What Went Wrong? Reflections on Science by Observation and *The Bell Curve**

Clark Glymour†

Departments of Philosophy, University of California-San Diego, Carnegie Mellon University

The Bell Curve aims to establish a set of causal claims. I argue that the methodology of *The Bell Curve* is typical of much of contemporary social science and is intrinsically defective. I claim better methods are available for causal inference from observational data, but that those methods would yield no causal conclusions from the data used in the formal analyses in *The Bell Curve*. Against the laissez-faire social policies advocated in the book, I claim that when combined with common sense and other information, the informal data mustered in *The Bell Curve* support a range of "liberal" social policies.

1. The Thesis of this Essay. Philosophical skepticism trades on two maneuvers: a focus on the worst case, and a demand that any method of forming belief find the truth in all logically possible circumstances. When action must be taken, skepticism is in league with obscurantism, with know-nothingism, and in opposition to forces that are more optimistic about the information that inquiry can provide to judgment. In this century, the principal tool of scientific optimism—although not always of social optimism—has been social statistics. Social statistics promised something less than a method of inquiry that is reliable in every possible circumstance, but something more than sheer ignorance; it promised methods that, under explicit and often plausible assumptions, but not in every logically possible circumstance, converge to the truth, whatever that may be, methods whose liability to error in the short run can be quantified and measured.

*Received November 1995; revised May 1997.

†Send reprint requests to the author, Department of Philosophy, University of California-San Diego, La Jolla, CA 92093

Philosophy of Science, 65 (March 1998) pp. 1–32. 0031-8248/98/6501-0001$2.00

That promise was kept for two statistical enterprises, hypothesis testing and parameter estimation, which for decades were the cynosure of professional statistical study, but it failed in the more important parts of social inquiry that decide which parameters to estimate and which hypotheses to test. To make those decisions with the same guarantee of conditional reliability requires methods of search and requires theoretical inquiries into the reliabilities of those methods. Social statistics produced and used a variety of procedures—factor analysis and regression are the principal examples—for searching for appropriate hypotheses, but no analysis of the conditions for their reliability. The reasons their reliabilities were insufficiently analyzed, and alternative methods not sought, are complex. They have to do with a positivism that, to this day, grips much of social statistics, and holds that causal hypotheses are intrinsically unscientific. Since almost all of the hypotheses of social inquiry are causal, this opinion requires a certain mental flexibility that inquiry into the reliabilities of methods of search for causal hypotheses would surely complicate. Perhaps an equally important reason is the view that causal hypotheses are theories, and theories are the special prerogative of experts, not of algorithms. These prejudices combined with a number of more technical disciplinary issues. For example, search methods are difficult to associate with any uniform measure of uncertainty, analogous to the standard error function for a parameter estimator, and social scientists and social statisticians have come to demand such measures without reflection. Again, disciplines are usually blind to their history, and although causal questions motivated much of the development of statistics, the paradigmatic tool for mathematical analysis in statistics is the theory of probability—there is no formal language in the subject for causal analysis (Pearl 1995). In some measure this state of affairs has been abetted by philosophy of science, which for generations taught that there could be no principles, no "logic," to scientific discovery.

The incoherence between practice and methodological theory would do little harm were the methods of searching for causal hypotheses that have developed in social statistics, and that are widely taught to social scientists, and widely used to justify conclusions, reliable under any set of conditions that might reasonably be assumed in the various domains to which the search methods are applied. They are not. We are left with enterprises that use the most rigorous possible methods to estimate parameters in causal models that are often produced by whimsy, prejudice, demonstrably unreliable search procedures, or, often without admission, by ad hoc search methods—sometimes reliable, sometimes not.

There is a remedy. Clear representations by directed graphs of causal hypotheses and their statistical implications, in train with rigorous investigation of search procedures, have been developed in the last decade in a thinly populated intersection of computer science, statistics, and philosophy. The empirical results obtained with these methods, including a number of cases in which the causal predictions were independently confirmed, have been good, perhaps surprisingly good. With exceptions, the reception by statisticians and philosophers has chiefly been hostile muddle or worse.

This paper illustrates some of the methodological difficulties of social statistics through a notorious example, *The Bell Curve*. I will use the directed graphical framework to explain the causal claims of the book and the unreliability of the search procedures on which the book relies. I will explain what search procedures that are asymptotically correct under plausible, general assumptions would say about the causal interpretation of the data the book uses. I will briefly describe some of the empirical research that has been carried out with new search procedures. I conclude with some remarks about social policy and about scientific strategy.

2. The Bell Curve. The *Bell Curve* is distinguished from a thousand and more efforts at social science chiefly by its length, popular style, ambition, and conclusions. The statistical methods of the book are multiple regression, logistic regression and factor analysis, techniques routinely taught to psychology and social science students in almost every graduate program in these subjects and routinely applied to make causal inferences from data of every kind. The methods and kinds of data of *The Bell Curve* are not very different in character from those in celebrated works of social statistics, for example the regression analyses in Peter Blau and Otis Dudley Duncan's *The American Occupational Structure* (1967), or the factor analyses in Melvin Kohn's *Class and Conformity* (1967); many statistical consultants use the same methods to guide business, military, and government policy on endless issues. One of the authors of *The Bell Curve*, Charles Murray, is a well-trained political scientist, and Herrnstein was a prominent psychologist; these are not *naïfs* or incompetents.

When Herrnstein and Murray write "cause" I take them to mean cause—something that varies in the population and whose variation produces variation in other variables, something that, if we could intervene and alter, would alter something else we did not directly wiggle. When they say genes cause IQ scores, I take them to mean that if somehow we could alter the relevant distribution of genes in the population, without altering directly anything else—the "environment"—then a different distribution of IQ scores would result if measured. That is how Sir Ronald Fisher (1958) thought of the causal role of genes in producing phenotypes, and it is how we think of causation in most other contexts.

Some statisticians, notably Paul Holland (1986), have claimed contrary to Fisher that it is nonsensical to talk of genes as causes. The thought seems to be that causation is a relation between individuals or between attributes of an individual, and I, for one, and you, for another, could not be who we are if our respective genetic structures were altered. The objection is wonderfully philosophical, Leibnizian even, but hardly persuasive in an age in which we can stick bits of DNA in chromosomes and re-identify the chromosome before and after the insertion.

