POLICY ANALYSIS WITH INCREDIBLE CERTITUDE

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

and

Leverhulme Visiting Professor, UCL

The Leverhulme Trust

POLICY ANALYSIS
WITH INCREDIBLE CERTITUDE

Charles F. Manski

Department of Economics & Institute for Policy Research
Northwestern University

and

Leverhulme Visiting Professor, UCL

Analyses of public policy regularly express certitude about the consequences of alternative policy choices.

Expressions of uncertainty are rare.

Yet, 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.

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.

I have shown how simple ideas in decision theory may be used to make reasonable policy choices.

See

Manski, C., *Identification for Prediction and Decision*, Harvard University Press, 2007.

Manski, C. "Choosing Treatment Policies under Ambiguity," *Annual Review of Economics*, Vol. 3, 2011.

I have warned against specific analytical practices that promote incredible certitude.

I have not previously sought to classify these practices and consider them in totality. I do so here.

A revised version of this paper will become the first chapter of a new book: *Public Policy in an Uncertain World*.

Typology of Incredible Practices

conventional certitudes

dueling certitudes

conflating science and advocacy

wishful extrapolation

*illogical certitude*s

media overreach

The Logic and Credibility of Empirical 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.

Non sequiturs yield misplaced certitude.

A fundamental difficulty in empirical research is to decide what assumptions to maintain.

With given data and no deductive errors, stronger assumptions yield stronger conclusions.

There is a tension between the strength of assumptions and their credibility. I have called this (Manski, 2003):

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.

Credibility is a subjective matter. Analysts should agree on the logic of inference, but they may disagree about the credibility of assumptions.

Disagreements occur often.

Disagreements can persist without resolution when assumptions are *nonrefutable*; that is, when multiple contradictory assumptions are all consistent with the available data.

An analyst can pose a nonrefutable assumption and adhere to it forever. He can displace the burden of proof, stating

"I will maintain this assumption until it is proved wrong."

Incentives for Certitude

A researcher can resolve the tension between the credibility and power of assumptions by posing assumptions of varying credibility and determining the conclusions that follow.

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, 1995, 2007):

“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, D.C. 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.’ “

A longtime econometrics colleague who frequently acts as a consultant stated the incentive argument this way:

"You can't give the client a bound. The client needs a point."

This comment reflects a common perception that policy makers are either psychologically unwilling or cognitively unable to cope with ambiguity.

Consultants often argue that pragmatism dictates provision of point predictions, even though these predictions may not be credible.

Support for Certitude in Philosophy of Science

The view that analysts should offer sharp predictions is not confined to presidents and consultants. It has a long history in the philosophy of science.

Milton Friedman (1953) expressed this perspective. He placed prediction as the central objective of science, writing

"The ultimate goal of a positive science is the development of a 'theory' or 'hypothesis' that yields valid and meaningful . . . predictions about phenomena not yet observed."

He went on to say:

“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.”

Friedman 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.

Conventional Certitudes

John Kenneth Galbraith popularized the term *conventional wisdom*. Wikipedia put it this way:

"Conventional wisdom is a term used to describe ideas or explanations that are generally accepted as true by the public or by experts in a field. . . . . Conventional wisdom is not necessarily true."

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

CBO Scoring of Legislation

Conventional certitude is exemplified by Congressional Budget Office (CBO) *scoring* of federal legislation.

The CBO was established in the Congressional Budget Act of 1974. The Act has been interpreted as mandating the CBO to provide point predictions (or scores) of the budgetary impact of legislation.

CBO scores are conveyed in letters that the Director writes to leaders of Congress.

They are not accompanied by measures of uncertainty, even though legislation may propose complex changes to law, whose budgetary implications may be difficult to foresee.

It is remarkable that CBO scores have achieved broad acceptance within American society.

The scores of pending legislation are used by both Democratic and Republican Members of Congress.

Media reports largely take them at face value.

The Patient Protection and Affordable Care Act of 2010

In March 2010 the CBO scored the combined consequences of the Patient Protection and Affordable Care Act and the Reconciliation Act of 2010. Director Douglas Elmendorf wrote to Nancy Pelosi:

"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."

The letter expressed no uncertainty and did not document the methodology generating the prediction.

Media reports largely accepted the CBO scores as fact.

