Institute for Policy Research
Northwestern University
Working Paper Series

WP-10-07

Policy Analysis with Incredible Certitude

Charles F. Manski
Department of Economics and Institute for Policy Research
Northwestern University

Version: May 2011

DRAFT

Abstract

Analyses of public policy regularly express certitude about the consequences of alternative policy choices. Yet policy predictions often are fragile, with conclusions resting on critical unsupported assumptions or leaps of logic. Then the certitude of policy analysis is not credible. I develop a typology of incredible analytical practices and gives illustrative cases. I call these practices *conventional certitude*, *dueling certitudes*, *conflating science and advocacy*, *wishful extrapolation*, *illogical certitude,* and *media overreach.*

Acknowledgments: *This paper was presented in February 2011 as a Leverhulme Lecture at the Institute for Fiscal Studies, in the framework of my Leverhulme Visiting Professorship at University College London. The research was supported in part by National Science Foundation grant SES-0911181. I am grateful for comments from Teresa Delgado, Joel Horowitz, Richard Lempert, Francesca Molinari, John Mullahy, Daniel Nagin, James Poterba, and two referees. I have also benefited from the opportunity to present parts of this work in seminars at NYU, the Congressional Budget Office, and the University of Essex.*

Analyses of public policy regularly express certitude about the consequences of alternative policy choices. Point predictions are common and expressions of uncertainty are rare. Yet policy predictions often are fragile. Conclusions may rest on critical unsupported assumptions or on leaps of logic. Then the certitude of policy analysis is not credible.

In a research program that began with Manski (1990, 1995) and continues to develop, I have studied identification problems that limit our ability to credibly predict policy outcomes. I have argued that analysts should acknowledge ambiguity rather than feign certitude, expressing partial knowledge by reporting interval rather than point predictions of policy outcomes. And I have shown how elementary principles of decision theory may be used to make reasonable policy choices using such interval predictions. Manski (2007) exposits core ideas and Manski (2006, 2009, 2010, 2011) report findings on various classes of policy choices.

In my work to date, I have warned against analytical practices that promote incredible certitude. I have not, however, sought to classify these practices and consider them in totality. I do so here. My hope is to move future policy analysis away from incredible certitude and towards honest portrayal of partial knowledge.

To begin, I distinguish the logic and the credibility of empirical research (Section 1) and cite arguments made for certitude (Section 2). I then develop a typology of practices that contribute to incredible certitude. I call these practices *conventional certitude* (Section 3), *dueling certitudes* (Section 4), *conflating science and advocacy* (Section 5), *wishful extrapolation* (Section 6), *illogical certitude* (Section 7), and *media overreach* (Section 8). To conclude I express my vision for credible policy analysis and raise some open questions (Section 9).

1. The Logic and Credibility of Empirical Research

Whether the context be policy analysis or any other form of research, the logic of inference is summarized by the relationship:

$$\text{assumptions} + \text{data} \Rightarrow \text{conclusions}.$$

Research is illogical if it commits deductive errors. These may be mundane mistakes in computation or algebra or, more seriously, they may be non sequiturs. Non sequiturs generate pseudo conclusions and, hence, yield misplaced certitude.

Holding fixed the available data, and presuming avoidance of deductive errors, stronger assumptions yield stronger conclusions. At the extreme, one may achieve certitude by posing sufficiently strong assumptions. The fundamental difficulty of research is to decide what assumptions to maintain.[1]

Given that strong conclusions are desirable, why not maintain strong assumptions? There is a tension between the strength of assumptions and their credibility. I have called this (Manski, 2003, p. 1):

The Law of Decreasing Credibility: The credibility of inference decreases with the strength of the assumptions maintained.

This "Law" implies that analysts face a dilemma as they decide what assumptions to maintain: Stronger assumptions yield conclusions that are more powerful but less credible.

---

[1] A reader has asked what role I see for theory in the logic of inference. I take "theory" and "assumptions" to be synonyms. I mainly use the latter term, reserving the former for broad systems of assumptions.

Credibility is a subjective matter. Whereas analysts should agree on the logic of inference, they may well disagree about the credibility of assumptions. Such disagreements occur often in practice. Indeed, they may persist without resolution.

Persistent disagreements are particularly common when assumptions are *nonrefutable*; that is, when multiple contradictory assumptions are all consistent with the available data. As a matter of logic alone, disregarding credibility, an analyst can pose a nonrefutable assumption and adhere to it forever in the absence of disproof. Indeed, he can displace the burden of proof, stating "I will maintain this assumption until it is proved wrong." Analysts often do just this. An observer may question the credibility of a non-refutable assumption, but not the logic of holding on to it.

When analysts largely agree on the credibility of certain assumptions, they may refer to this agreement as "scientific consensus." Persons sometimes push the envelope and refer to a scientific consensus as a "fact" or a "scientific truth." This is overreach. Consensus does not imply truth.

2. Incentives for Certitude

In principle, a researcher can resolve the tension between the credibility and power of assumptions by posing alternative assumptions of varying credibility and determining the conclusions that follow in each case. In practice, policy analysis tends to sacrifice credibility in return for strong conclusions. Why so?

A proximate answer is that analysts respond to incentives. I have earlier put it this way (Manski, 2007, pp. 7-8):

> The scientific community rewards those who produce strong novel findings. The public, impatient for solutions to its pressing concerns, rewards those who offer simple analyses leading to unequivocal policy recommendations. These incentives make it tempting for researchers to maintain

assumptions far stronger than they can persuasively defend, in order to draw strong conclusions.

> The pressure to produce an answer, without qualifications, seems particularly intense in the environs of Washington, DC A perhaps apocryphal, but quite believable, story circulates about an economist's attempt to describe his uncertainty about a forecast to President Lyndon B. Johnson. The economist presented his forecast as a likely range of values for the quantity under discussion. Johnson is said to have replied, "Ranges are for cattle. Give me a number."

When a President as forceful as LBJ seeks a point prediction with no expression of uncertainty, it is understandable that his advisors feel compelled to comply.

Jerry Hausman, a longtime econometrics colleague, stated the incentive argument this way at a conference in 1988, when I presented in public my initial findings on interval prediction with credible assumptions: "You can't give the client a bound. The client needs a point." This comment reflects a perception that I have found to be common among economic consultants. They contend that policy makers are either psychologically or cognitively unable to cope with ambiguity. Hence, they argue that pragmatism dictates provision of point predictions, even though these predictions may not be credible.

This psychological-cognitive argument for certitude begins from the reasonable premise that policy makers, like other humans, have bounded rationality. However, I think it too strong to draw the general conclusion that "The client needs a point." It may be that some persons think in purely deterministic terms, but a considerable body of research measuring expectations shows that most make sensible probabilistic predictions when asked to do so.[2] I see no reason to expect that policy makers are less capable than ordinary

[2] The review article of Manski (2004) describes the emergence of this field of empirical research and summarizes applications ranging from worker perceptions of job insecurity and student perceptions of the returns to schooling through personal expectations of income, Social Security benefits, and returns to mutual-fund investments. Hurd (2009) and Delavande, Giné, and McKenzie (2011) subsequently review additional parts of the large literature.

people.

2.1. Support for Certitude in Philosophy of Science

The view that analysts should offer point predictions is not confined to Presidents of the United States and economic consultants. It has a long history in the philosophy of science.

Over fifty years ago, Milton Friedman expressed this perspective in an influential methodological essay. Friedman (1953) placed prediction as the central objective of science, writing (p. 5): "The ultimate goal of a positive science is the development of a 'theory' or 'hypothesis' that yields valid and meaningful (i.e. not truistic) predictions about phenomena not yet observed." He went on to say (p. 10):

> The choice among alternative hypotheses equally consistent with the available evidence must to some extent be arbitrary, though there is general agreement that relevant considerations are suggested by the criteria 'simplicity' and 'fruitfulness,' themselves notions that defy completely objective specification.

Thus, Friedman counselled scientists to choose one hypothesis, even though this may require the use of "to some extent . . . arbitrary" criteria. He did not explain why scientists should choose a single hypothesis out of many. He did not entertain the idea that scientists might offer predictions under the range of plausible hypotheses that are consistent with the available evidence.

The idea that a scientist should choose one hypothesis among those consistent with the data is not peculiar to Friedman. Researchers wanting to justify adherence to a particular hypothesis sometime refer to *Ockham's Razor*, the medieval philosophical notion that "Plurality should not be posited without necessity." The *Encyclopaedia Britannica* gives the usual modern interpretation, stating:[3] "The principle

[3] *Encyclopaedia Britannica Online*.
<http://www.britannica.com/EBchecked/topic/424706/Ockhams-razor>. Accessed June 25, 2010.

gives precedence to simplicity; of two competing theories, the simplest explanation of an entity is to be preferred." The philosopher Richard Swinburne writes (Swinburne, 1997, p. 1):

> I seek…to show that—other things being equal—the simplest hypothesis proposed as an explanation of phenomena is more likely to be the true one than is any other available hypothesis, that its predictions are more likely to be true than those of any other available hypothesis, and that it is an ultimate a priori epistemic principle that simplicity is evidence for truth.

The choice criterion offered here is as imprecise as the one given by Friedman. What do Britannica and Swinburne mean by "simplicity?"

