. In this case, the experimenter added a drop of legalization to five states in 1970, and then added a drop to the remaining states a scant three years later. Of course, it wouldn't be clear that even in this experiment you could detect an "effect" on crime, unless the effect was large relative to the variation across the petri dishes expected in the absence of any experiment.[49]

Though one would not know from reading *Freakonomics*, Donohue and Levitt (2001) argue that this research design is

[49] Indeed, this or similar identification strategy is employed in such work as Kerwin Kofi Charles and Melvin Stephens (2006), Jonathan Gruber, Phillip Levine, and Douglas Staiger (1999), Marianne Bitler and Madeline Zavodny (2002), as well as Joyce (2004). Gruber, Levine, and Staiger (1999) detect a rather small (and brief) effect on the total number of children born from this identification strategy. Note of course, that such an experiment would provide us essentially no information on the "mechanisms"—the "effect of abortion legalization" could be a complicated interaction of many things having little to do with selective abortion or cohort size per se. Merely the *option* of having an abortion might change outcomes for many reasons.

inadequate.[50] Consequently, much of this is beside the point. Donohue and Levitt (2001) argue that evidence from such a research design is only "suggestive."

The bulk of their argument centers on their attempts to "more systematically" analyze the relationship with an analysis of state level crime data on lagged "abortion rates."

Consider equation (1) from Donohue and Levitt (2001):

$$A_t \equiv \mathit{Effective\ Abortion}_t = \sum_a \mathit{Abortion}^*_{t-a} \frac{\mathit{Arrests}_a}{\mathit{Arrests}_{total}}.$$

They label A_t the "effective abortion rate." The "*a*" subscript denotes a particular "age" group.[51] Using data on state *s* at time *t*, they then divide this by the number of live births to get an "effective abortion ratio":

$$\mathcal{A}_{st} = \frac{A_{st}}{LB_{st}}.$$

Much of the more "systematic" evidence on the link between abortion legalization and crime is a result of regressions of the form:

$$\text{(2)} \qquad \log \mathit{Crime\ Per\ Capita}_{st} = \beta_1 \mathcal{A}_{st} + X_{st}\theta + \gamma_s + \lambda_t + \varepsilon_{st},$$

where each observation is the relevant state/year average or value. The X_{st} are a set of covariates, γ are a set of state dummy variables and λ_t are a set of year fixed effects. The ε is a random disturbance that is presumably uncorrelated with any of the regressors. Up to a constant that differs by states, absent variation in *X* or the (modified) abortion ratio, it is assumed that trends across state in crime would be the same.[52]

Stipulating that all of the data used to generate this specification are fine[53], I find it impossible to interpret the coefficients at all. In common econometric parlance, the abortion ratio is "endogenous." Indeed, some work has looked at the effect of economic and other conditions on abortion (Rebecca M. Blank, Christine C. George, and Rebecca A. London 1996): that is, something akin to *A* is the *dependent* variable in the regression. Donohue and Levitt (2001), however, spend surprisingly little time discussing the issue.[54]

What are the "ground rules" that a skeptical, but persuadable person should use for evaluating this regression? Other than that "the coefficients look reasonable," what would speak to the credibility of the research design, or what should lead me to reject it?

Not obvious is the notion that we should be reassured about the existence of an "abortion–crime" link because the OLS coefficient on *A* in a regression like equation (2) is robust to the inclusion of some covariates or slight modifications of the sample. One "intuition" that motivates investigating whether a result is "robust" to the inclusion

[50] They argue *against* the identification strategy both on a priori grounds and on ex post grounds (the implausibility of the results so obtained). In Donohue and Levitt (2001), for example, when they deploy that identification strategy, they report that "the cumulative decrease in crime between 1982–97 for early-legalizing states compared with the rest of the nation is 16.2 percent greater for murder, 30.4 percent greater for violent crime, and 35.3 percent greater for property crime. Realistically, these crime decreases are too large to be attributed to the three-year head start in the early-legalizing states." The reservations in Donohue and Levitt (2001) about the estimates generated with this identification strategy do not appear in *Freakonomics* which selectively discusses some comparison between early and late legalizing states.

[51] The asterisk appears to be undefined in the text and may be a typographical error.

[52] As it turns out, the description of the regressions in the text and the actual regressions run are not always the same. See Foote and Goetz (forthcoming).