A *New York Times* article reported:

"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.

The certitude that CBO expressed when predicting budgetary impacts ten years into the future gave way to considerable uncertainty when considering longer horizons.

Elmendorf wrote to Pelosi:

"CBO has developed a rough outlook for the decade following the 2010-2019 period. . . . Our analysis indicates that H.R. 3590 . . . . . 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."

Thus, the CBO acknowledged uncertainty when asked to predict 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?

It is not reasonable to set a discontinuity at ten years, with certitude expressed up to that point and uncertainty only beyond it.

I expect the CBO knew the ten-year prediction was only a rough estimate.

However, it felt compelled to express certitude when providing ten-year predictions, which play a formal role in the Congressional budget process.

Interval Scoring

The CBO has established an admirable reputation for impartiality.

One may argue that it is best to leave well enough alone and have the CBO express certitude when it scores legislation, even if the certitude is conventional rather than credible.

I worry that the existing social contract to take CBO scores at face value will eventually break down.

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.

A simple approach would be to provide interval forecasts of the budgetary impacts of legislation.

The CBO would 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 forecast.

Dueling Certitudes

A rare commentator who rejected the CBO's score for the health care legislation was Douglas Holtz-Eakin, a former Director of the CBO. He wrote

"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 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 questioned the assumptions maintained by the CBO, and he asserted that a different result emerges under alternative assumptions that he preferred.

Each score makes sense in its own terms, each combining available data with assumptions to draw logically valid conclusions. Yet the two scores are sharply contradictory.

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 at RAND and the other at the Institute for Defense Analyses (IDA).

The two studies posed the same objective—reduction in cocaine consumption by one percent. Both predicted the cost of using certain policies to achieve this objective.

RAND and IDA used different assumptions and data to reach dramatically different conclusions.

The RAND study specified a model of the supply and demand for cocaine. It used the model to evaluate alternative policies and reached this conclusion:

"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. The least costly supply-control program (domestic enforcement) costs 7.3 times as much as treatment to achieve the same consumption reduction."

The IDA study examined the time-series association between source-zone interdiction activities and retail cocaine prices. It reached this conclusion:

"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 RAND study was used to argue that funding should be shifted towards drug treatment and away from activities to reduce drug production or to interdict drug shipments.

The IDA study was used to argue that interdiction activities should be funded at present levels or higher.

The National Research Council Assessment

The National Research Council Committee on Data and Research for Policy on Illegal Drugs assessed the RAND and IDA studies.

The Committee concluded that neither study provides a credible estimate of the cost of using alternative policies to reduce cocaine consumption.

I consider the many differences between the RAND and IDA studies 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.

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

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.

Conflating Science and Advocacy

Recall: assumptions + data ⇒ conclusions.

The directionality ordinarily runs from left to right. One poses assumptions and derives conclusions.

One can reverse directionality, seeking assumptions that imply predetermined conclusions. This characterizes advocacy.

Some policy analysis is advocacy wrapped in the rhetoric of science.

Friedman and Educational Vouchers

Proponents of vouchers have argued that American school finance policy limits the available educational options and impedes the development of superior alternatives.

Government operation of free public schools, they say, should be replaced by vouchers permitting students to choose any school meeting specified standards.

The awakening of modern interest is usually credited to Friedman (1955).

He cited no empirical evidence relating school finance to educational outcomes. He posed a purely theoretical classical economic argument for vouchers, writing

“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 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.”

“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 argued that the “three major grounds on which government intervention is to be justified” justify government supply of educational vouchers but not government operation of free public schools.

Repeatedly, he entertained a ground for government operation of schools and then dismissed it.

Here is an excerpt from his discussion of the neighborhood-effects argument:

“One argument from the “neighborhood 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.”

Friedman cited no empirical evidence regarding neighborhood effects, nor did he call for research. He simply stated “it is by no means clear” and “it is not at all clear” that neighborhood 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 reverse the burden of proof, arguing that the existing system should be retained in the absence of evidence. The result would be dueling certitudes.

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 depends on the magnitudes of the market imperfections and neighborhood effects that Friedman noted as possible justifications for government intervention.

And it would have to observe that information about these matters was almost entirely lacking when Friedman wrote in the mid-1950s.