However one may operationalise the various philosophical dicta for choosing a single hypothesis, the relevance of philosophical thinking to policy analysis is not evident. In policy analysis, knowledge is instrumental to the objective of making good decisions. When philosophers discuss the logical foundations and human construction of knowledge, they do so without posing this or another explicit objective. Does use of criteria such as "simplicity" to choose one hypothesis among those consistent with the data promote good policy making? This is the relevant question for policy analysis. To the best of my knowledge, thinking in philosophy has not addressed it.

3. Conventional Certitude

John Kenneth Galbraith popularized the term *conventional wisdom*, writing (Galbraith, 1958, chap. 2): "It will be convenient to have a name for the ideas which are esteemed at any time for their acceptability, and it should be a term that emphasizes this predictability. I shall refer to these ideas henceforth as the conventional wisdom." In 2010, Wikipedia nicely put it this way: [4]

[4] http://en.wikipedia.org/wiki/Conventional_wisdom. Accessed May 8, 2010.

> Conventional wisdom (CW) is a term used to describe ideas or explanations that are generally accepted as true by the public or by experts in a field. The term implies that the ideas or explanations, though widely held, are unexamined and, hence, may be reevaluated upon further examination or as events unfold. . . . . Conventional wisdom is not necessarily true.

I shall similarly use the term *conventional certitude* to describe predictions that are generally accepted as true, but that are not necessarily true.

3.1. CBO Scoring of Pending Legislation

Governments regularly produce official forecasts of unknown accuracy. Some such forecasts become conventional certitudes. In the United States today, conventional certitude is exemplified by Congressional Budget Office (CBO) *scoring* of pending federal legislation. I will use CBO scoring as an extended case study.

The CBO was established in the Congressional Budget Act of 1974. Section 402 states (Committee on the Budget, United States House of Representatives, 2008, p. 39-40):

> The Director of the Congressional Budget Office shall, to the extent practicable, prepare for each bill or resolution of a public character reported by any committee of the House of Representatives or the Senate (except the Committee on Appropriations of each House), and submit to such committee—(1) an estimate of the costs which would be incurred in carrying out such bill or resolution in the fiscal year in which it is to become effective and in each of the 4 fiscal years following such fiscal year, together with the basis for each such estimate;

This language has been interpreted as mandating the CBO to provide point predictions (aka scores) of the budgetary impact of pending legislation. Whereas the 1974 legislation called for prediction five years into the future, the more recent practice has been to forecast ten years out. CBO scores are conveyed in letters

that the Director writes to leaders of Congress and chairs of Congressional committees. They are not accompanied by measures of uncertainty, even though legislation often proposes complex changes to federal law, whose budgetary implications must be difficult to foresee.

Serious policy analysts recognize that scores for complex legislation are fragile numbers, derived from numerous untenable assumptions. Considering the related matter of scoring the effects of tax changes on federal revenues, Auerbach (1996) wrote (p. 156): "in many instances, the uncertainty is so great that one honestly could report a number either twice or half the size of the estimate actually reported."

Credible scoring is particularly difficult to achieve when proposed legislation may significantly affect the behaviour of individuals and firms, by changing the incentives they face to work, hire, make purchases, and so on. Academic economists, who have the luxury of studying subjects for years, have worked long and hard to learn how specific elements of public policy affect individual and firm behaviour, but with only limited success. CBO analysts face the more difficult challenge of forecasting the effects of the many policy changes that may be embodied in complex legislation, and they must do so under extreme time pressure.

In light of the above, it is remarkable that CBO scores have achieved broad acceptance within American society. In our highly contentious political age, the scores of pending legislation have been eagerly awaited by both Democratic and Republican Members of Congress. And media reports largely take them at face value.

3.1.1. Scoring the Patient Protection and Affordable Care Act of 2010. CBO scoring of the major health care legislation enacted in 2009–2010 illustrates well current practice. Throughout the legislative process, Congress and the media paid close attention to the scores of alternative bills considered by various Congressional committees. A culminating event occurred on March 18, 2010 when the CBO, assisted by staff of the Joint Committee on Taxation (JCT), provided a preliminary score for the combined consequences

of the Patient Protection and Affordable Care Act and the Reconciliation Act of 2010. CBO director Douglas Elmendorf wrote to House of Representatives Speaker Nancy Pelosi as follows (Elmendorf, 2010a, p.2): "CBO and JCT estimate that enacting both pieces of legislation . . . . would produce a net reduction of changes in federal deficits of $138 billion over the 2010–2019 period as a result of changes in direct spending and revenue."

Anyone seriously contemplating the many changes to federal law embodied in this legislation should recognize that the $138 billion prediction of deficit reduction can be no more than a very rough estimate. However, the twenty-five page letter from Elmendorf to Pelosi expressed no uncertainty and did not document the methodology generating the prediction.

Media reports largely accepted the CBO scores as fact, the hallmark of conventional certitude. For example, a March 18, 2010 *New York Times* article documenting how CBO scoring was critical in shaping the legislation reported (Herszenhorn, 2010): "A preliminary cost estimate of the final legislation, released by the Congressional Budget Office on Thursday, showed that the President got almost exactly what he wanted: a $940 billion price tag for the new insurance coverage provisions in the bill, and the reduction of future federal deficits of $138 billion over 10 years." The *Times* article did not question the validity of the $940 and $138 billion figures.

Interestingly, the certitude that CBO expressed when predicting budgetary impacts ten years into the future gave way to considerable uncertainty when considering longer horizons. In his letter to Pelosi, Director Elmendorf wrote (p. 3):

> Although CBO does not generally provide cost estimates beyond the 10-year budget projection period, certain Congressional rules require some information about the budgetary impact of legislation in subsequent decades. . . . . Therefore, CBO has developed a rough outlook for the decade following the 2010-2019 period. . . . Our analysis indicates that H.R. 3590, as passed by the Senate, would reduce federal budget deficits over the ensuing decade relative to those projected

> under current law—with a total effect during that decade that is in a broad range between one-quarter percent and one-half percent of gross domestic product (GDP).

Further insight into the distinction that the CBO drew between the ten-year and longer horizons emerges from a March 19 letter that the Director wrote to Congressman Paul Ryan. He wrote (Elmensdorf, 2010b, p. 3):

> A detailed year-by-year projection, like those that CBO prepares for the 10-year budget window, would not be meaningful over a longer horizon because the uncertainties involved are simply too great. Among other factors, a wide range of changes could occur—in people's health, in the sources and extent of their insurance coverage, and in the delivery of medical care (such as advances in medical research, technological developments, and changes in physicians' practice patterns) —that are likely to be significant but are very difficult to predict, both under current law and under any proposal.

Thus, the CBO was quick to acknowledge uncertainty when asked to predict the budgetary impact of the health care legislation more than ten years out, phrasing its forecast as a "broad range" rather than as a point estimate.

Why did the CBO express uncertainty only when making predictions beyond the ten-year horizon? Longer term predictions may be more uncertain than shorter-term ones, but it is not reasonable to set a discontinuity at ten years, with certitude expressed up to that point and uncertainty only beyond it. The potential behavioural changes cited by Elmendorf in his letter to Ryan, particularly changes in insurance coverage and in physicians' practice patterns, could occur soon after passage of the new legislation.

Having discussed scoring practices with various CBO personnel, I am confident that Director Elmendorf recognized the ten-year prediction sent to Speaker Pelosi was at most a rough estimate. However, he felt compelled to adhere to the established CBO practice of expressing certitude when providing ten-year predictions, which play a formal role in the Congressional budget process.

A similar tension between unofficial recognition of uncertainty and official expression of certitude

is evident in Foster (2010), a United States Department of Health and Human Services (HHS) document that reports independent estimates of the budgetary implications of the health care legislation. The HHS document, like the CBO letter, provides point estimates with no accompanying measures of uncertainty. However, HHS verbally cautions that the estimates are uncertain, stating (p. 19):

> Due to the very substantial challenges inherent in modelling national health reform legislation, our estimates will vary from those of other experts and agencies. Differences in results from one estimating entity to another may tend to cause confusion among policy makers. These differences, however, provide a useful reminder that all such estimates are uncertain and that actual future impacts could differ significantly from the estimates of any given organization. Indeed, the future costs and coverage effects could lie outside of the range of estimates provided by the various estimators.

3.1.2. Credible Interval Scoring. Since its creation in 1974, the CBO has established and maintained an admirable reputation for impartiality. Perhaps it is best to leave well enough alone and have the CBO continue to express certitude when it scores pending legislation, even if the certitude is only conventional rather than credible.

I understand the temptation to leave well enough alone, but I think it unwise to try to do so. I would like to believe that Congress will make better decisions if the CBO provides it with credible forecasts of budgetary impacts. Whether or not this is a reasonable expectation, I worry that the prevailing norm to take CBO scores seriously will eventually break down. Conventional certitudes that lack foundation cannot last indefinitely. I think it better for the CBO to preemptively act to protect its reputation than to have some disgruntled group in Congress or the media declare that the emperor has no clothes.

It has been suggested that, when performing its official function of scoring legislation, the CBO is required to provide no more than a single point estimate. For example, in a 2005 article, CBO analyst

Benjamin Page wrote (Page, 2005, p. 437):

> Scoring has a specific meaning in the context of the federal budget process. Under the Congressional Budget Act of 1974, the Congressional Budget Office provides a cost estimate, or "score," for each piece of legislation that is reported by a Congressional committee. . . . By its nature, the cost estimate must be a single point estimate.