There are two parts to the causal argument of *The Bell Curve*. One part argues that there is a feature of people, general intelligence, that is principally responsible for how people perform on IQ tests. The other part argues that this feature, as measured by IQ tests, causes a lot of other things. The first part is argued by appeal to factor analysis, the second part by regression. Because the hypotheses are causal, there is no substitute for making the causal claims explicit, and for that I will use graphical causal models. They explicitly represent important distinctions that are often lost when the discussion is couched in more typical formalisms.

3. Factor Analysis. Herrnstein and Murray rely on factor analytic studies to justify the claim that there is a single unobserved cause—which they, following Charles Spearman, call *g* for general intelligence—whose variation in the human population is responsible for most of the variation in scores on IQ tests. I want instead to consider the very idea of factor analysis as a reliable method for discovering the unobserved.

The issue is one of those delicate cases where it is important to say the right thing for the right reason. Stephen Jay Gould (in Fraser 1995) says the right thing about factor analysis—it is unreliable—but partly for the wrong reasons: that there exist alternative, distinct causal structures that are "statistically equivalent" and that entities and processes postulated because they explained observed correlations should not be "reified," that is, should not be taken seriously and literally. In the generality they are given, if not intended, Gould's reasons would be the end of science, including his own. Atoms, molecules, gravitational fields, the orbits of the planets, even the reality of the past, are all beyond the eye- and earshot that led our scientific ancestors, and lead us still, to believe in them. Physicists and philosophers of science have known for much of this century that standard physical theories—Newtonian gravitational theory, for example—admit alternative theories with different entities that equally save the phenomena. An objection that, when applied even-handedly, indicts factor analysis along with the best of our science leaves factor analysis in excellent company. The problems of factor analysis are more particular: they are the kinds of alternatives factor analytic procedures allow, the kind of restrictions the factor analytic tradition employs to eliminate alternatives, and, in consequence, the want of correspondence between factor analytic results and actual structures from which data are generated.

Herrnstein and Murray's history of factor analysis requires a correction. They say that Spearman introduced the concept of general intelligence upon noticing that scores on his "mental" tests were all strongly positively correlated. Not exactly. Spearman developed his argument in various roughly equivalent forms over half a century,[1] but it came to this: The correlations of any four mental test scores i, j, k, l satisfy three equations:

$$\rho_{ij}\ \rho_{kl} = \rho_{il}\ \rho_{jk} = \rho_{ik}\ \rho_{jl}$$

Spearman observed that these "tetrad" equations are implied by any linear structure in which scores on tests are all influenced by a single common cause, and otherwise sources of variation in test scores are uncorrelated. Graphically:

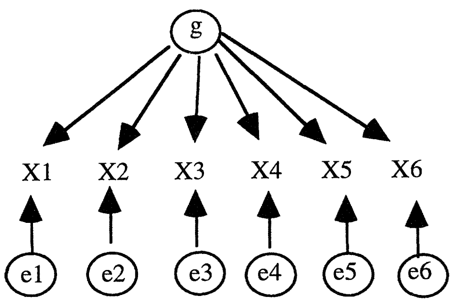

Figure 1

Spearman realized that certain alternative structures would also generate the tetrad equations, for example

1. Beginning with Spearman 1904 and ending (so far as I know) with Jones and Spearman 1950.

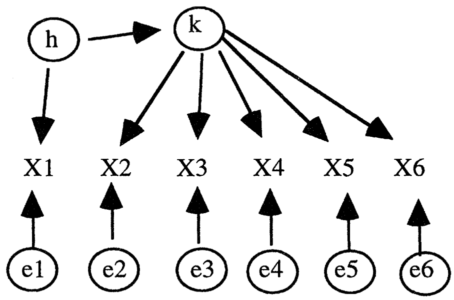

Figure 2

but he thought of such structures as simply finer hypotheses about the structure of general intelligence, g.

Spearman must have known that structures with still more latent variables can account for the data. The tetrads, for example, can be made to vanish by suitable choice of the linear coefficients when there are two or more common latent factors affecting the measured variables. Such models might be rejected on the grounds that models that postulate fewer unobserved causes are more likely to be true than those that save the same phenomena by postulating more unobserved causes, but that is a very strong assumption. A much weaker one would serve the purpose: Factor models assume that observed variables that do not influence one another are independent conditional on all of their common causes, an assumption that is a special case of what Terry Speed has called the "Markov condition" for directed graphical models. The rank constraints—of which vanishing tetrads are a special case—used in factor analysis are implied by conditional independencies in factor models, conditional independencies guaranteed by the topological structure of the graph of the model, no matter what values the linear coefficients or "factor loadings" may have. To exclude more latent variables when fewer will do, Spearman needed only to assume that vanishing tetrads do not depend on the constraints on the numerical values of the linear coefficients or "factor loadings," but are implied by the underlying causal structure. A general version of this second assumption has been called "faithfulness." It is known that the set of values of linear parameters (coefficients and variances) that generate probability distributions unfaithful to a directed graph is measure zero in the natural measure on parameter space.

To illustrate the point compare the structures:

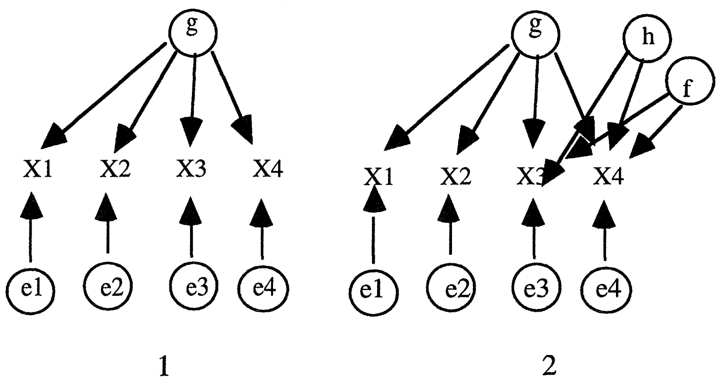

Figure 3

Let the factor loadings of g in graphs 1 and 2 be a_i, the factor loadings of h in graph 2 be b_i and the factor loadings of f be c_i, where the index is over the measured variable connected to the factor. Then in graph 1 the vanishing tetrad differences follow from the commutativity of multiplication, that is, that $a_i a_j a_k a_l = a_i a_k a_j a_l$. In graph 2, however the tetrad equation $\rho_{12}\,\rho_{34} = \rho_{13}\,\rho_{24}$ requires that $a_1 a_2(a_3 a_4 + b_3 b_4 + c_3 c_4) = a_1 a_3 a_2 a_4$, that is, $b_3 b_4 = -c_3 c_4$.