[53] This is perhaps more than we should stipulate to: our knowledge of the number of illegal abortions today or abortions that preceded abortion legalization in the 1970s is meager at best. Moreover, Donohue and Levitt (2001) and other researchers typically do not have data on the amount of crime committed by individuals of a given age. At best one has very crude proxies. See Charles and Stephens (2006) or Joyce (2004) for discussion.

[54] In the published version of the paper, the word "endogeneity" appears only regarding a discussion of two right hand side variables—number of police and prisons—which are "lagged to minimize endogeneity." The word "exogeneity" appears in a confusing discussion about the difference between high and low abortion states (p. 401).

of a large number of explanatory variables comes from the RCT. On average, if we repeat the experiment, the answer we get from including covariates and from excluding covariates should be the same.

On the other hand, clearly it makes no sense to think of $\mathcal{A}$ as "randomly assigned." Indeed, if abortion legalization is all about "selection"—i.e., the difference in the crime propensities of those born and those not born—pure random assignment of abortion (a thought too grotesque to even contemplate) would not merely leave the statistical problem unsolved; it would answer a different (even more uninteresting) question. For example, in one version of the Donohue–Levitt story, abortion matters for crime because it is the consequence of *choice* made by women to *selectively* abort some fetuses and not others. "Random abortion" would, on the other hand, would produce no "selection effect"—studying such "random" variation in abortion ratios would be silent about the putative effects of legalizing abortion.

If thinking about the regression as an approximation to some sort of RCT doesn't help, how is one to even *assess* or interpret the specification? What covariates "should" be included? Missing from this research is *either* a similarity to the simple type of question I described in section 3 *or* an explicit model of the link between abortion legalization and cohort size. With an explicit structural model, one might in principle be able to wrap one's mind around what question is being asked. The mere presence of an explicit model might help, although we would still be faced with the task of putting it to some sort of test. Absent that, it is hard to understand why this (or similar evidence) should persuade anyone (one way or the other.)

4.4 *Catching Cheaters*

I have suggested that a focus on actual policies and their potentially predictable consequences often goes a long way to making some queries intelligible. Even where questions about causes are not the only ones in question, such considerations can remain important.

One example is the discussion entitled "What Do School Teacher's and Sumo Wrestler's Have in Common," the authors discuss some work by Levitt on detecting "teacher cheating." In the telling, the cast of heroes includes the CEO of the Chicago Public School system, and the villains include the school teachers and their labor union ("When [Duncan] took over the public schools, his allegiance lay more with the schoolchildren and their families than with teachers and their unions," p. 36.)

The basic method is to analyze the *pattern* of test answers. Answers that depart from the posited (ad hoc) data generation process are flagged as "cheating." For obvious reasons, at no point in the process are actual data on *observed* teacher cheating used. As a consequence, the algorithm described has no way of discriminating between the case in which a teacher selectively "corrects" a subset of answers for a class, from those cases in which the students (unknown to the teacher) have obtained copies of a subset of the answers, to name one (perhaps unlikely) situation. At a most basic level, of course, there is no *perfect* way to "detect teacher cheating" with statistical analysis.[55]

Indeed, the chapter indicates that the "teacher cheating" algorithm was not the sole method used to assess guilt (as one hopes) but remarks with little further curiosity that "the evidence was strong enough

[55] To make this clear, consider an analysis made by officials responsible for New York's Powerball lottery. In the March 30, 2005, drawing, a startling number of persons (110) got five out of six numbers correct. According to a news report (Jennifer 8 Lee, "Who Needs Giacomo? Bet on the Fortune Cookie," *New York Times*, May 11, 2005), past experience with the lottery had led them to believe that, in the 29 states where the game is played, the average number of winners would be more like four or five. After considering numerous hypotheses including fraud, lottery officials finally concluded that some of the winners had chosen their number on the basis of a fortune cookie. Lottery investigators finally even managed to locate the fortune cookie maker who verified that his factory had produced the fortune cookies with the winning number.

only to get rid of a dozen of them." Given the rest of the discussion, this might come as quite a surprise. Why would such a clever algorithm work so poorly in a situation when there was so much cheating?

Anything but a perfect "test" for the existence or "nonexistence" of something (a virus, for example) commits two types of error—in unhelpful terminology, Type I and Type II. I find the legal metaphor the easiest way to remember the distinction. The legal system in the United States putatively attempts to minimize Type I error—sending an innocent person to jail. Type II error is the opposite mistake—exonerating the guilty. In practice, there is a trade-off between the two types. One way to avoid Type II error is to declare everyone guilty; declare everyone innocent and one avoids Type I error at the expense of Type II error.