However, my reading of the Congressional Budget Act suggests that the CBO is not prohibited from presenting measures of uncertainty when performing its official function of scoring.[5]

Presuming that the CBO can express uncertainty, how should it do so? There is no uniquely correct answer to this question, and alternatives may range from verbal descriptors to provision of probabilistic predictions. Aiming to balance simplicity and informativeness, I suggest provision of interval predictions

---

[5] A document on the Congressional budget describes the process for modifying the CBO scoring procedure. Committee on the Budget, United States House of Representatives (2008) states (p. 156):

> These budget scorekeeping guidelines are to be used by the House and Senate Budget Committees, the Congressional Budget Office, and the Office of Management and Budget (the ''scorekeepers'') in measuring compliance with the Congressional Budget Act of 1974 (CBA), as amended, and GRH 2 as amended. The purpose of the guidelines is to ensure that the scorekeepers measure the effects of legislation on the deficit consistent with established scorekeeping conventions and with the specific requirements in those Acts regarding discretionary spending, direct spending, and receipts. These rules shall be reviewed annually by the scorekeepers and revised as necessary to adhere to the purpose. These rules shall not be changed unless all of the scorekeepers agree. New accounts or activities shall be classified only after consultation among the scorekeepers. Accounts and activities shall not be reclassified unless all of the scorekeepers agree.

This passage indicates that the CBO cannot unilaterally change its scoring procedure, but that change can occur if the various "scorekeepers" agree.

of the budgetary impacts of legislation. Stripped to its essentials, computation of an interval prediction just requires that the CBO produce two scores for a bill, a low score and a high score, and report both. If the CBO must provide a point prediction for official purposes, it can continue to do so, with some convention used to locate the point within the interval prediction.

This idea is not entirely new. A version of it was briefly entertained by Alan Auerbach in the article mentioned earlier. Auerbach wrote "Presumably, forecasters could offer their own subjective confidence intervals for the estimates they produce, and this extra information ought to be helpful for policymakers." He went on to caution "However, there is also the question of how well legislators without formal statistical training would grasp the notion of a confidence interval."

The CBO need not describe its interval predictions as statistical confidence intervals. The main sources of uncertainty about budgetary impacts are not statistical in nature. They are rather that analysts are not sure what assumptions are realistic when they make predictions. A CBO interval prediction would be more appropriately described as the result of a sensitivity analysis, the sensitivity being to variation across alternative plausible assumptions.

3.1.3. Can Congress Cope with Uncertainty? I have received disparate reactions when I have suggested interval CBO scoring to other economists and policy analysts. Academics usually react positively, but persons who have worked within the federal government tend to be sceptical. Indeed, former CBO director Douglas Holtz-Eakin told me that he expected Congress would be highly displeased if the CBO were to provide it with interval scores.

The arguments that I have heard against interval scoring have been of two types. One is the psychological-cognitive argument discussed in Section 2. The other begins by observing that Congress is not an individual, but rather a collection of persons with differing beliefs and objectives who must jointly make policy choices in a political decision process. Thus, Congressional decision making should be

conceptualized as a noncooperative game. In a game, the usual economic presumption that information is a good need not apply. Players possessing more complete information may adopt strategies that yield better or worse outcomes. It depends on the structure of the game and the objectives of the players.

Viewing Congressional policy choice as a game legitimately counters wishful thinking that a Congress receiving credible scores would necessarily make better decisions. However, game theory does not generically support the contention that current CBO practice should be preferred to credible scoring. Whether game theory can generate useful normative conclusions about scoring is an open question.

3.1.4. British Norms. Curiously, the antipathy towards measurement of government forecast uncertainty evident in Washington, DC is not as apparent in London, United Kingdom. Since 1996, the Bank of England has regularly published probabilistic inflation forecasts presented visually as a "fan chart;" see Britton, Fisher, and Whitley (1998). The fan chart provides a succinct and informative measurement of forecast uncertainty.

More recently, it has become official government to require an Impact Assessment (IA) for legislation submitted to Parliament. The originating government agency must state upper and lower bounds for the benefits and costs of the proposal, as well as a central estimate. The government specifically asks that a sensitivity analysis be performed, providing this guidance to agencies in its Impact Assessment Toolkit: "The 'Summary Analysis and Evidence' page of the IA template asks you to highlight key assumptions underpinning the analysis; sensitivities of the estimates to changes in the assumptions; and risks, and how significant they might be, to policy delivery.[6]

The norms for government forecasting in the United Kingdom thus differ from those in the United

[6] See www.bis.gov.uk/assets/biscore/better-regulation/docs/i/11-518-impact-assessment-toolkit.pdf, page 63, accessed April 20, 2011. I am grateful to Teresa Delgado of the United Kingdom Department for Environment, Food and Rural Affairs for bringing this to my attention.

States. I do not have a clear sense why this is the case.

4. Dueling Certitudes

A rare commentator who rejected the CBO prediction that the health care legislation would reduce the budget deficit by $138 billion was Douglas Holtz-Eakin, its former Director. He dismissed the CBO score and offered his own, writing (Holtz-Eakin, 2010): "In reality, if you strip out all the gimmicks and budgetary games and rework the calculus, a wholly different picture emerges: The health care reform legislation would raise, not lower, federal deficits, by $562 billion." The CBO and Holtz-Eakin scores differed hugely, by $700 billion. Yet they shared the common feature of certitude. Both were presented as exact, with no expression of uncertainty.

This provides an example of *dueling certitudes*. Hotz-Eakin did not assert that the CBO committed a deductive error. He instead questioned the assumptions maintained by the CBO in performing its derivation, and he asserted that a very different result emerges under alternative assumptions that he preferred. Each score may make sense in its own terms, each combining available data with assumptions to draw logically valid conclusions. Yet the two scores are sharply contradictory.

Anyone familiar with the style of policy analysis regularly practiced within the Washington Beltway, and often well beyond it, will immediately recognize the phenomenon of dueling certitudes. To illustrate, I will draw on my experience a decade ago chairing a National Research Council committee on illegal drug policy (National Research Council, 1999, 2001).

4.1. The RAND and IDA Reports on Illegal Drug Policy

During the mid-1990s, two studies of cocaine control policy played prominent roles in discussions of federal policy towards illegal drugs. One was performed by analysts at RAND (Rydell and Everingham, 1994) and the other by analysts at the Institute for Defense Analyses (IDA) (Crane, Rivolo, and Comfort, 1997). The two studies posed similar hypothetical objectives for cocaine-control policy, namely reduction in cocaine consumption in the United States by one percent. Both studies predicted the monetary cost of using certain policies to achieve this objective. However, RAND and IDA used different assumptions and data to reach dramatically different policy conclusions.

The authors of the RAND study specified a model of the supply and demand for cocaine that aimed to formally characterize the complex interaction of producers and users and the subtle process through which alternative cocaine-control policies may affect consumption and prices. They used this model to evaluate various demand-control and supply-control policies and reached this conclusion (p. xiii):

> The analytical goal is to make the discounted sum of cocaine reductions over 15 years equal to 1 percent of current annual consumption. The most cost-effective program is the one that achieves this goal for the least additional control-program expenditure in the first projection year. The additional spending required to achieve the specified consumption reduction is $783 million for source-country control, $366 million for interdiction, $246 million for domestic enforcement, or $34 million for treatment (see Figure S.3). The least costly supply-control program (domestic enforcement) costs 7.3 times as much as treatment to achieve the same consumption reduction.

The authors of the IDA study examined the time-series association between source-zone interdiction activities and retail cocaine prices. They reached an entirely different policy conclusion (p. 3):

> A rough estimate of cost-effectiveness indicates that the cost of decreasing cocaine use by one percent through the use of source-zone interdiction efforts is on the order of a few tens of millions

> of dollars per year and not on the order of a billion dollars as reported in previous research [the RAND study]. The differences are primarily attributed to a failure in the earlier research to account for the major costs imposed on traffickers by interdiction operations and overestimation of the costs of conducting interdiction operations.

Thus, the IDA study specifically rebutted a key finding of the RAND study.

When they appeared, the RAND and IDA studies drew attention in the ongoing struggle over federal funding of drug control activities. The RAND study was used to argue that funding should be shifted towards drug treatment programs and away from activities to reduce drug production or to interdict drug shipments. The IDA study, undertaken in part as a re-analysis of the RAND research, was used to argue that interdiction activities should be funded at present levels or higher.

Lee Brown, then director of The Office of National Drug Control Policy (SNDCP), used the RAND study to argue for drug treatment at a Congressional hearing (Subcommittee on National Security, International Affairs, and Criminal Justice, 1996, p. 61):

> Let me now talk about what we know works in addressing the drug problem. There is compelling evidence that treatment is cost-effective and provides significant benefits to public safety. In June 1994, a RAND Corporation study concluded that drug treatment is the most cost-effective drug control intervention.