In the absence of further substantive assumptions, however, neither faithfulness nor the much stronger simplicity assumption would lead from tetrad constraints to Spearman's latent common cause models. Quite different structures also imply his tetrad equations, for example:

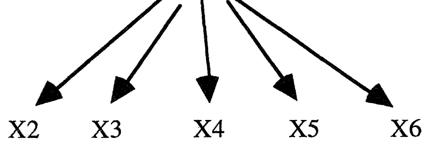

Figure 4

where I have omitted the error terms. The vanishing of all tetrads guarantees that a single common cause suffices; it does not guarantee that the common cause is unmeasured. Figures 1 and 4 are, however, distinguished by the vanishing partial correlations they require among measured variables; Figure 1 requires none, Figure 4 requires that all partials on X1 vanish. But Figure 5 is statistically indistinguishable from Figure 1:

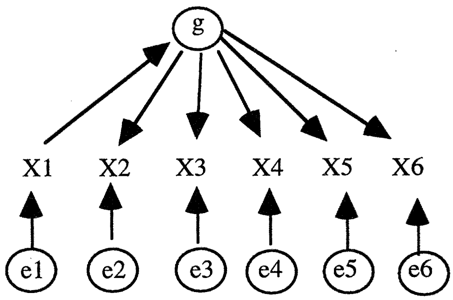

Figure 5

So far as I know, Spearman and his followers never considered these matters.

Spearman's original mental tests did not prove well correlated with teachers' and others' judgments of intelligence, and they were replaced by tests in Binet and Simon's mode. These tests in turn had more complicated correlation structures, and typically all tetrads did not vanish. Spearman's followers, notably Karl Holzinger, began the practice of assuming a single common cause, g, and then introducing additional common causes as they were needed to account for residual correlation and prevent the implication of tetrad equations not approximated in the data. Their procedure guaranteed that if most of the correlation among measures could be attributed to one common cause, it would be, even if alternative structures and factor loadings were consistent with the data. Reliability was never an issue.[2]

Thurstone (1947) said he discovered factor analysis when he realized the tetrads were merely the determinant of a second order minor. The mathematical idea in factor analysis was that the rank of the correlation matrix gives information about the minimum number of latent

2. For a more detailed history, see Glymour et al. 1987.

common causes needed to reproduce the matrix. The procedural idea was a method—the centroid method—of forming from the covariances a particular linear causal model, in which all of the correlations of measured variables are due to latent common causes. Thurstone realized that the models his procedure produced were not the only possible linear, latent variable explanations of the data from which they started, and that in fact any non-singular linear combination of latent factors obtained by his centroid method would do as well.

Thurstone's problem is fairly compared to John Dalton's. Thurstone had no means of uniquely determining the latent factor loadings and relations, and Dalton had no means of determining relative atomic weights. Both sought to remove or at least reduce underdetermination with a simplicity principle.[3] In graphical terms, Thurstone's proposal is to find the linear combination of latents that produces the fewest total number of directed edges from latent factors to measured variables. Thurstone thought such a "simple structure" is unique for each correlation matrix, but it is not. More important, why should we think actual mental structures obey Thurstone's rule of simplicity any more than atoms obey Dalton's? Unlike faithfulness, simple structure has no special measure theoretic virtue and no special stability properties.

Thurstone's factor analysis rapidly displaced Spearman's methods. Reliability does not seem to have been one of the reasons. Guilford, who discusses both in his *Psychometric Methods,* recommends factor analysis over tetrad analysis on grounds of computational tractability.

There is no proof of the correctness of any factor analytic procedure in identifying causal structure in the large sample limit. In general, factor analytic procedures are not correct even with perfect data from the marginals of distributions faithful to the actual structure. There is not even a proof that the procedures find a correct simple structure when one exists. I know of no serious study of the reliability of any exploratory factor analysis procedures with simulated data. A serious study would generate random graphs with latent variables, randomly parameterize them, generate random samples (of various sizes) from each model so produced, run the sample data through factor analytic procedures, and measure the average errors in model specification produced by the factor methods.[4]

3. Glymour 1980 gives an account of Dalton's simplicity principle and its empirical difficulties

4. It might be said that such a study would be otiose: we know the factor analytic procedures are not reliable. But we do not know how unreliable, and in what circumstances, or whether there are heuristics that can be used with factor analysis to improve reliability.

Consider the more fundamental problem of the existence of asymptotically correct, computable procedures that will, under appropriate distribution assumptions give information about latent structure when the Markov and faithfulness assumptions, or similar structural and stability assumptions appropriate to causal systems,[5] are met. There is nowhere in the statistical literature, at least to my knowledge, any informative mathematical study characterizing the bounds on reliably extracting information about latent structure from measured covariances under general distribution assumptions such as linearity and normality and measure one assumptions such as faithfulness. We do not have a classification of the structures for which faithful distributions will generate any given set of constraints on the covariance matrix of measured variables. Worse, we do not even have a classification of the constraints on the covariance of measured variables that can be implied (via the Markov condition) by partial correlation constraints in latent variable models. Analogous lacunae exist for other factor models, those with discrete observed variables and discrete or continuous latent variables. So far as Reliable Search is concerned, psychometrics is a century of sleep.

There is a little work on the fundamental questions, but so little that it is almost anomalous. Junker and Ellis (1995) have recently provided necessary and sufficient conditions for the existence of a unidimensional latent variable model of any real valued measures—representing the most recent of a sequence of papers by several authors focused on when their exists a single common cause of observed measures. It has been shown (Spirtes et al. 1993) that if the investigator provides a correct, initial division of variables into disjoint clusters such that the members of each cluster share at least a distinct latent common cause, then under certain assumptions, including linearity, unidimensional measurement models may be found for each latent, and from such models and the data, some causal relations among latents may reliably be found.