This is of course relevant if one is interested in the causal impact of implementing a cheating detection algorithm. Here I focus only on the narrowest causal question: how many innocents are punished by such a system? There are others to be sure.

If the fact that only a "handful" were caught is a surprise to the reader, it wouldn't be a surprise to those familiar with the findings of Amos Tversky and Daniel Kahneman (1974) who argue that people are frequently inattentive to "base rates" (although that interpretation is the subject of a lively debate.) The canonical problem can be illustrated by making a few assumptions about the algorithm discussed in *Freakonomics*. Suppose that the probability of one's cheating being detected, given that you cheat is 0.90—the probability of Type II error is 0.1. Also assume that the probability the algorithm incorrectly identifies you as a cheater when you are not is .06—Type I error. Further suppose that 4 percent of teachers cheat—this is the crucial "base rate." Slightly more formally:

$\Pr(D|C) \equiv \Pr($Detected Cheating by Algorithm|Engaged in Cheating$) = .90$

$\Pr(D|{\sim}C) \equiv \Pr($Detected Cheating by Algorithm|Not Engaged in Cheating$) = .06$

$\Pr(C) \equiv \Pr($Engaged in Cheating$) = .04$.

I wasn't able to locate the actual numbers in *Freakonomics* and the ones I have chosen seem a bit optimistic for the algorithm they describe (albeit a bit pessimistic about the fraction of cheating teachers). If they were correct, however, it would explain why only a handful of those identified by the algorithm were finally identified as cheaters—despite the large pool of potential cheaters. Many statistically naive readers might conclude that virtually all of those identified as guilty were indeed guilty. The test looks pretty accurate. Few detected cheaters are innocent, and cheaters have a good chance of being caught. However, even in this example, of the roughly 9 percent of teachers classified as cheating on the basis of the algorithm, the majority (about 62 percent) would actually be innocent. This strikes me as a frighteningly high percentage, but perhaps others will disagree.[56] A more thoughtful analysis would go even further: does it treat different but morally homogeneous groups differently? It would almost certainly give one a moment's pause if an algorithm was only (or mostly) able to detect cheating among the lowest paid teachers with the most difficult students, as well as being scarcely able to detect cheating among the most affluent. *Freakonomics* unfortunately discusses none of these issues.

5. *The Power of Theory*

If what we mean by theory is an articulation of the premises of the questions we are

[56] The calculation is:

$$\begin{aligned}1-\Pr(C|D) &= 1-\left\{\frac{(\Pr(D|C)\cdot\Pr(C)}{\Pr(D|C)\cdot\Pr(C)+\Pr(D|{\sim}C)\cdot(1-\Pr(C))}\right\}\\ &= 1-\frac{.9(.04)}{.9(.04)+.06(.96)}\\ &= 1-0.385\\ &= 0.615.\end{aligned}$$

It should also be noted that the usual way to minimize this problem is to test the teacher more than once: this works, of course, only in the highly improbable case that one might consider errors from the proposed procedure as independent.

attempting to answer, it is fair to say that a little bit of theory can go a long way.

Despite *Freakonomics*'s various encomiums to the "powerful toolkit of economics" (apart from "regression" which manages to rise to a level no higher than "art"), I could detect very little in the research that depended on "economic theory" in anything but the most superficial way. Perhaps that is all for the best. Cartwright (2007d) makes a convincing argument that the "trouble with [more mathematically sophisticated theory models] is not that [they are] too rigorous, but rather . . . not rigorous enough." Unrealistic assumptions are necessary in all theorizing, but the problem is that the alleged "tendencies" isolated in such models are usually model-dependent in a way that vitiates any "external validity."[57] It's one thing to find that a model helps illuminate some small aspect of human responses in a few contexts. In my field of labor economics, for example, the canonical labor–leisure model might help me predict how working mothers in female headed households adjust their hours of market work to a small change in an existing government welfare program, but I wouldn't use the model to tell me much else about working mothers.

If were one to judge Economics by reading *Freakonomics* alone, however, it would appear that some economists are successfully going after much bigger game, with economic theory leading the way. How else could a study about professional Sumo speak to anything but the unusual context of Sumo? Presumably some "model" takes us from a finding for wrestlers to something of more general interest. On the other hand, all we learn about economic theory is that it appears to be the premise that "incentives matter." Whatever enthusiasm one might have for the power of that insight, it is not clear what an incentive "is." The helpful index to the book lists the following: *incentives, bright line versus murky, as a cornerstone of modern life, criminal, definitions of, discovery and understanding, economic, of experts, invention and enactment of, moral, negative versus positive, power of, of real estate agents, schemes based on, of schoolteachers, social, study, tinkering with, trade-offs inherent in.*

Indeed, in Dubner and Levitt's hands, the assertion that incentives are the "cornerstone of modern life" often comes off as a two part tautology. The first part of the tautology is: "when incentives matter, they matter." The second part of the tautology is that when incentives don't matter, it is because of "moral incentives."