In a subsequent hearing specifically devoted to the IDA study, Subcommittee Chair William Zeliff used the study to argue for interdiction (Subcommittee on National Security, International Affairs, and Criminal Justice, 1998, p. 1):

> We are holding these hearings today to review a study on drug policy, a study we believe to have significant findings, prepared by an independent group, the Institute for Defense Analysis, at the request of Secretary of Defense Perry in 1994. . . .[T]he subcommittee has questioned for some time the administration's strong reliance on treatment as the key to winning our Nation's drug war, and

> furthermore this subcommittee has questioned the wisdom of drastically cutting to the bone interdiction programs in order to support major increases in hard-core drug addiction treatment programs. The basis for this change in strategy has been the administration's reliance on the 1994 RAND study.

4.1.1. The National Research Council Assessment. At the request of SNDCP, the National Research Council Committee on Data and Research for Policy on Illegal Drugs (henceforth, the Committee) assessed the RAND and IDA studies. This assessment was published as a committee report (National Research Council, 1999).

After examining the assumptions, data, methods, and findings of the two studies, the Committee concluded that neither constitutes a persuasive basis for the formation of cocaine control policy. The Committee summarized its assessment of the RAND study as follows (p. 28):

> The RAND study is best thought of as conceptual research offering a coherent way to think about the cocaine problem. The study documents a significant effort to identify and model important elements of the market for cocaine. It represents a serious attempt to formally characterize the complex interaction of producers and users and the subtle process through which alternative cocaine-control policies may affect consumption and prices. The study establishes an important point of departure for the development of richer models of the market for cocaine and for empirical research applying such models to evaluate alternative policies.
>
> However, the RAND study does not yield usable empirical findings on the relative cost-effectiveness of alternative policies in reducing cocaine consumption. The study makes many unsubstantiated assumptions about the processes through which cocaine is produced, distributed, and consumed. Plausible changes in these assumptions can change not only the quantitative findings reported, but also the main qualitative conclusions of the study. . . . . Hence the study's findings do

not constitute a persuasive basis for the formation of cocaine control policy.

It summarized its assessment of the IDA study this way (p.43):

> The IDA study is best thought of as a descriptive time-series analysis of statistics relevant to analysis of the market for cocaine in the United States. The study makes a useful contribution by displaying a wealth of empirical time-series evidence on cocaine prices, purity, and use since 1980. Efforts to understand the operation of the market for cocaine must be cognizant of the empirical data. The IDA study presents many of those data and calls attention to some intriguing empirical associations among the various series.
>
> However, the IDA study does not yield useful empirical findings on the cost-effectiveness of interdiction policies to reduce cocaine consumption. Major flaws in the assumptions, data, and methods of the study make it impossible to accept the IDA findings as a basis for the assessment of interdiction policies. For example, the conclusions drawn from the data rest on the assumption that all time-series deviations in cocaine price from an exponential decay path should be attributed to interdiction events, not to other forces acting on the market for cocaine. Numerous problems diminish the credibility of the cocaine price series developed in the study, and an absence of information prevents assessment of the procedure for selecting interdiction events.

Thus, the Committee concluded that neither the RAND nor the IDA study provides a credible estimate of what it would cost to use alternative policies to reduce cocaine consumption in the United States.

When I think now about the RAND and IDA studies, I consider their many specific differences to be less salient than their shared lack of credibility. Each study may be coherent internally, but each rests on such a fragile foundation of weak data and unsubstantiated assumptions as to undermine its findings. To its great frustration, the NRC committee had to conclude that the nation should not draw even the most tentative policy lessons from either study. Neither yields usable findings.

What troubles me most about both studies is their injudicious efforts to draw strong policy

conclusions. It is not necessarily problematic for researchers to try to make sense of weak data and to entertain unsubstantiated conjectures. However, the strength of the conclusions drawn in a study should be commensurate with the quality of the evidence. When researchers overreach, they not only give away their own credibility but they diminish public trust in science more generally. The damage to public trust is particularly severe when researchers inappropriately draw strong conclusions about matters as contentious as drug policy.

4.2. Analysis without Certitude: Sentencing and Recidivism

I would like to be able to discuss an analysis of illegal drug policy that does not line up on one side or the other of the debate between treatment and law enforcement. However, the dueling certitudes illustrated by the RAND and IDA reports are characteristic of the study of drug policy. Indeed, dueling certitudes are common in analysis of criminal justice policy more broadly.

It need not be this way. Rather than make strong unsubstantiated assumptions that yield strong incredible conclusions on one or the other side of a policy debate, analysts could aim to illuminate how the assumptions posed determine the conclusions drawn. To show how this may be accomplished, Manski and Nagin (1998) considered how sentencing of convicted juvenile offenders affects recidivism. I summarize our work here.

4.2.1. Background. Ample observational data are available on the outcomes experienced by juvenile offenders given the sentences that they actually receive. However, researchers have long debated the counterfactual outcomes that offenders would experience if they were to receive other sentences. There has been particular disagreement about the relative merits of confinement in residential treatment facilities and diversion to nonresidential treatment.

Confinement has been favoured by the "medical model" of deviance, which views deviance as symptomatic of an underlying pathology that requires treatment. In this view, the juvenile justice system should determine the needs of the child and direct the treatment resources of the state to ameliorating those needs. Confinement is thought beneficial because it enables treatment.

Non-confinement has been favoured by criminologists who are sceptical of the ability of the justice system to deliver effective treatment. This skepticism stems in part from the "labelling" view of deviance. According to this view, a constellation of negative consequences may flow from official processing of a juvenile as deviant, even with a therapeutic intent. Confinement in a residential facility may make it more likely that the person thinks of himself as deviant, may exclude him from the normal routines of life, and may place him into closer affinity with deviant others who may reinforce negative feelings the person has about himself. Given these concerns, labelling theorists have promoted the "secondary deviance" hypothesis, which holds that confinement is more likely to lead to recidivism than is nonresidential treatment.

To adjudicate between the competing predictions of the medical model and the secondary deviance hypothesis, it would be useful to perform experiments that randomly assign some offenders to confinement and others to nonresidential treatment. However, experimentation with criminal justice policy is difficult to implement. Hence, empirical research on sentencing and recidivism has relied on observational data. Analysts have typically combined the available data with the strong but suspect assumption that judges randomly sentence offenders conditional on covariates that are observable to researchers.

4.2.2. Our Analysis. Manski and Nagin (1998) implemented a cautious mode of "layered" analysis that begins with no assumptions about how judges sentence offenders and then moves from weak, highly credible assumptions to stronger, less credible ones. Exploiting the rich event-history data on juvenile offenders collected by the state of Utah, we presented several sets of findings and showed how conclusions about sentencing policy vary depending on the assumptions made.

We first reported interval predictions of recidivism obtained without making any assumptions at all about the manner in which judges choose sentences. We then presented interval predictions obtained under two alternative models of judicial decision making. The *outcome optimization* model assumes judges make sentencing decisions that minimize the chance of recidivism. The *skimming* model assumes that judges classify offenders as "higher risk" or "lower risk," sentencing only the former to residential confinement. Each model expresses an easily understood hypothesis about judicial decision making. Finally, we brought to bear further assumptions in the form of *exclusion restrictions*, which posit that specified sub-populations of offenders respond to sentencing similarly but face different sentencing selection rules.

The empirical findings turned out to depend critically on the assumptions imposed. With nothing assumed about sentencing rules or response, only weak conclusions could be drawn about the recidivism implications of the two sentencing options. With assumptions made about judicial decision making, the results were far more informative. If one believes that Utah judges choose sentences to minimize recidivism, the empirical results point to the conclusion that a policy of mandatory residential confinement would exacerbate criminality relative to one of mandatory nonresidential treatment. If one believes that judges behave in accord with the skimming model, the results suggest the opposite conclusion, namely that mandatory confinement has an ameliorative effect relative to mandatory nonresidential treatment. Imposition of an exclusion restriction strengthened each of these opposing conclusions.

Abstracting from the specifics of our juvenile-justice application, we viewed our analysis as demonstrating the value of reporting layered findings. Holding fixed the available data and presuming the absence of deductive errors, dueling certitudes can occur if analysts make conflicting strong assumptions. Reporting layered findings makes clear how the conclusions drawn depend on the assumptions posed.

5. Conflating Science and Advocacy

I earlier summarized the logic of inference by the relationship: assumptions + data ⇒ conclusions. Holding fixed the available data, the scientific method supposes that the directionality of inference runs from left to right. One poses assumptions and derives conclusions. However, one can reverse the directionality, seeking assumptions that imply predetermined conclusions. The latter practice characterizes advocacy.

Policy analysts inevitably portray their deliberative processes as scientific. Yet some analysis may be advocacy wrapped in the rhetoric of science. Studies published by certain think tanks seem almost inevitably to reach strong liberal or conservative policy conclusions. The conclusions of some academic researchers are similarly predictable. Perhaps these analysts begin without pre-conceptions and are led by the logic of inference to draw strong conclusions. Or they may begin with conclusions they find congenial and work backwards to support them.

In the late 1980s, when I visited Washington often as Director of the Institute for Research on Poverty, a thoughtful senior Congressional staffer told me that he found it prudent to view all policy analysis as advocacy. Scott Lilly remarked that he preferred to read studies performed by think tanks with established reputations as advocates to ones performed by ostensibly neutral academic researchers. He said that he often felt able to learn from the think-tank studies, because he was aware of the biases of the authors. In contrast, he found it difficult to learn from academic research by authors who may have biases but attempt to conceal them.