Stephen Jay Gould (in Fraser 1995) claims that one of the essential premises of *The Bell Curve* is that there is a single common factor, g, responsible for performance on intelligence tests. No doubt Herrnstein and Murray make that assumption, but it is largely inessential to their argument. If IQ scores measured a pastiche of substantially heritable

5. For example, the Markov condition fails for feedback systems, although a condition equivalent to it for acyclic graphs holds for linear cyclic systems representing feedback or "non-recursive" models. See Spirtes 1995.

features that doom people to misery, the argument of the Bell Curve would be much the same. So the more important questions concern causal relations between whatever it is IQ measures and various social outcomes. Which brings us to regression.

4. Regression and Discovery. Herrnstein and Murray begin the second part of their book with a description of some of their methods and what the methods are used to do. I ask the reader to keep in mind their account from pages 72–75. I have numbered their paragraphs for subsequent reference:

(1) The basic tool for multivariate analysis in the social sciences is known as regression analysis. The many forms of regression analysis have a common structure. There is a result to explain, the dependent variable. There are some things that might be the causes, the independent variables. Regression analysis tells how much each cause actually affects the result, taking the role of all the other hypothesized causes into account—an enormously useful thing for a statistical procedure to do, hence its widespread use.

(2) In most of the chapters of Part II, we will be looking at a variety of social behaviors, ranging from crime to childbearing to unemployment to citizenship. In each instance, we will look first at the direct relationship of cognitive ability to that behavior. After observing a statistical connection, the next question to come to mind is, What else might be another source of the relationship?

(3) In the case of IQ the obvious answer is socioeconomic status. . . . Our measure of SES is an index combining indicators of parental education, income, and occupational prestige . . . Our basic procedure has been to run regression analyses in which the independent variables include IQ and parental SES. The result is a statement of the form "Here is the relationship of IQ to social behavior X after the effects of socioeconomic background have been extracted," or vice versa. . . .

The causal picture Herrnstein and Murray seem to have in mind is this:

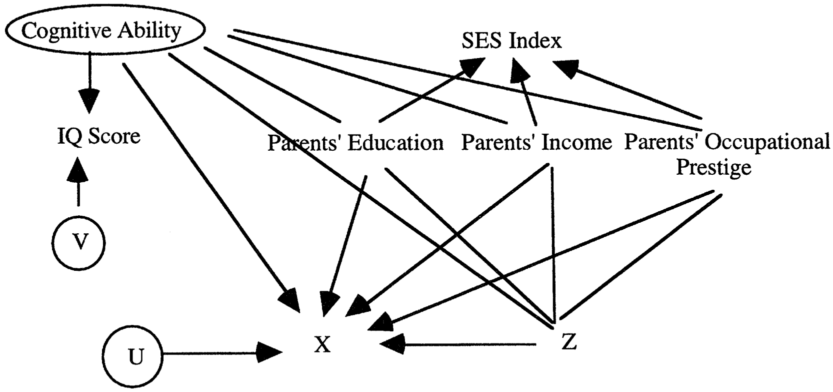

Figure 6

where the features in circles or ovals are unobserved, and the lines without arrows indicate statistical associations that may be due to influences in one direction or the other, or to unobserved common causes, or both. Z varies from case to case; often it is age.

If this were the correct causal story, then provided that very little of the variation in IQ scores between individuals were due to V, one could estimate the influence of cognitive ability on X by the two methods Herrnstein and Murray use: multiple regression of X on IQ and SES index when the dependencies are all linear, and by logistic regression on those variables under other distribution assumptions. By "could estimate" I mean the expected values of estimates of parameters would equal their true values.

I will sometimes simplify the diagram in Figure 6 as Herrnstein and Murray simplify their discussion:

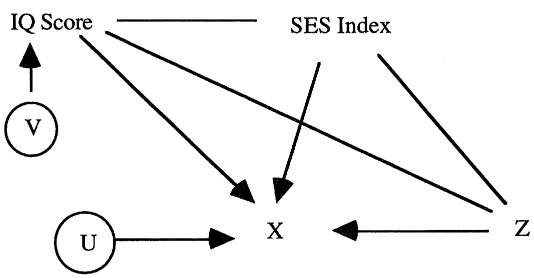

Figure 7

Under the assumptions just mentioned, if estimates of the influence of IQ score based on the causal model of Figure 6 are correct, so are estimates of IQ based on the simpler surrogate structure of Figure 7.

Now the standard objection to assuming something like the structure of Figure 6 or Figure 7 is put in terms of "correlated error." The objection is that in the corresponding regression equation, the error term, U, for X may be correlated with any of IQ, SES, and Z, that such correlation cannot be detected from the data, and that when it exists the regression estimates of the influence of cognitive ability on X will be incorrect. "Correlated error" is jargon—euphemism, really—for those who want to avoid saying what they are talking about, namely causal relations. Unless correlations arise by sheer chance, the correlation of U and IQ, say, must be due to some common causal pathway connecting IQ scores with whatever features are disguised by the variable U. A "correlated error" between a regressor such as IQ and an outcome variable, X, is the manifestation of some unknown cause or causes influencing both variables.[6]

Suppose something else, denoted by W—mother's character, attention to small children, the number of siblings, the place in birth order, the presence of two parents, a scholarly tradition, a strong parental positive attitude towards learning, where rather than how long parents went to school, whatever, influences both cognitive ability and X. Then the regression estimates of the influence of cognitive ability on X based on the model in Figure 6 will compound that influence with the association between cognitive ability and X produced by W.