Less than a theory, perhaps, it describes a world view that evokes a sort neo-Skinnerian behaviorism that in popular writing was most cogently demolished by Noam Chomsky (1971). For example, it was quite easy for me to get confused, when reading *Freakonomics*, about whether negative and positive incentives were merely synonyms for the Skinnerian notions of negative and positive reinforcement.

Perhaps I read more into the use of the word incentives than is there. However consider Dubner and Levitt's description of the "typical economist's view" of incentives:

> Economists love incentives. They love to dream them up and enact them, study them, and tinker with them. The typical economist believes the world has not yet invented a problem that he can not fix if given a free hand to design the proper incentive scheme. His solution may not always be pretty—it may involve coercion or exorbitant penalties or the violation of civil liberties—but the original problem, rest assured, will be fixed. An incentive is a bullet, a lever, a key: an often tiny object with astonishing power to change a situation (p. 20).

Nonetheless, as elastic as the notion of incentives is, I think it is still way too narrow. Speaking of B. F. Skinner's views of the power of "reinforcement," Chomsky's (1971) words about B. F. Skinner seem to apply with equal force to Dubner and

[57] This gloss is clearly inadequate. See Cartwright (2007d). See also Alexandrova (2006).

Levitt's typical economist: "Humans are not merely dull mechanisms formed by a history of reinforcement and behaving predictably with no intrinsic needs apart from the need for physiological satiation. Then humans are not fit subjects for manipulation, and we will seek to design a social order accordingly."

I do not mean to suggest that Dubner and Levitt believe that humans are merely "dull mechanisms" formed only by a history of "incentives." My point is merely that as a framework for understanding human behavior the typical economists' focus on "incentives" ignores much that is important. James M. Buchanan, for example, writes that "any person's ideal situation is one that allows him full freedom of action and inhibits the behavior of others so as to force adherence to his own desires. That is to say, each person seeks mastery over a world of slaves."[58]

I hope I never live to meet this sort of *homo economicus*. At a minimum, Buchanan's ideal appears more like a dystopian nightmare, notwithstanding the fact that his description of *homo economicus* follows logically from his premises about human motivation. It also highlights the problem with the view that "theory is evidence too" (June O'Neill as cited in Angus Deaton 1996)—most commonly our models are perfect environments to do "experiments" on people we would never hope to meet, in situations in which they could never find themselves. If even "the laws of physics lie"[59] and physicists often fruitfully use different and mutually inconsistent models, it seems far more modesty about economic theory/models is warranted than *Freakonomics* seems to suggest. This is not to deny that simple economic models can be put to good use, but economic theory—whatever it is—shouldn't be epistemologically privileged. Even if social scientists were to learn something "deep," "fundamental," or "primordial" about human behavior that was previously unknown to the skilled novelist, it is unlikely to inform us very much about the type of policies we might like to pursue. As Lecam (1977) observed about a genuine example from real science, "even those physicists who are most fascinated by the kinetic theory of gases would hesitate to use it to compute the size of wood beams for their own abode." Simply put questions about the predictable consequences of our actions are not well-answered by untested or untestable insights from some "general" economic theory; rather we might learn a little about the predictable consequences of our actions—if we are lucky—by formulating ideas that can be put to a test.

[58] Buchanan (1975), page 92; Chapter 6, "The Paradox of 'Being Governed'" at Buchanan (1999) http://www.econlib.org/LIBRARY/Buchanan/buchCv7Contents.html.

[59] See Cartwright (1983), especially essay 2 ("The Truth Doesn't Explain Much") and essay 3 ("Do the Laws of Physics State the Facts?") for a discussion of the case of physics.

References

Alexandrova, Anna. 2006. "Connecting Economic Models to the Real World: Game Theory and the FCC Spectrum Auctions." *Philosophy of the Social Sciences*, 36(2): 173–92.

Almond, Douglas V., Kenneth Y. Chay, and Michael Greenstone. 2003. "Civil Rights, the War on Poverty, and Black–White Convergence in Infant Mortality in Mississippi." Unpublished.