Milton Friedman, whom I have previously quoted, had a seductive ability to conflate science and advocacy. I give one illustration here.[7]

[7] See Krugman (2007) for a broader portrait of Friedman as scientist and advocate.

5.1. Friedman and Educational Vouchers

Proponents of educational vouchers for school attendance have argued that American school finance policy limits the options available to students and impedes the development of superior educational alternatives. Government operation of free public schools, they say, should be replaced by vouchers permitting students to choose any school meeting specified standards. The voucher idea has a long history. Tom Paine proposed a voucher plan in 1792, in *The Rights of Man*. The awakening of modern interest is usually credited to Friedman (1955,1962). His writing on the subject is emblematic of analysis that conflates science and advocacy

Friedman cited no empirical evidence relating school finance to educational outcomes. He posed a purely theoretical classical economic argument for vouchers, which began as follows (Friedman, 1955):

> The role assigned to government in any particular field depends, of course, on the principles accepted for the organization of society in general. In what follows, I shall assume a society that takes freedom of the individual, or more realistically the family, as its ultimate objective, and seeks to further this objective by relying primarily on voluntary exchange among individuals for the organization of economic activity. In such a free private enterprise exchange economy, government's primary role is to preserve the rules of the game by enforcing contracts, preventing coercion, and keeping markets free. Beyond this, there are only three major grounds on which government intervention is to be justified. One is "natural monopoly" or similar market imperfection which makes effective competition (and therefore thoroughly voluntary ex change) impossible. A second is the existence of substantial "neighborhood effects," i.e., the action of one individual imposes significant costs on other individuals for which it is not feasible to make him compensate them or yields significant gains to them for which it is not feasible to make them compensate him– circumstances that again make voluntary exchange impossible. The third derives from an ambiguity

> in the ultimate objective rather than from the difficulty of achieving it by voluntary exchange, namely, paternalistic concern for children and other irresponsible individuals.

He went on to argue that the "three major grounds on which government intervention is to be justified" may justify government supply of educational vouchers but not government operation of free public schools, which he referred to as "nationalization" of the education industry.

Repeatedly, Friedman entertained a ground for government operation of schools and then dismissed it. Here is an excerpt from his discussion of the neighbourhood-effects argument:

> One argument from the "neighbourhood effect" for nationalizing education is that it might otherwise be impossible to provide the common core of values deemed requisite for social stability. . . . This argument has considerable force. But it is by no means clear . . . . that it is valid. . . . .
>
> Another special case of the argument that governmentally conducted schools are necessary to keep education a unifying force is that private schools would tend to exacerbate class distinctions. Given greater freedom about where to send their children, parents of a kind would flock together and so prevent a healthy intermingling of children from decidedly different backgrounds. Again, whether or not this argument is valid in principle, it is not at all clear that the stated results would follow.

This passage is intriguing. Friedman cited no empirical evidence regarding neighbourhood effects, nor did he call for research on the subject. Instead, he simply stated "it is by no means clear" and "it is not at all clear" that neighbourhood effects warrant public schooling.

Rhetorically, Friedman placed the burden of proof on free public schooling, effectively asserting that vouchers are the preferred policy in the absence of evidence to the contrary. This is the rhetoric of advocacy, not science. An advocate for public schooling could just as well reverse the burden of proof, arguing that the existing educational system should be retained in the absence of evidence to the contrary. The result would be dueling certitudes.

As I have discussed in Manski (1992), a scientific analysis would have to acknowledge that

economic theory per se does not suffice to draw conclusions about the optimal design of educational systems. It would have to stress that the merits of alternative designs depend on the magnitudes and natures of the market imperfections and neighbourhood effects that Friedman noted as possible justifications for government intervention. And it would have to observe that knowledge about these matters was almost entirely lacking when Friedman wrote in the mid-1950s. Indeed, much of the needed knowledge remains lacking today.

6. Wishful Extrapolation

The Second Edition of the *Oxford English Dictionary* (*OED*) defines *extrapolation* as "the drawing of a conclusion about some future or hypothetical situation based on observed tendencies." Extrapolation in this sense is essential to the use of data in policy analysis. Policy analysis is not just historical study of observed tendencies. A central objective is to inform policy choice by predicting the outcomes that would occur if past policies were to be continued or alternative ones were to be enacted.

While I am hesitant to second-guess the *OED*, I think it important to observe that its definition of extrapolation is incomplete. The logic of inference does not enable any conclusions about future or hypothetical situations to be drawn based on observed tendencies per se. Assumptions are essential. Thus, I will amend the *OED* definition and say that extrapolation is "the drawing of a conclusion about some future or hypothetical situation based on observed tendencies and maintained assumptions."

Given available data, the credibility of extrapolation depends on what assumptions are maintained. Researchers often use untenable assumptions to extrapolate. I will refer to this manifestation of incredible certitude as *wishful extrapolation*.

Perhaps the most common extrapolation practice is to assume that a future or hypothetical situation

is identical to an observed one in some respect. Analysts regularly make such invariance assumptions, sometimes with good reason but often without basis. Economists drawing broad conclusions from specific findings sometimes say that their extrapolations are "stylized facts."

Certain invariance assumptions achieve the status of conventional certitudes, giving analysts license to pose them without fear that their validity will be questioned. To illustrate, I will discuss extrapolation from randomized experiments, with particular attention to the drug approval process of the Food and Drug Administration (FDA).

6.1. Extrapolation from Randomized Experiments

The appeal of randomized experiments is that they often deliver credible certitude about the outcomes of policies within a population under study. Standard experimental protocol calls for specification of a study population from which random samples of persons are drawn to form treatment groups. All members of a treatment group are assigned the same treatment.

Assume that treatment response is *individualistic*; that is, each person's outcome depends only on his own treatment, not on those received by other members of the study population. Then the distribution of outcomes realized by a treatment group is the same (up to random sampling variation) as would occur if this treatment were assigned to all members of the population. Thus, when the assumption of individualistic treatment response is credible, a randomized experiment enables one to draw credible sharp conclusions about the outcomes that would occur if a policy were to be applied to the entire study population.

A common problem of policy analysis is to extrapolate experimental findings. To accomplish this, analysts regularly assume that the distribution of outcomes that would occur under a policy of interest would be the same as the distribution of outcomes realized by a specific experimental treatment group. This invariance assumption sometimes is reasonable, but it may be wishful extrapolation.

There are many reasons why the invariance assumption may be suspect. I will discuss three here. The use of randomized experiments to inform policy choice has been particularly important in medicine. I will use the drug approval process of the Food and Drug Administration (FDA) to illustrate.

6.1.1. The Study Population and the Population of Interest. The study populations of randomized experiments often differ from the population of policy interest. Participation in experiments cannot be mandated in democracies. Hence, study populations consist of persons who volunteer to participate. Experiments reveal the distribution of treatment response among these volunteers, not within the population to whom a policy would be applied.

Consider the randomized clinical trials (RCTs) performed by pharmaceutical firms to obtain FDA approval to market new drugs. The volunteer participants in these trials may not be representative of the relevant patient population. The volunteers are persons who respond to the financial and medical incentives offered by pharmaceutical firms. Financial incentives may be payment to participate in a trial or receipt of free treatments. The medical incentive is that participation in a trial gives a person a chance of receiving a new drug that is not otherwise available.

The study population materially differs from the relevant patient population if treatment response in the group who volunteer for a trial differs from treatment response among those who do not volunteer. When the FDA uses trial data to make drug approval decisions, it implicitly assumes that treatment response in the patient population is similar to that observed in the trial.

6.1.2. The Experimental Treatments and the Treatments of Interest. The treatments assigned in experiments often differ from those that would be assigned in actual policies. Consider again the RCTs performed for drug approval. These trials are normally double-blinded, neither the patient nor his physician knowing the assigned treatment. Hence, a trial reveals the distribution of response in a setting where patients and

physicians are uncertain what drug a patient receives. It does not reveal what response would be in a real clinical setting where patients and physicians would have this information and be able to react to it.

Another source of difference between the treatments assigned in experiments and those that would be assigned in actual policies occurs when evaluating vaccines for prevention of infectious disease. The assumption of individualistic treatment response traditionally made in analysis of experiments does not hold when considering vaccines, which may not only protect the person vaccinated but also lower the rate at which unvaccinated persons becomes infected. A vaccine is *internally* effective if it generates an immune response that prevents a vaccinated person from become ill or infectious. It is *externally* effective to the extent that it prevents transmission of disease to members of the population who are unvaccinated or unsuccessfully vaccinated.

A standard RCT enables evaluation of internal effectiveness, but does not reveal the external effect of applying different vaccination rates to the population. If the group vaccinated in an experiment is small relative to the size of the population, the vaccination rate is essentially zero. If a trial vaccinates a non-negligible fraction of the population, the findings only reveal the external effectiveness of the chosen vaccination rate. It does not reveal what the population illness rate would be with other vaccination rates.

6.1.3. The Outcomes Measured in Experiments and the Outcomes of Interest. A serious measurement problem occurs when studies have short durations. We often want to learn long-term outcomes of treatments, but short studies reveal only immediate outcomes. Credible extrapolation from such *surrogate outcomes* to the long-term outcomes of interest can be highly challenging.