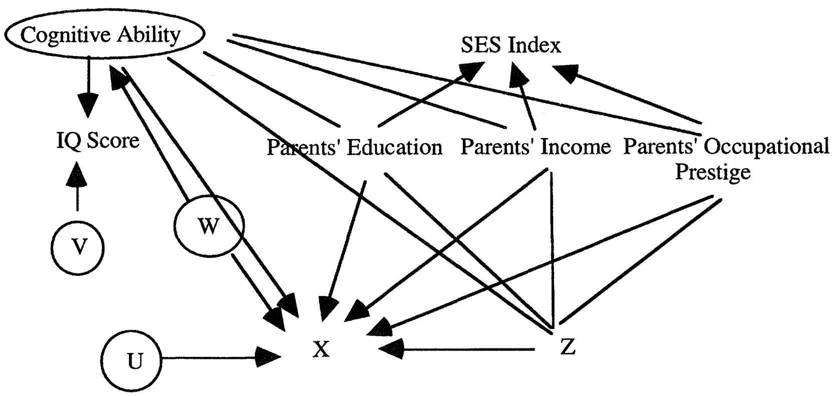

Figure 8

6. There is an open technical issue here. There are cases in which a covariance matrix generated by a model with correlated error cannot be reproduced by that model but with each correlated error replaced by a distinct latent variable and the latent variables are uncorrelated—the question is whether such matrices can always be reproduced from an appropriate latent variable structure.

or more briefly:

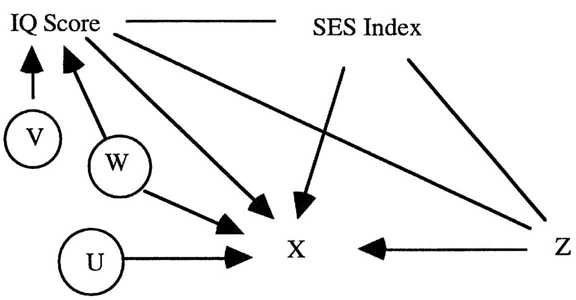

Figure 9

Here is how Herrnstein and Murray respond:

(4) We can already hear critics saying, "If only they had added this other variable to the analysis, they would have seen that intelligence has nothing to do with X." A major part of our analysis accordingly has been to anticipate what other variables might be invoked and seeing if they do in fact attenuate the relationship of IQ to any given social behavior.

This sounds quite sensible, until one notes that none of the possible confounding variables suggested above, nor many others that can easily be imagined, are considered in *The Bell Curve*, and until one reads the following:

(5) At this point, however, statistical analysis can become a bottomless pit . . . Our principle was to explore additional dynamics where there was another factor that was not only conceivably important but for clear logical reasons might be important because of dynamics having little or nothing to do with IQ. This last proviso is crucial, for one of the common misuses of regression analysis is to introduce an additional variable that in reality is mostly another expression of variables that are already in the equation.

There is a legitimate concern in this remark, which does not, however, excuse the neglect: If W is an *effect* of cognitive ability, then including W among the regressors will omit the mechanism that involves W and will lead to an incorrect estimate of the overall influence of cognitive ability on X:

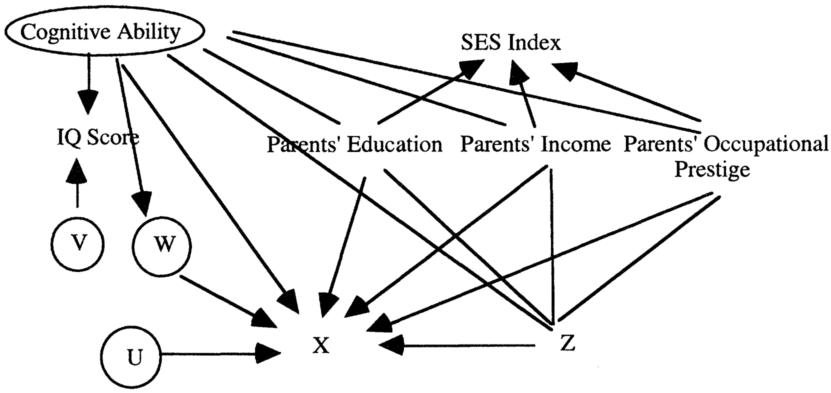

Figure 10

Against Herrnstein and Murray's remark in paragraph (5), however, it is exactly the presence of other variables that are common causes of X and of cognitive ability, and therefore "having to do" with cognitive ability, that lead to the "correlated errors" problem in estimating the influence of cognitive ability on X. Omitting such variables, if they exist, ensures that the regression estimates of effects are wrong. *The surprising fact is that the regression estimates may very well be wrong even if such variables are included in the regression.* That requires some explanation.

The authors of *The Bell Curve* have been criticized for omitting the subjects' educations from their set of regressors, an omission about which I will have more to say later. But their analysis would have been no better for including education. Suppose the true causal structure were as in Figure 10, with W representing years of education. Then multiple regression with education included would mistake the influence of cognitive ability on X, because it would leave out all pathways from cognitive ability to X that pass through W. At least, one might say, a regression that includes education would tell us how much cognitive ability influences X *other than* through mechanisms involving education, SES and Z. But even that is not so. Depending on whether there are additional unmeasured common causes of education and X, the error in the estimate of the separate effect of cognitive ability on X might be positive or negative, and arbitrarily large. There are circumstances, arguably quite common circumstances, in which assumptions about distribution families (normal, etc.) are satisfied, and there is no "correlated error" between an outcome variable X and a regressor such as cognitive ability—that is, there is no unmeasured common cause of

X and the regressor—but regression nonetheless mistakes the influence of the regressor on the outcome. Suppose the actual structure is as in Figure 11:

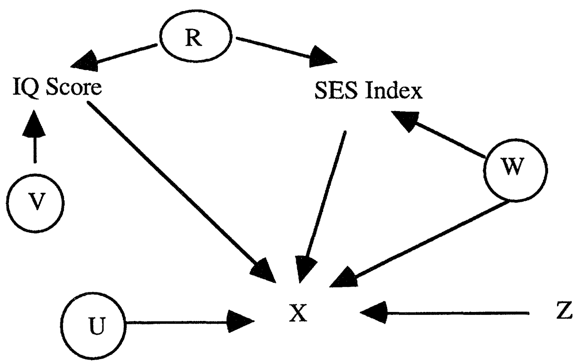

Figure 11

Notice that there is no unmeasured common cause of IQ and X, no correlation of the error term with IQ in the regression equation for X, but the error term in the regression equation is correlated with another regressor, SES. In that case, multiple regression of X on IQ, SES and Z will give an incorrect estimate of the influence of IQ on X. The error of the estimate can be arbitrarily large and either positive or negative, depending on the values of the parameters associated with the unmeasured R and W variables. For all we know, the subjects in the data Herrnstein and Murray study are rich in such R's and W's.

Critics have noted that the SES index Murray and Herrnstein use is rather lame, but the criticism is largely beside the point. Suppose they had used a better index, compounded of more measured features of the subjects and their families. The variables in SES indices may be strongly correlated, but they typically have no single common cause—those Murray and Herrnstein use demonstrably do not.[7] So a better index would add a lot of causally disparate measures together. Wouldn't that make it all the more likely that there are unmeasured variables, structurally like W in Figure 10, influencing X and also influencing one or more of the components of SES? I think so.