Anonymous. 1986. "On the Physical Death of Jesus Christ." *Journal of the American Medical Association*, 255(20): 2752–60.

Armon, Carmel, and Randolph W. Evans. 2005. "Addendum to Assessment: Prevention of Post-lumbar Puncture Headaches: Report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology." *Neurology*, 65(4): 510–12.

Berry, Scott M., and Joseph B. Kadane. 1997. "Optimal Bayesian Randomization." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4): 813–19.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94(4): 991–1013.

Bitler, Marianne, and Madeline Zavodny. 2002. "Did Abortion Legalization Reduce the Number of Unwanted Children? Evidence from Adoptions." *Perspectives on Sexual and Reproductive Health*, 34(1): 25–33.

Blank, Rebecca M., Christine C. George, and Rebecca A. London. 1996. "State Abortion Rates: The Impact of Policies, Providers, Politics, Demographics, and Economic Environment." *Journal of Health*

Economics, 15(5): 513–53.

Brenner, B. 2005. “Did Jesus Die of Pulmonary Embolism?” *Journal of Thrombosis and Haemostasis*, 3(9): 2130–31.

Buchanan, James M. 1975. *The Limits of Liberty: Between Anarchy and Leviathan*. Chicago: University of Chicago Press.

Buchanan, James M. 1999. *The Limits of Liberty: Between Anarchy and Leviathan*. Indianapolis: Liberty Fund, Inc.

Card, David Edward, and Alan B. Krueger. 1995. *Myth and Measurement: The New Economics of the Minimum Wage*. Princeton: Princeton University Press.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford and New York: Oxford University Press.

Cartwright, Nancy. 1999. “The Vanity of Rigour in Economics: Theoretical Models and Galilean Experiments.” Centre for the Philosophy of Natural and Social Science Discussion Paper, no. 43/99.

Cartwright, Nancy. 2003a. “Causation: One Word; Many Things.” Causality: Metaphysics and Methods Technical Report, no. CTR 07-03.

Cartwright, Nancy. 2003b. “Counterfactuals in Economics: A Commentary.” In *Proceedings from INPC 2003: Explanation and Causation: Topics in Contemporary Philosophy*, *Volume 4*, ed. O’Rourke. Cambridge and London: MIT Press.

Cartwright, Nancy. 2007a. “Causation: One Word, Many Things.” In *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge and New York: Cambridge University Press, 9–23.

Cartwright, Nancy. 2007b. “Counterfactuals in Economics: A Commentary.” In *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge and New York: Cambridge University Press, 236–61.

Cartwright, Nancy. 2007c. *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge and New York: Cambridge University Press.

Cartwright, Nancy. 2007d. “The Vanity of Rigour in Economics: Theoretical Models and Galilean Experiments.” In *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge and New York: Cambridge University Press, 217–35.

Center for Health Statistics. 1998. “Data Matters: Multiple Cause of Death in California.” U.S. Department of Health and Human Services, Centers for Disease Control and Prevention Technical Report.

Charles, Kerwin Kofi, and Melvin Stephens, Jr. 2006. “Abortion Legalization and Adolescent Substance Use.” *Journal of Law and Economics*, 49(2): 481–505.

Chomsky, Noam. 1971. “The Case against B.F. Skinner.” *New York Review of Books*, December 30.

Cournot, M., J. C. Marquié, D. Ansiau, C. Martinaud, H. Fonds, J. Ferrières, and J. B. Ruidavets. 2006. “Relation between Body Mass Index and Cognitive Function in Healthy Middle-Aged Men and Women.” *Neurology*, 67(7): 1208–14

Cullen, Julie Berry, Brian A. Jacob, and Steven D. Levitt. 2003. “The Effect of School Choice on Student Outcomes: Evidence from Randomized Lotteries.” NBER Working Papers, no. 10113.

Deaton, Angus. 1996. “Letters from America: The Minimum Wage.” *Newsletter of the Royal Economic Society*, 95: 13.

DeVany, Arthur. Forthcoming. “Steroids, Home Runs and the Law of Genius.” *Economic Inquiry*.

Dickens, Charles. 1854. *Hard Times*. London: Bradbury & Evans.

DiNardo, John. 2006a. “Freakonomics: Scholarship in the Service of Storytelling.” *American Law and Economics Review*, 8(3): 615–26.