Again, the RCTs for drug approval provide a good illustration. The most lengthy, called *phase 3 trials*, typically run for only two to three years. When trials are not long enough to observe the health outcomes of real interest, the practice is to measure surrogate outcomes and base drug approval decisions on their values. For example, treatments for heart disease may be evaluated using data on patient cholesterol

levels and blood pressure rather than data on heart attacks and life span. In such cases, which occur regularly, the trials used in drug approval only reveal the distribution of surrogate outcomes in the study population, not the distribution of outcomes of real health interest.

Health researchers have called attention to the difficulty of extrapolating from surrogate outcomes to health outcomes of interest. Fleming and Demets (1996), who review the prevalent use of surrogate outcomes in phase 3 trials evaluating drug treatments for heart disease, cancer, HIV/AIDS, osteoporosis, and other diseases, write (p. 605): "Surrogate end points are rarely, if ever, adequate substitutes for the definitive clinical outcome in phase 3 trials."

6.1.4. The FDA and Conventional Certitude. The FDA drug approval process clearly values credibility, as shown in its insistence on evidence from RCTs and on trial sizes adequate to bound the statistical uncertainty of findings. However, the FDA makes considerable use of conventional certitudes when it attempts to extrapolate from RCT data to predict the effectiveness and safety of new drugs in practice.

The approval process essentially assumes that treatment response in the relevant patient population with be similar to response in the study population. It assumes that response in clinical practice will be similar to response with double-blinded treatment assignment. And it assumes that effectiveness measured by outcomes of interest will be similar to effectiveness measured by surrogate outcomes. These assumptions often are unsubstantiated and sometimes may not be true, but they have become enshrined by long use.

6.2. Campbell and the Primacy of Internal Validity

The FDA is not alone in downplaying the problem of extrapolation from experiments. Elevation of concern with inference in the study population over extrapolation to contexts of policy interest is also characteristic of the social-science research paradigm emerging from the influential work of Donald

Campbell.

Campbell distinguished between the internal and external validity of a study of treatment response. A study is said to have *internal validity* if its findings for the study population are credible. It has *external validity* if an invariance assumption permits credible extrapolation. Campbell discussed both forms of validity, but he argued that studies should be judged primarily by their internal validity and only secondarily by their external validity (Campbell and Stanley, 1963; Campbell, 1984).

This perspective has been used to argue for the primacy of experimental research over observational studies, whatever the study population may be. The reason given is that properly executed randomized experiments have high internal validity. This perspective has also been used to argue that observational studies are most credible when they most closely approximate randomized experiments.

These ideas have noticeably affected governmental decision making. A prominent case is the FDA drug approval process, which only makes use of experimental evidence. Another American example is the Education Sciences Reform Act of 2002 (Public Law 107-279), which provides funds for improvement of federal educational research. The Act defines a scientifically valid educational evaluation to be one that "employs experimental designs using random assignment, when feasible, and other research methodologies that allow for the strongest possible causal inferences when random assignment is not feasible." The term "strongest possible causal inference" has been interpreted to mean the highest possible internal validity. No weight is given to external validity.

Unfortunately, analyses of experimental data have tended to be silent on the problem of extrapolating from the experiments performed to policies of interest. For example, the influential analyses of welfare reform experiments reported in Gueron and Pauly (1991) only described the mean outcomes experienced by the various treatment groups. One can use the reported experimental findings to predict policy outcomes only if one is willing to take the findings at face value, accepting their internal validity and not questioning their external validity. One is at a loss to interpret the findings otherwise.

From the perspective of policy choice, it makes no sense to value one type of validity above the other. What matters is the informativeness of a study for policy making, which depends jointly on internal and external validity. Hence, research should strive to measure both types of validity. Whether the available data are experimental or observational, the result of credible policy analysis will typically be interval rather than point predictions of policy outcomes. I earlier described an illustrative case study of observational data performed in this manner, the Manski and Nagin (1998) study of sentencing and recidivism. Illustrative studies using experimental data to generate interval predictions of policy outcomes include Manski (1997) and Pepper (2003).

7. Illogical Certitude

I have thus far discussed research practices that are not credible but are logical. Deductive errors, particularly non sequiturs, also contribute to incredible certitude. A common non sequitur occurs when a researcher performs a statistical test of some null hypothesis, finds that the hypothesis is not rejected, and interprets non-rejection as proof that the hypothesis is correct. Texts on statistics routinely caution that non-rejection does not prove a null hypothesis is correct. It only indicates the absence of strong evidence that the hypothesis is incorrect. Nevertheless, researchers sometimes confuse statistical non-rejection with proof.

A more exotic non sequitur has persisted in research on the heritability of human traits, which has often been wrongly interpreted to have implications for social policy. I will use this as an extended case study.

7.1. Heritability

Heritability has been a persistent topic of study and controversy since the latter third of the 19th century. The beginning of formal research is usually attributed to the British scientist Francis Galton, who appears to have been the first to attempt to distinguish the roles of "nature" and "nurture." About one hundred years after Galton started his studies, controversy about the heritability of IQ flared in the 1960s and 1970s. This subject has been particularly heated because some social scientists have sought to connect heritability of IQ with social policy, asserting that policy can do little to ameliorate inequality of achievement if IQ is largely heritable.

Considering the state of thinking in the late 1970s, Goldberger (1979) began a cogent critique of research on heritability this way: (p. 327):

> When we look across a national population, we see large differences in intelligence as measured by IQ tests. To what extent are those differences the result of differences in genetic make-up, and to what extent are they the result of differences in life experience? What proportion of the variance in IQ test scores is attributable to genetic variance, and what proportion to environmental variance? This question has fascinated mankind—or at least the Anglo-American academic sub-species—for several generations. The fascination, I suppose, arises from the notion that the answer has some relevance to social policy: if IQ variance is largely genetic, then it is natural, just and immutable; but if IQ variance is largely environmental, then it is unnatural, unjust and easily eradicated.

Goldberger concluded that heritability, whether it be of IQ or other traits, is irrelevant to social policy. I will explain why here. However, I first need to explain what the heritability statistic measures and how it has been interpreted.

Lay people often use the word "heritability" in the loose sense of the *Oxford English Dictionary*, which defines it as "The quality of being heritable, or capable of being inherited." However, formal research

on heritability uses the word in a specific technical way. Stripped to its essentials, heritability research seeks to perform an analysis of variance.

Consider a population of persons. Researchers pose an equation of the form

$$\text{outcome} = \text{genetic factors} + \text{environmental factors}$$

or, more succinctly, $y = g + e$. Here, y is a personal outcome (or phenotype), g symbolizes genetic factors, and e symbolizes environmental factors. It is commonly assumed that g and e are uncorrelated across the population. Then the ratio of the population variance of g to the variance of y is called the heritability of y. Researchers say that heritability gives the fraction of the variation in the outcome "explained by" or "due to" genetic factors.

If y, g, and e were observable variables, this would be all there is to the methodology of heritability research. However, only outcomes are observable. The quantities g and e are not observable summary statistics for a person's genome and the environment. They are unobservable metaphors. The somewhat mystifying technical intricacies of heritability research—its reliance on outcome data for biological relatives, usually twins, and on various strong assumptions—derives from the desire of researchers to make heritability estimable despite the fact that g and e are metaphorical.

Suppose that a researcher obtains data on the outcomes experienced by twins or other relatives, makes enough assumptions, and reports an estimate of the heritability of the outcome. What does this number tell us? Researchers often say that heritability measures the relative importance of genetic and environmental factors. Loose use of the word "importance" is unfortunately common in empirical social science research. A prominent example is *The Bell Curve*, where Herrnstein and Murray (1994, Chap. 5) proclaimed (p. 135): "Cognitive ability is more important than parental SES in determining poverty." Goldberger and Manski (1995) critique the analysis that underlies this and similar assertions.

7.1.1. Heritability and Social Policy. What has made research on heritability particularly controversial has been the inclination of researchers such as Herrnstein and Murray to interpret the magnitude of heritability estimates as indicators of the potential responsiveness of personal achievement to social policy. Large estimates of heritability have been interpreted as implying small potential policy effectiveness.

A notable example was given by Goldberger (1979). Discussing a *London Times* report of research relating genetics to earnings and drawing implications for social policy, he wrote (p. 337):

> For a more recent source we turn to the front page of The Times (13 May 1977), where under the heading "Twins show heredity link with earnings" the social policy correspondent Neville Hodgkinson reported:
>
> > A study of more than two thousand pairs of twins indicates that genetic factors play a huge role in determining an individual's earning capacity . . . . According to some British researchers, the study provides the best evidence to date in the protracted debate over the respective contributions of genetics and environment to an individual's fate . . . . The findings are significant for matters of social policy because of the implication that attempts to make society more equal by breaking "cycles of disadvantage" . . . . are likely to have much less effect than has commonly been supposed.
>
> Professor Hans Eysenck was so moved by the twin study that he immediately announced to Hodgkinson that it "really tells the [Royal] Commission [on the Distribution of Income and Wealth] that they might as well pack up" (The Times, 13 May 1977).