Adding extra variables to their study would not necessarily improve

7. Herrnstein and Murray give a correlation matrix for their four SES variables. The TETRAD II program will automatically test for vanishing tetrad differences not implied by vanishing partial correlations in the matrix. If there is a single common cause there should be three such differences. There are none.

the accuracy of their estimates and might make them much worse, but leaving extra variables out may result in terribly inaccurate estimates.

Herrnstein and Murray remark that an obvious additional variable to control for is education, but they do not, first because years of education are caused by both SES and IQ, second because the effect of education on other variables is not linear and depends on whether certain milestones—graduations—have been passed, third because the correlation of education with SES and IQ makes for unstable estimates of regression coefficients, and fourth

(6) to take education's regression coefficient seriously tacitly assumes that education and intelligence could vary independently and produce similar results. No one can believe this to be true in general: indisputably giving nineteen years of education to a person with an IQ of 75 is not going to have the same impact on life as it would for a person with an IQ of I25.

(7) Our solution to this situation is to report the role of cognitive ability for two sub populations of the NLSY that each have the same level of education: a high school diploma, no more and no less in one group; a bachelor's degree, no more and no less, in the other. This is a simple, but we believe reasonable, way of bounding the degree to which cognitive ability makes a difference independent of education.

The third reason is unconvincing, since SES and IQ are already strongly correlated. The last reason, in paragraph (6), is unconvincing as a ground for omitting education from the analysis, but correct in supposing there is an interaction. The interesting thing, however, is the alternative procedure suggested in paragraph (7), since it reveals a problem about causal inference that may trouble a great deal of work in social science and psychology.

Herrnstein and Murray make it plain—they even draw the graph—that they have in mind a particular causal picture:

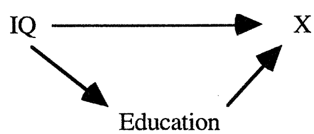

Figure 12

If this is the correct structure, then if there is no interaction between IQ and education in their influence on X, one way to estimate the direct effect of IQ on X is to condition on any value of education. The point of conditioning on two values of education, I take it, is to give us some idea of how much the interaction makes this estimate unstable.

Here is the problem: What if Figure 12 is *not* the correct causal structure? What if, instead, the correct causal structure is

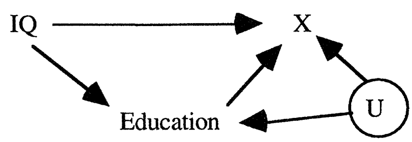

Figure 13

whatever U may be. In that case, the association between IQ and X conditional on a value of education will not be a measure of the direct influence of IQ on X, and the error can be as large as you please, positive or negative, depending on U and the parameters associated with it.

This sort of problem, *sample selection bias,* can occur whenever membership in a sample is influenced by variables whose influence on one another is under investigation. It may happen, for example, when using a sample of hospitalized patients, or when using college students as subjects in psychological experiments, or when subjects in a longitudinal study are lost, simply by using a subsample determined by values of a variable with complex causal relations, as Herrnstein and Murray do.

5. The Problems of Causal Inference. Herrnstein and Murray use the tools their professions, and social statistics generally, gave to them. The tools are incompetent for the use Herrnstein and Murray put them to, but what else were they to do? What else can anyone do who is trying to understand the causal structure at work in processes that cannot be controlled experimentally?

Consider for a moment some of the difficulties in the problem of trying to infer causation from observed correlations:

0. Little may actually be known beforehand about the causal relations, or absence of causal relations among variables. In typical social studies, time order often provides the only reliable information—negative information, at that—about cause and effect.
1. Observed associations may be due to unmeasured or unrecorded common causes.
2. A vast number of alternative possible hypotheses—the larger the number of measured variables, the more astronomical the set of

possible causal structures. When latent variables are allowed the number of possible causal structures is literally infinite.

3. Several or even a great many hypothetical structures may equally account for the same correlations, no matter how large the sample, and in finite samples a great many models may fit the data quite well.
4. The sample may be unrepresentative of a larger population because membership in the sample is influenced by some of the very features whose causal relations are the object of study.
5. The sample may be unrepresentative by chance.
6. Values for sundry variables may be unrecorded for some units in the sample.
7. The joint distribution of variables may not be well approximated by any of the familiar distributions of statistics. In particular, there may be combinations of continuous variables and variables that take only a finite set of values.
8. Relations among variables may be complicated by feedback, as between education and IQ.

Many of the same difficulties beset causal inference in experimental contexts, even though experimental design aims to remove the possibility of confounding common causes of treatment and to maximize prior knowledge of the causal structure of the experimental system. Psychological experiments often concern unobserved and uncontrolled features; clinical experiments sometimes try to investigate multiple treatments and multiple outcomes simultaneously, with entirely parallel problems about confounding and feedback, especially in longitudinal studies. Sample selection and attrition in experiments, especially experiments with humans, can create selection bias as in (4) and can result in missing values. The distribution of treatments in experiments is controlled by the experimenter but the distribution of outcomes, which may conform to no familiar pattern, is not.

We can imagine a black-box that addresses these problems. Data and relevant beliefs are put in, causal information comes out, and inside the box the problems just listed are taken account of. The box is imaginary, of course. There are no methods available that more or less automatically address all of these problems. There is no computer program that will take the data and prior knowledge, automatically take account of missing values, distributions, possible selection bias, possible feedback, and possible latent variables, and reliably and informatively give back the possible causal explanations that produce good approximations to the data, information about error bounds, or posterior probabilities. But we can think of the box as an ideal, not only for inference but also for forcing practitioners to cleanly separate the claims they make before examining the data from the claims they believe are warranted by the data. How close do the methods used by Herrnstein and Murray and other social scientists come to the ideal box? And how close could they come were they to use available, if nonstandard, methods?