DiNardo, John. 2006b. “Natural Experiments.” In *The New Palgrave Dictionary of Economics*, ed. Steven N. Durlauf and Lawrence E. Blume. Houndmills, U.K. and New York: Palgrave Macmillan.

Donohue, John J., III, and Steven D. Levitt. 2001. “The Impact of Legalized Abortion on Crime.” *Quarterly Journal of Economics*, 116(2): 379–420.

Edgeworth, Francis Ysidro. 1885. “The Calculus of Probabilities Applied to Psychic Research, I.” *Proceedings of the Society for Psychical Research*, 3: 190–99.

Edgeworth, Francis Ysidro. 1887. “The Calculus of Probabilities Applied to Psychic Research, II.” *Proceedings of the Society for Psychical Research*, 4: 189–208.

Edwards, W. D., W. J. Gabel, and F. E. Hosmer. 1986. “On the Physical Death of Jesus Christ.” *Journal of the American Medical Association*, 255(11): 1455–63.

Fagot, Anne M. 1980. “Probabilities and Causes: On Life Tables, Causes of Death, and Etiological Diagnoses.” In *Probabalistic Thinking, Thermodynamics and the Interaction of the History and Philosophy of Science*, ed. Jakko Hintikka, David Gruender, and Evandro Agazzi. Studies in Epistemology, Logic, Methodology, and Philosophy of Science, vol. 146. Dordrecht; Boston and London: D. Reidel Publishing Company, 41–104.

Food and Drug Administration. 1997. “FDA Announces Withdrawal of Fenfluramine and Dexfenfluramine (Fen-Phen).” Press Release: September 15.

Foote, Christopher L., and Christopher F. Goetz. Forthcoming. “The Impact of Legalized Abortion on Crime: Comment.” *Quarterly Journal of Economics*.

Freedman, David A. 1999. “From Association to Causation: Some Remarks on the History of Statistics.” *Statistical Science*, 14(3): 243–58.

Fryer, Roland G., Jr., and Steven D. Levitt. 2004a. “The Causes and Consequences of Distinctively Black Names.” *Quarterly Journal of Economics*, 119(3): 767–805.

Fryer, Roland G., Jr., and Steven D. Levitt. 2004b. “Understanding the Black–White Test Score Gap in the First Two Years of School.” *Review of Economics and Statistics*, 86(2): 447–64.

Gosling, C. G. 1987. “Comments on ‘The Physical Death of Jesus Christ.’” *Journal of Biocommunication*, 14(4): 2–3.

Granger, Clive. 1986. “Statistics and Causal Inference: Comment.” *Journal of the American Statistical Association*, 81(396): 967–68.

Graunt, John. 1676. *Natural and Political Observations Mentioned in a Following Index and Made upon the Bills of Mortality. With Reference to the Government,* *Religion, Trade, Growth, Air, Diseases, and the Several Changes of the Said City*, First edition. John Martyn.

Graunt, John. *Natural and Political Observations Mentioned in a Following Index and Made upon the Bills of Mortality. With Reference to the Government, Religion, Trade, Growth, Air, Diseases, and the Several Changes of the Said City.*, first ed., John Martyn, 1676. Rendered into HTML Format by Ed Stephan 25 Jan 96. Accessed June 1, 2005 from http://www.ac.wwu.edu/~stephan/Graunt/bills .html.

Grob, Gerald N. 1978. *Edward Jarvis and the Medical World of Nineteenth-Century America*. Knoxville: University of Tennessee Press.

Gronniger, Jerome Timothy. 2003. "Fat and Happy: Dissecting the Obesity–Mortality Relationship." Unpublished.

Gronniger, Jerome Timothy. 2006. "A Semiparametric Analysis of the Relationship of Body Mass Index to Mortality." *American Journal of Public Health*, 96(1): 173–78.

Gruber, Jonathan, Phillip Levine, and Douglas Staiger. 1999. "Abortion Legalization and Child Living Circumstances: Who Is the 'Marginal Child?'" *Quarterly Journal of Economics*, 114(1): 263–91.

Guerry, André-Michel. 1883. *Essai sur la statistique moral de la France. A Translation of André-Michel Guerry's Essay on the Moral Statistics of France: A Sociological Report to the French Academy of Science* edited and Translated by Hugh P. Whitt and Victor W. Reinking, 2002.

Hacking, Ian. 1965. *The Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hacking, Ian. 1975. *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge: Cambridge University Press.

Hacking, Ian. 1983a. "Nineteenth Century Cracks in the Concept of Determinism." *Journal of the History of Ideas*, 44(3): 455–75.