Commenting on Eysenck, Goldberger continued (p. 337):

> (A powerful intellect was at work. In the same vein, if it were shown that a large proportion of the variance in eyesight were due to genetic causes, then the Royal Commission on the Distribution of Eyeglasses might as well pack up. And if it were shown that most of the variation in rainfall is due to natural causes, then the Royal Commission on the Distribution of Umbrellas could pack up too.)

This parenthetical passage, displaying Goldberger's characteristic combination of utter seriousness and devastating wit, shows the absurdity of considering heritability estimates to be policy relevant. Goldberger concluded (p. 346): "On this assessment, heritability estimates serve no worthwhile purpose."

It is important to understand that Goldberger's conclusion did not rest on the metaphorical nature of g and e in heritability research. It was based, more fundamentally, on the fact that variance decompositions do not yield estimands of policy relevance.

To place heritability research on the best imaginable footing, suppose that g and e are not metaphors but rather are observable summary statistics for a person's genome and environment. Suppose that the equation $y = g + e$ is a physical law showing how the genome and environment combine to determine outcomes. Also suppose that g and e are uncorrelated in the population, as is typically assumed in heritability research. Then a researcher who observes the population may directly compute the heritability of y, without the need for special data on twins or obscure assumptions.

At one extreme, suppose that the population is composed entirely of clones who face diverse environments. Then the variance of g is zero, implying that heritability is zero. At the other extreme, suppose that the population is composed of genetically diverse persons who share the same environment. Then the variance of e is zero, implying that heritability is one.

What does this have to do with policy analysis? Nothing. Policy analysis asks what would happen to outcomes if an intervention, such as distribution of eyeglasses, were to change persons' environments in some manner. Heritability is uninformative about this.

Due to the work of Goldberger and others, it was recognized more than thirty years ago that heritability research is irrelevant to policy.[8] Nevertheless, some have continued to assert its relevance. For

[8] While Goldberger got to the heart of the logical problem with heritability research in a particularly succinct and effective way, he was not alone in grasping the irrelevance of heritability to policy. Writing contemporaneously, the statistician Oscar Kempthorne summarized his view of the matter this way

example, Herrnstein and Murray did so in *The Bell Curve*, referring to (p. 109): "the limits that heritability puts on the ability to manipulate intelligence." Research on the heritability of all sorts of outcomes continues to appear regularly today. Recent studies tend not to explicitly refer to policy, but neither do they provide any other articulate interpretation of the heritability statistics they report. The work goes on, but I do not know why.

8. Media Overreach

Elected officials, civil servants, and the public rarely learn of policy analysis from the original sources. The writing in journal articles and research reports is usually too technical and jargon-laden for non-professionals to decipher. Broad audiences may learn of new findings from newspapers, magazines, and electronic media. The journalists and editors who decide what analysis warrants coverage and how to report it therefore have considerable power to influence societal perspectives.

Some media coverage of policy analysis is serious and informative, but overreach is common. Journalists and editors seem to rarely err on the side of overly cautious reporting. The prevailing view seems to be that certitude sells.

(Kempthorne, 1978, p. 1):

> The conclusion is that the heredity-IQ controversy has been a "tale full of sound and fury, signifying nothing." To suppose that one can establish effects of an intervention process when it does not occur in the data is plainly ludicrous.

8.1. "The Case for $320,000 Kindergarten Teachers"

A conspicuous instance of media overreach appeared on the front page of the *New York Times* on July 28, 2010, in an article with the above title. There the *Times* economics columnist David Leonhardt reported on research investigating how students' kindergarten experiences affect their income as adults. Leonhardt began his article with the question "How much do your kindergarten teacher and classmates affect the rest of your life?" He then called attention to new work by a group of six economists that attempts to answer the question, at least with regard to adult income.

Characterizing the study's findings as "fairly explosive," Leonhardt focussed most attention on the impact of good teaching. Referring by name to Raj Chetty, one of the authors, he wrote

> Mr. Chetty and his colleagues . . . . estimate that a standout kindergarten teacher is worth about $320,000 a year. That's the present value of the additional money that a full class of students can expect to earn over their careers.

Leonhardt concluded by making a policy recommendation, stating:

> Obviously, great kindergarten teachers are not going to start making $320,000 anytime soon. Still, school administrators can do more than they're doing. They can pay their best teachers more . . . . and give them the support they deserve. . . . Given today's budget pressures, finding the money for any new programs will be difficult. But that's all the more reason to focus our scarce resources on investments whose benefits won't simply fade away.

I have called Leonhardt's article media overreach. My reason was hinted at by Leonhardt when he wrote that the new study was "not yet peer-reviewed." In fact, the study did not even exist as a publicly available working paper when Leonhardt wrote his article. All that existed for public distribution was a set of slides dated July 2010 for a conference presentation made by the authors. A bullet point on the final page of the slides estimates the value of good kindergarten teaching to be $320,000. The slides do not provide

sufficient information about the study's data and assumptions to enable an observer to assess the credibility of this estimate.[9]

When Leonhardt wrote his article, the community of researchers in the economics of education had not yet had the opportunity to read or react to the new study, never mind to review it for publication. Nevertheless, Leonhardt touted the findings as definitive and used them to recommend policy. Surely this is incredible certitude. I think it highly premature for a major national newspaper to report at all on new research at such an early stage, and bizarre to place the report on the front page.

8.2. Peer Review and Credible Reporting

The 2010 New York Times article on kindergarten teaching is a striking case of reporting on research prior to peer review, but it is not unique. For example, the 1977 London Times article on heritability cited in Section 7 reported the findings of an unpublished draft research paper.

Premature media reporting on research would lessen to some degree if the media would refrain from covering research that has not yet been vetted within the scientific community through an established peer-review process. However, journalists should not trust peer review per se to certify the logic or credibility of research. Anyone with experience submitting or reviewing articles for publication becomes aware that peer review is an imperfect human enterprise. Weak studies may be accepted for publication and strong studies rejected, even when peer reviewers do their best to evaluate research objectively. The trustworthiness of peer review is diminished further when reviewers use the process to push their own advocacy agendas, accepting studies whose conclusions they favour.

---

[9] The work was presented on July 27, 2010 at the Summer Institute of the National Bureau of Economic Research. The slides were available on Raj Chetty's web page at http://obs.rc.fas.harvard.edu/chetty/STAR_slides.pdf. Accessed August 3, 2010.

It is unquestionably difficult for journalists and editors, who cannot possibly be sufficiently expert to evaluate personally all policy analysis, to decide what studies to report and how to frame their coverage. Yet there are straightforward actions that they can take to mitigate media overreach. First and perhaps foremost, they can scrutinize research reports to assess whether and how the authors express uncertainty about their findings. They should be deeply sceptical of studies that assert certitude. When authors express uncertainty, journalists should pay close attention to what they say.

Second, journalists should not rely fully on what authors say about their own work. They should seek perspectives from relevant reputable researchers who are not closely associated with the authors. Careful journalists already do this, but the practice should become standard.

9. Credible Policy Analysis

This paper has developed a typology of analytical practices that contribute to incredible certitude. The phenomena discussed here are common attributes of policy studies. I have presented illustrative cases that I think to be instructive. Readers may have their own favourite illustrations to offer. Readers may also wish to refine or add to the typology of practices.

I have asserted that incredible certitude is harmful to policy choice, but it is not enough to criticize. I must suggest a constructive alternative. I wrote in Section 2 that an analyst can resolve the tension between the credibility and power of assumptions by posing alternative assumptions of varying credibility and determining the conclusions that follow in each case. I gave an example in Section 4, when I discussed the Manski and Nagin (1998) study of sentencing and recidivism. To reiterate, we implemented a "layered" analysis. This began with no assumptions about how judges sentence offenders and then moved from weak, highly credible assumptions to stronger, less credible ones. We presented several sets of findings, in the form

of interval predictions of policy outcomes. These findings showed how conclusions about sentencing policy vary depending on the assumptions made.

A researcher who performs an instructive layered policy analysis and exposits the work clearly may see himself as having accomplished the objective of informing policy choice. There remains the question of how policy makers may use the information provided. When the policy maker is a planner with well-defined beliefs and social welfare function, decision theory provides an appropriate framework for credible policy choice. Decision theory does not offer a consensus prescription for policy choice with partial knowledge, but it is unified in supposing that choice should reflect the knowledge that the decision maker actually has. It does not prescribe incredible certitude.

How should a planner with partial knowledge act? Decision theory gives a simple partial answer, but no complete answer.

The partial answer is that a planner should not choose a *dominated* policy. Contemplating some policy D, a planner might find that there exists another feasible policy, say C, that yields higher welfare than D in every scenario the planner thinks feasible. Then policy D is said to be dominated by C.

Decision theory prescribes that one should not choose a dominated policy. This is common sense. Uncertainty is inconsequential when evaluating a dominated policy. Although a planner may not be able to predict exact outcomes, he surely can do better than choose a dominated policy. Manski (2006, 2010) present illustrative applications, the former to police search policy and the latter to vaccination policy.

The hard part of planning with partial knowledge is choice among undominated policies. There is no one right way to make this choice. Consequently, decision theory cannot provide a consensus prescription. Instead, it suggests a variety of approaches that might be deemed reasonable.

Economists are most familiar with the Bayesian branch of decision theory, which supposes that beliefs are probabilistic and applies the subjective expected utility criterion. Some policy-choice applications are Meltzer (2001) to medical decision making, Dehejia (2005) to anti-poverty programs, and Nordhaus

(2008) to climate policy.