Let us leave aside the problems 4 through 9, and suppose that our samples are nice, distributed nicely—normally, say—there are no missing values and no feedback, and no sample selection bias. Consider for a moment in this context using regression to decide a simpler question than estimating the influence of cognitive ability on X from ideal data on X, cognitive ability, and a definite set of other regressors: Does cognitive ability have any influence *at all* on X? Multiple regression will lead to a negative answer when the partial regression coefficient for cognitive ability is not significantly different from zero. That is, under a normal distribution, essentially an assumption connecting the absence of causal influence with a conditional independence fact, namely that *cognitive ability does not (directly) influence X if and only if cognitive ability and X are independent conditional on the set of all of the other regressors.*

We have in effect observed by example in the previous section that the italicized principle is false, in fact intensely false. Indeed, without *a priori* causal knowledge, there is no way to get reliable causal information of any sort from multiple regression. If one should be so fortunate as to know independently of the data analysis that there are no common causes of any of the regressors and the outcome variable, and the outcome variable is not a cause of any of the regressors, then under appropriate distribution assumptions, regression gives the right answer. Otherwise, not.

Regression does a funny thing: to evaluate the influence of one regressor on X, it conditions on *all other* regressors, but not on any proper subsets of other regressors. Stepwise regression procedures typically do investigate the dependence of a regressor and X conditional on various subsets of other regressors, but they do so in completely ad hoc ways, with no demonstrable connection between the procedures and getting to the truth about causal structure. Regression and stepwise regression reflect intuitions from experimental design and elsewhere that absence of causation has something to do with conditional independence. They simply do not get the something right. The correct relationship is far more complicated.

Fifteen years ago, Terry Speed and his student, Harry Kiiveri (1982), introduced a correct relation, which, with some historical inaccuracy, they called a Markov condition. Speed has since testified to the correctness of the principle in the most infamous trial of our time. We considered the condition informally in discussing factor analysis; now consider it a little more exactly. Understanding the condition requires the notion of one variable Y, say, being a *direct cause* of another, X, relative to a set of variables **D** to which X and Y both belong. Y is a direct cause of X relative to **D** if there is a causal pathway from Y to X that does not contain any other variable in **D**—in other words, there is no set of variables in **D** such that if we were to intervene to fix values for variables in that set, variations in Y would no longer influence X. We need one further preliminary definition: I will say that any set, **D**, of variables is *causally sufficient*, provided that for every pair, X, Y of variables in **D**, if the directed graph representing all causal connections among variables in **D** contains a variable Z which is the source of two non-intersecting (except at Z) directed paths, respectively to X and to Y, then Z is in **D.**

Causal Markov Condition: For any variable X, and any set of variables **Z** that are not effects of X (and does not have X as a member), and any causally sufficient set **D** of variables including **Z** and having X as a member, X is independent of **Z** conditional on the set of members of **D** that are direct causes of X—the set of parents of X in the directed graph of causal relations.

When true, the Markov condition gives a sufficient causal condition for conditional independence. A converse condition gives necessity:

Faithfulness Condition:[8] All conditional independencies in a causal system result from the Causal Markov condition.

The scope of the Markov condition is occasionally misunderstood by philosophical commentations: *As a formal principle about directed graphs and probability distributions, the Markov Condition is necessarily true if exogenous variables (including errors or noises) are independent, and each variable is a deterministic function of its parents (including among parents, any errors or noises). The form of the functional dependence is irrelevant.*

In systems whose causal structure is represented by a direct acyclic graph and the system generates a probability distributions meeting the Markov condition for that graph, the faithfulness condition can only fail if two variables are connected by two or more causal pathways (either from one variable to another or from a third variable to both) that exactly, perfectly, cancel one another, or if some of the relations among variables (excluding error terms) are deterministic. In practice both the Markov and faithfulness conditions are consistent with almost every causal model in the social scientific literature, non-linear models

8. The formulation of the condition is due to Pearl (1988).

included, that does not purport to represent feedback or reciprocal influence.

We can use these two conditions to discover what the conditional independencies implied by the structures that Herrnstein and Murray postulate could *possibly* tell us, by any method whatsoever, about those structures. That is, we will suppose their causal story is correct and ask whether they could reasonably infer it is correct from data that nicely agrees with it. To do so we need some simple representations of ignorance about causal structures.[9] Here is a convenient code:

X o-o Y	X is a cause of Y, or Y is a cause of Y, or there is a common unmeasured cause of X and of Y, or one of X, Y causes the other and there is also an unmeasured common cause.
X o→Y	X is a cause of Y, or there is a common unmeasured cause of X and of Y, or both
X↔Y	There is a common unmeasured cause of X and of Y, but neither X nor Y influences the other.

With these conventions, here is what the conditional independencies implied by the causal hypothesis of Figure 7 tell us about causal relations:

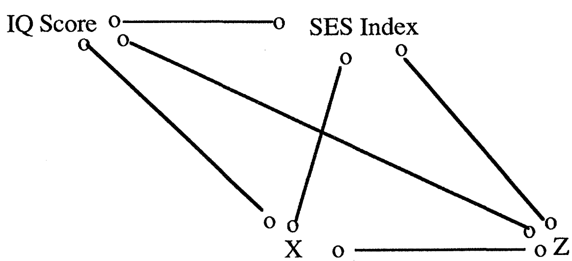

Figure 14

Nothing at all about whether IQ score is a cause of X. Suppose that common sense tells us that X is not a cause of the other variables. That does not help much. The result is:

9. The discussion here applies results in Spirtes et. al. 1993.

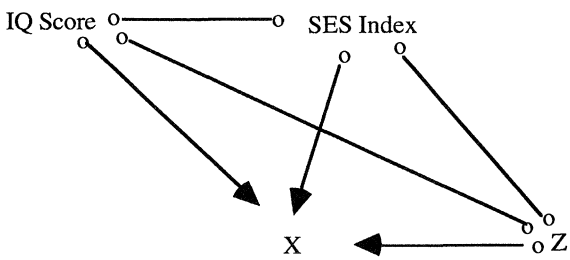

Figure 15

We still cannot tell whether IQ has any influence at all on X. For all we know, from the data and the prior knowledge, the association between IQ scores and X is produced entirely by the variation of unmeasured factors that influence both IQ score and X. The sizes and signs of the observed covariances in this case would give no other extra information about the actual causal structure.