Hacking, Ian. 1983b. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

Hacking, Ian. 1988. "Telepathy: Origins of Randomization in Experimental Design." *Isis*, 79(3): 427–51.

Hacking, Ian. 1990. *The Taming of Chance*. Ideas in Context, no. 17. Cambridge: Cambridge University Press.

Hacking, Ian. 1995. *Rewriting the Soul: Multiple Personality and the Sciences of Memory*. Princeton: Princeton University Press.

Hacking, Ian. 2000. *The Social Construction of What?* Cambridge and London: Harvard University Press.

Hanke, Lewis. 1935. *The First Social Experiments in America: A Study in the Development of Spanish Indian Policy in the Sixteenth Century*. Cambridge, Mass. and London: Harvard University Press.

Harville, David A. 1975. "Experimental Randomization: Who Needs It?" *American Statistician*, 29 (1): 27–31.

Heckman, James J. 1995. "Lessons from the Bell Curve." *Journal of Political Economy*, 103(5): 1091–1120.

Heckman, James J. 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective." *Quarterly Journal of Economics*, 115(1): 45–97.

Heckman, James J. 2005. "The Scientific Model of Causality." *Sociological Methodology*, 35(1): 1–98.

Heckman, James J., and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, 73(3): 669–738.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association*, 81(396): 945–60.

Horowitz, Joel L., and Charles F. Manski. 1998. "Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations." *Journal of Econometrics*, 84(1): 37–58.

Joyce, Ted. 2004. "Further Tests of Abortion and Crime." NBER Working Papers, no. 10564.

Kennan, John. 1989. "Simultaneous Equations Bias in Disaggregated Econometric Models." *Review of Economic Studies*, 56(1): 151–56.

Keynes, John Maynard. 1921. *A Treatise on Probability*. London: Macmillan.

Laplace, Pierre Simon de. *Essai Philosophique sur les Probabilités*, sixth ed., New York: John Wiley & Sons, 1795. Translated as "Philosophical Essay on Probabilities" from the Sixth French Edition by Frederick Wilson Truscott and Frederick Lincoln Emory and First U.S. edition Published in 1902.

Lecam, Lucien. 1977. "A Note on Metastatistics or 'An Essay toward Stating a Problem in the Doctrine of Chances.'" *Synthese*, 36(1): 133–60.

Levitt, Steven D., and Stephen J. Dubner. 2005. *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. New York: Harper Collins.

Levitt, Steven D., and Stephen J. Dubner. 2006a. "Freakonomics 2.0." http://freakonomics.blogs.nytimes.com/2006/09/20/freakonomics-20/. Accessed November 1, 2007.

Levitt, Steven D., and Stephen J. Dubner. 2006b. "Hoodwinked." *New York Times Magazine*, January 8.

Lillard, Lee A. 1998. "The Market for Sex: Street Prostitution in Los Angeles." Unpublished.

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies*, 60(3): 531–42.

Mayo, Deborah G. 1996. *Error and the Growth of Experimental Knowledge*. Science and Its Conceptual Foundations series. Chicago: University of Chicago Press.

Mayo, Deborah G., and Michael Kruse. 2002. "Principles of Inference and Their Consequences." In *Foundations of Bayesianism*, ed. D. Corfield and J. Williamson. Applied Logic Series, vol. 24. Dordrecht: Kluwer Academic, 381–404.

Moffitt, Robert. 2005. "Remarks on the Analysis of Causal Relationships in Population Research." *Demography*, 42(1): 91–108.

National Center for Health Statistics. 2006. "Instructions for Classifying Underlying Cause-of-Death." NCHS Instruction Manual Part 2-a.

Pearl, Judea. 1997. "The New Challenge: From a

Century of Statistics to the Age of Causation." *Computing Science and Statistics*, 29(2): 415–23.

Pearson, Karl. 1892. *The Grammar of Science*. A. London and C. Black.

Pearson, Karl. 1930. *Life, Letters and Labours of Francis Galton*. Vol. 3A, *Correlation, Personal Identification and Eugenics*. Cambridge: Cambridge University Press.

Peirce, Charles Sanders. 1958. *Collected Papers*. Vol. 7–8, ed. A. Burks. Cambridge: Harvard University Press.

Reid, Sue Titus. 1985. *Crime and Criminology*, Second edition. New York: Holt, Rinehart and Winston.