Another branch is the theory of decision making under ambiguity, which does not presume probabilistic beliefs. Prominent suggestions for decision making under ambiguity include the maximin and the minimax-regret criteria. Manski (2006, 2010) apply both criteria to search and vaccination policy. Hansen and Sargent (2007) and Barlevy (2011) apply the maximin criterion to macroeconomic policy.

There remains an open question of what constitutes effective analysis when policy making is not adequately approximated by decision theory. The psychological-cognitive argument for certitude cited in Section 2 views policy makers as so boundedly rational that incredible certitude is more useful than credible policy analysis. I do not find this conclusion credible, but I have to acknowledge that it is not refutable with available data.

A different question concerns the nature of effective policy analysis in political settings, where agents with differing beliefs and objectives jointly make policy choices. I observed in Section 3 that the study of political decision making requires viewing policy choice as a game rather than as a problem of individual decision making. That said, it is not clear what guidance game theory can credibly provide.

References

Auerbach, A. (1996). 'Dynamic Revenue Estimation.' *Journal of Economic Perspectives*, vol. 10, no. 1, (Winter), pp. 141-157.

Barlevy G. (2011). 'Robustness and Macroeconomic Policy.' *Annual Review of Economics*, vol. 3, forthcoming.

Britton, E., Fisher, P. and Whitley, J. (1998). 'The *Inflation Report* Projections: Understanding the Fan Chart.' *Bank of England Quarterly Bulletin*, (February), pp. 30-37.

Campbell, D. (1984). 'Can We Be Scientific in Applied Social Science?' *Evaluation Studies Review Annual*, vol. 9, pp. 26–48.

Campbell, D. and Stanley, J. (1963). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Committee on the Budget, United States House of Representatives (2008), *Compilation of Laws and Rules Relating to the Congressional Budget Process*, Serial No. CP-3, Washington, DC: United States Government Printing Office.

Crane, B, Rivolo, A. and Comfort, G. (1997). *An Empirical Examination of Counterdrug Interdiction Program Effectiveness*. IDA paper P-3219, Alexandria, VA: Institute for Defense Analyses.

Dehejia, R. (2005). 'Program Evaluation as a Decision Problem.' *Journal of Econometrics*, vol. 125, nos. 1-2, (March-April), pp. 141-173.

Delavande, A., Giné, X. and McKenzie, D. (2011). 'Measuring Subjective Expectations in Developing Countries: A Critical Review and New Evidence.' *Journal of Development Economics*, vol. 94, no. 2, (March), pp. 151-163.

Elmendorf, D. (2010a). letter to Honorable Nancy Pelosi, Speaker, United States House of Representatives. Congressional Budget Office, March 18. http://www.cbo.gov/ftpdocs/113xx/doc11355/hr4872.pdf.

Elmendorf, D. (2010b). letter to Honorable Paul Ryan, United States House of Representatives. Congressional Budget Office, March 19. http://www.cbo.gov/ftpdocs/113xx/doc11376/RyanLtrhr4872.pdf

Fleming, T. and Demets, D. (1996). 'Surrogate End Points in Clinical Trials: Are We Being Misled?' *Annals of Internal Medicine.* Vol. 125, no. 7, pp. 605-613.

Foster, R. (2010). *Estimated Financial Effects of the 'Patient Protection and Affordable Care Act.' as Amended*, Office of the Actuary, Centers for Medicare & Medicaid Services, United States Department of Health and Human Services, April 22. https://www.cms.gov/ActuarialStudies/Downloads/PPACA_2010-04-22.pdf

Friedman, M. (1953). *Essays in Positive Economics*. Chicago: University of Chicago Press.

Friedman, M.(1955). 'The Role of Government in Education.' in R. Solo (editor), *Economics and the Public*

Interest, New Brunswick: Rutgers University Press.

Friedman, M. (1962). *Capitalism and Freedom*. Chicago: University of Chicago Press.

Galbraith, J. (1958). *The Affluent Society*. New York: Mentor Book.

Goldberger, A. (1979). 'Heritability.' *Economica*, vol. 46, no. 184, pp. 327-347.

Goldberger, A. and Manski, C. (1995). 'Review Article: *The Bell Curve* by Herrnstein and Murray.' *Journal of Economic Literature,* vol. 33, no. 2, pp. 762-776.

Gueron, J. and Pauly, E. (1991). *From Welfare to Work*. New York: Russell Sage Foundation.

Hansen, L. and Sargent, T. (2007). *Robustness*. Princeton: Princeton University Press.

Herrnstein, R. and Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life*. New York: The Free Press.

Herszenhorn, D. (2010). 'Fine-Tuning Led to Health Bill's $940 Billion Price Tag.' *The New York Times*, March 18.

Holtz-Eakin, D. (2010). 'The Real Arithmetic of Health Care Reform.' *The New York Times*, March 21.

Hurd, M. (2009). 'Subjective Probabilities in Household Surveys.' *Annual Review of Economics*, vol. 1, pp.

543-564.

Kempthorne, O. (1978). 'Logical, Epistemological, and Statistical Aspects of Nature-Nurture Data Interpretation.' *Biometrics*, vol. 34, no.1, (March), pp. 1-23.

Krugman, P. (2007). 'Who Was Milton Friedman?' *New York Review of Books*, February 15.

Manski C. (1990). 'Nonparametric Bounds on Treatment Effects.' *American Economic Review Papers and Proceedings*. Vol. 80, no. 3, (May), pp. 319-323.

Manski, C. (1992). 'School *Choice* (Vouchers) and Social Mobility.' *Economics of Education Review*, vol. 11, no. 4, pp. 351-369.

Manski C. (1995). *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Manski, C. (1997). 'The Mixing Problem in Programme Evaluation.' *Review of Economic Studies*, vol. 64, no. 4, pp. 537–553.

Manski, C. (2003). *Partial Identification of Probability Distributions*. New York: Springer-Verlag.

Manski, C. (2004). 'Measuring Expectations.' *Econometrica*, vol. 72, no. 5, pp. 1329-1376.

Manski C. (2006). 'Search Profiling with Partial Knowledge of Deterrence.' ECONOMIC JOURNAL, vol. 116, no. 515, pp. F385-F401

Manski C. (2007). *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press.

Manski C. (2009). 'Diversified Treatment under Ambiguity.' *International Economic Review*, vol. 50, no. 4, pp. 1013-1041.

Manski C. (2010). 'Vaccination with Partial Knowledge of External Effectiveness.' *Proceedings of the National Academy of Sciences*. Vol. 107, no. 9, pp. 3953-3960.

Manski, C. (2011). 'Choosing Treatment Policies under Ambiguity.' *Annual Review of Economics*, vol. 3, forthcoming.

Manski, C. and D. Nagin (1998). 'Bounding Disagreements About Treatment Effects: A Case Study of Sentencing and Recidivism.' *Sociological Methodology*, vol. 28, pp. 99-137.

Meltzer D. (2001). 'Addressing Uncertainty in Medical Cost-Effectiveness: Implications of Expected Utility Maximization for Methods to Perform Sensitivity Analysis and the Use of Cost-Effectiveness Analysis to Set Priorities for Medical Research.' *Journal of Health Economics*, vol. 20, no. 1, (January), pp. 109-129

National Research Council (1999). *Assessment of Two Cost-Effectiveness Studies on Cocaine Control Policy*. Committee on Data and Research for Policy on Illegal Drugs, Charles F. Manski, John V. Pepper, and Yonette Thomas, editors. Committee on Law and Justice and Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, Washington, DC: National Academy Press.

National Research Council (2001). *Informing America's Policy on Illegal Drugs: What We Don't Know*

Keeps Hurting Us. Committee on Data and Research for Policy on Illegal Drugs, Charles F. Manski, John V. Pepper, and Carol V. Petrie, editors. Committee on Law and Justice and Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, Washington, DC: National Academy Press.

Nordhaus, W. (2008). *A Question of Balance: Weighing the Options on Global Warming Policy*. New Haven, CT: Yale University Press.

Pepper, J. (2003). 'Using Experiments to Evaluate Performance Standards: What Do Welfare-to-Work Demonstrations Reveal to Welfare Reformers?' *Journal of Human Resources*, Vol. 38, no. 4, pp. 860–880.

Page, R. (2005). 'CBO's Analysis of the Macroeconomic Effects of the President's Budget.' *American Economic Review Papers and Proceedings*, vol. 95, no. 3, (May), pp. 437-440.

Rydell, C. and S. Everingham (1994). *Controlling Cocaine*. Report prepared for the Office of National Drug Control Policy and the United States Army, Santa Monica, CA: RAND Corporation.

Swinburne, R. (1997). *Simplicity as Evidence for Truth*. Milwaukee: Marquette University Press.

Subcommittee on National Security, International Affairs, and Criminal Justice (1996). *Hearing Before the Committee on Governmental Reform and Oversight*. United States House of Representatives, Washington, DC: U.S. Government Printing Office.

Subcommittee on National Security, International Affairs, and Criminal Justice (1998). *Hearing Before the*

Committee on Governmental Reform and Oversight. United States House of Representatives, Washington, DC: U.S. Government Printing Office.