Wishful Extrapolation

The *Oxford English Dictionary* defines *extrapolation* as

"the drawing of a conclusion about some future or hypothetical situation based on observed tendencies."

Extrapolation is essential to the use of data in policy analysis.

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.

The *OED* 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**."

Researchers often use untenable assumptions to extrapolate. This manifestation of incredible certitude is *wishful extrapolation*.

Extrapolation from Randomized Experiments

Randomized experiments may yield credible certitude about treatment response in a study population. (*Internal validity*)

A common problem is to extrapolate findings. *(External validity*)

Analysts often assume that the outcomes that would occur under a policy of interest are the same as the outcomes that actually occur in a treatment group.

This *invariance* assumption may be reasonable or may be wishful extrapolation.

The FDA Drug Approval Process

Consider the randomized clinical trials (RCTs) performed to obtain FDA approval to market new drugs.

The Study Population and the Population of Interest

The study population is composed of volunteers, who may not be representative of the relevant patient population.

When the FDA uses trial data to approve drugs, it implicitly assumes that treatment response in the patient population is similar to that observed in the trial. This invariance assumption may or may not be accurate.

Experimental Treatments and Treatments of Interest

The drug treatments assigned in RCTs differ from those that would be assigned in practice.

Drug trials are double-blinded, neither the patient nor physician knowing the assigned treatment.

A trial reveals response when 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.

Measured Outcomes and Outcomes of Interest

We often want to learn long-term outcomes of treatments. RCTs often have short durations. Credible extrapolation from measured *surrogate outcomes* to outcomes of interest can be challenging.

The most lengthy RCTs for drug approval, called *phase 3 trials*, typically run two to three years. A standard practice is to measure surrogate outcomes and base drug approval decisions on their values.

Illogical Certitude

Deductive errors, particularly non sequiturs, contribute to incredible certitude.

A common non sequitur occurs when a researcher performs a statistical test of a null hypothesis, finds that the hypothesis is not rejected, and interprets non-rejection as proof that the hypothesis is correct.

Texts on statistics caution that non-rejection does not prove a null hypothesis is correct. Nevertheless, researchers sometimes confuse statistical non-rejection with proof.

Heritability

An exotic non sequitur has persisted in research on the heritability of human traits.

Heritability has been a topic of study and controversy since the latter third of the 19th century.

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.

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.”

Formal research on heritability uses the word in a specific technical way. Heritability research seeks to perform an analysis of variance.

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

outcome = genetic factors + environmental factors

or y = g + e. Here, y is a personal outcome, 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 variance of g to the variance of y is called the heritability of y.

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. g and e are unobservable metaphors.

The 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.

Heritability and 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:

“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 passage shows the absurdity of considering heritability estimates to be policy relevant.

Goldberger concluded:

"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.

Media Overreach

The public rarely learn of policy analysis from the original sources. They learn of new findings from the media.

The journalists and editors who decide what analysis warrants coverage and how to report it have considerable power to influence societal perspectives.

Some media coverage of policy analysis is serious and informative, but overreach is common.

The prevailing view seems to be that certitude sells.

“The Case for $320,000 Kindergarten Teachers”

A conspicuous instance appeared on the front page of the *New York Times* on July 28, 2010. The *Times* economics columnist David Leonhardt reported on research investigating how students’ kindergarten experiences affect their income as adults. He began with the question

“How much do your kindergarten teacher and classmates affect the rest of your life?”

He then called attention to new work that attempts to answer the question, at least with regard to adult income.

Leonhardt focused on the impact of good teaching. Referring 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."

He concluded with 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.

This is media overreach. Leonhardt 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.

What existed 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.

The research community has 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.

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.

Credible Policy Analysis

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 stated earlier 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.

My research provides illustrative case studies of this approach.

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.

Yet it is unified in supposing that choice should reflect the beliefs that the decision maker actually holds, not incredible certitude.

Economists are most familiar with the Bayesian branch of decision theory, which supposes that beliefs are probabilistic and applies the expected utility criterion.

Another branch is the theory of decision making under ambiguity, which does not presume probabilistic beliefs and may apply the maximin or minimax-regret criterion.

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 views policy makers as so boundedly rational that incredible certitude is more useful than credible policy analysis.

A different question concerns the nature of effective policy analysis in political settings, where multiple agents with differing beliefs and objectives jointly make policy choices.