Reiss, Julian, and Nancy Cartwright. 2004. "Uncertainty in Econometrics: Evaluating Policy Counterfactuals." In *Economic Policy under Uncertainty: The Role of Truth and Accountability in Policy Advice*, ed. P. Mooslechner, H. Schuberth, and M. Schürz. Cheltenham, U.K. and Northampton, Mass.: Elgar, 204–32.

Reuter, Peter, Robert MacCoun, and Patrick Murphy. 1990. "Money from Crime: A Study of the Economics of Drug Dealing in Washington, D.C." RAND Research Report, no. R-3894-RF

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics*, 6(1): 34–58.

Sakula, A. 1991. "A Hundred Years of Lumbar Puncture: 1891–1991." *Journal of the Royal College of Physicians of London*, 25(2): 171–75.

Saliba, W. R. 2006. "Did Jesus Die of Pulmonary Embolism?" *Journal of Thrombosis and Haemostasis*, 4(4): 891–92.

Savage, Leonard J. 1961. "Discussion." In *The Foundations of Statistical Inference: A Discussion*, ed. G. A. Barnard and D. R. Cox. London and Colchester: Spottiswoode Ballantyne and Company, 62–103.

Savage, Leonard J., M. S. Bartlett, G. A. Barnard, D. R. Cox, E. S. Pearson, C. A. B. Smith et al. "The Foundations of Statistical Inference: A Discussion." In G. A. Barnard and D. R. Cox, eds., *The Foundations of Statistical Inference: A Discussion*, Meuthen's Monographs on Applied Probability and Statistics Spottiswoode Ballantyne & Co. Ltd. London & Colchester 1962. A Disucssion Opened by L. J. Savage at the Joint Statistics Seminar of Birbeck and Imperial Colleges. Discussants also Include H. Ruben, I. J. Good, D. V. Lindley, P. Armitage, C. B. Winsten, R. Syski, E. D. Van Rest and G. M. Jenkins.

Scheiber, Noam. 2007. "How *Freakonomics* Is Ruining the Dismal Science." *The New Republic*, April 2.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Simon, Herbert A. 1953. "Causal Ordering and Identifiability." In *Studies in Econometric Method*, ed. W. Hood and T. Koopmans. Cowles Commission for Research in Economics Monograph, no. 14. New York: Wiley, 49–74.

Siskind, Frederic B. 1977. "Minimum Wage Legislation in the United States: Comment." *Economic Inquiry*, 15(1): 135–38.

Stigler, Stephen M. 1978. "Mathematical Statistics in the Early States." *Annals of Statistics*, 6(2): 239–65.

Suppes, Patrick. 1982. "Arguements for Randomizing." *Philosophy of Science Association Proceedings*, 2: 464–75.

Swijtink, Zeno G. 1982. "A Bayesian Justification of Experimental Randomization." *Philosophy of Science Association Proceedings*, 1: 159–68.

Thacher, David. 2001. "Policing Is Not a Treatment: Alternatives to the Medical Model of Policy Research." *Journal of Research in Crime and Delinquency*, 38(4): 387–415.

Thacher, David. 2002. "From Racial Profiling to Racial Equality: Rethinking Equity in Police Stops and Searches." Gerald R. Ford School of Public Policy Working Paper, no. 02-006.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgement under Uncertainty: Heuristics and Biases." *Science*, 185(4157): 1124–31.

Ur Rehman, H. 2005. "Did Jesus Christ Die of Pulmonary Embolism? A Rebuttal." *Journal of Thrombosis and Haemostasis*, 3(9): 2131–33.

Voltaire. 1796. *The History of Candid; or All for the Best*. London: C. Cooke.

Welch, Finis R.. 1974. "Minimum Wage Legislation in the United States." *Economic Inquiry*, 12(3): 285–318.

Welch, Finis R. 1976. "Minimum Wage Legislation in the United States." In *Evaluating the Labor Market Effects of Social Programs*, ed. Orley Ashenfelter and James Blum. Princeton: Princeton University Press.

Welch, Finis R. 1977. "Minimum Wage Legislation in the United States: Reply." *Economic Inquiry*, 15(1): 139–42.

Williams, Richard H., Bruno D. Zumbo, Donald Ross, and Donald W. Zimmerman. 2003. "On the Intellectual Versatility of Karl Pearson." *Human Nature Review*, 3: 296–301.

Zellner, Arnold. 1984. "Causality and Econometrics." In *Basic Issues in Econometrics*, ed. A. Zellner. Chicago and London: University of Chicago Press, 35–74.