The Markov and faithfulness conditions also entail that there are possible causal relations we can determine from observed associations, provided we have none of the problems (4)–(9) listed above. For example, suppose we have measures of A, B, C and D, and their causal relations are actually as in Figure 16:

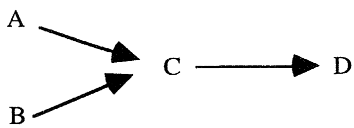

Figure 16

Then according to the two conditions, we can determine from independence facts that:

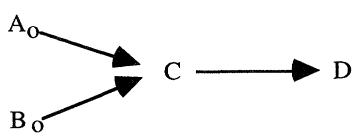

Figure 17

and so that C is actually a cause of D. Moreover, there is a certain robustness to the determination, for if we were to decide that the independencies corresponding to Figure 16 obtain when in fact they do not quite because of a small common cause of C and D, if the association of A and C or B and C is large, then (in the linear case at any rate) the estimate of the influence of C on D obtained using the result in Figure 17 will be a good approximation to the truth. There is here a general moral—almost never observed—about the kind of data one should seek if causal relations are to be inferred from observed data.

The Markov and faithfulness conditions can just as well give us information about the presence of unmeasured common causes (see Spirtes et al. 1993 for details).

These remarks would be of little practical use if in any application one were required to prove some intricate theorem, distinct for almost every case, characterizing the structures consistent with prior knowledge and the patterns of independence and conditional independence found in the data. No such effort is necessary. There is a general algorithm,[10] commercially available in the TETRAD II program, that does the computations for any case. The finite sample reliabilities of the program have been pretty thoroughly investigated with simulated data; it is known, for example, that they tend to give too many false arrowheads, and that for reasonable sample sizes they rarely produce false adjacencies. The procedures are rarely used, certainly not by Herrnstein and Murray or their critics. Were they used, social scientists would at least be forced to be entirely explicit about the causal assumptions that they have forced on their data analysis.

In keeping with social scientific tradition, Herrnstein and Murray give endless pages of statistical conclusions but their data are all but hidden; one has to go to the original sources and know the sample they selected from it. Although Herrnstein and Murray report any number of linear regressions with results determined entirely by a simple covariance matrix, they give only one such matrix in the entire book, and no count data. Even so, we have excellent reason to think that scientific searches applied to the data they use would turn up structures such as those in Figure 18, permitting no causal inferences of the kind Murray and Herrnstein wish to draw.

6. Some Empirical Cases. Of course it is easy to avoid making erroneous causal inferences: make none at all. The challenge is to find methods that make correct causal inferences where possible while avoiding, insofar as possible, incorrect causal inferences. The positive value of the search methods can be illustrated by considering a problem posed by the intelligence test Herrnstein and Murray use, the Armed Forces Qualification Test.

Before 1989 the AFQT score was a weighted sum of several component tests, and was itself a part of a larger battery that included tests

10. Given in Spirtes et. al. 1993.

that were not included in AFQT. Consider data for AFQT and the following seven tests from a sample of 6224 subjects in 1987:

Arithmetical Reasoning (*AR*)
Numerical Operations (*NO*)
Word Knowledge (*WK*)
Mathematical Knowledge (*MK*)
Electronics Information (*EI*)
General Science (*GS*)
Mechanical Comprehension (*MC*)

Here is the problem. Given the information above, and the correlations, and nothing more, determine which of the seven tests are components of AFQT.[11] The problem is at least partly linear, because we know AFQT depends linearly on some of these tests, perhaps not others, and also on some tests whose scores are not in the data. Of course we do not know that relations among the scores on various tests are produced by linear processes.

If you run the correlations through a linear regression package you will find that two of the seven tests have insignificant regression coefficients, two have significant negative coefficients, and three (NO, WK and AR) have significant positive coefficients. The tests with negative coefficients might be components of AFQT that are negatively correlated with test components not included in the data set, or for some reasons these tests may be weighted with a minus sign in the AFQT score, as they would be if higher scores on the tests indicate less ability.

The TETRAD II program permits the user to determine the structures compatible with background knowledge and a data set under various assumptions, including the assumption that no latent variables are acting, or alternatively without that restriction. Assuming there are no latent variables, the procedure finds that only four of the seven tests are components of AFQT; one of the tests with a significant but negative regression coefficient is eliminated in addition to the two tests with insignificant regression coefficients. The assumption that there are no latent variables is not very plausible, and removing it, the TETRAD II program finds that only NO, WK, and AR are adjacent to AFQT, although the adjacency is a double-headed arrow in all three cases. Recognizing that procedure rarely makes false positive adjacencies, and it is liable to false positive arrowheads, one concludes that the components of AFQT among the seven tests are probably NO, WK and AR, which is the correct answer, one we found with an experimental version of the TETRAD program using only the information

11. The covariance matrix is given in Spirtes et al. 1993.

given here, not only initially in ignorance of the actual components of AFQT, but burdened with the false information that all seven tests are components.

Several other applications have been made of the techniques, for example:

1. Spirtes et al. (1993) used published data on a small observational sample of Spartina grass from the Cape Fear estuary to correctly predict—contrary both to regression results and expert opinion—the outcome of an unpublished greenhouse experiment on the influence of salinity, pH and aeration on growth.
2. Druzdzel and Glymour (1994) used data from the US News and World Report survey of American colleges and universities to predict the effect on dropout rates of manipulating average SAT scores of freshman classes. The prediction was confirmed at Carnegie Mellon University.
3. Waldemark used the techniques to recalibrate a mass spectrometer aboard a Swedish satellite, reducing errors by half.
4. Shipley (1995, 1997, in review) used the techniques to model a variety of biological problems, and developed adaptations of them for small sample problems.
5. Akleman et al. (1997) have found that the graphical model search techniques do as well or better than standard time series regression techniques based on statistical loss functions at out of sample predictions for data on exchange rates and corn prices.

7. Projects and Attitudes. The TETRAD II program represents only one of several approaches to automated, reliable model specification, steps toward the ideal inference box. Bayesian methods are under investigation, as are methods that combine Bayesian methods with the constraint-based procedures used in the TETRAD programs. Procedures have been described, and under explicit assumptions proved correct, that make a more limited set of causal inference when sample selection bias may be present. The same is true when feedback may be present in linear systems.[12] Research is under way on understanding constraints other than conditional independence relations.

There may never be an inference box that addresses all of the problems of causal inference from observational data, but there can certainly be boxes that can help social and behavioral scientists do better than they will armed only with their preconceptions, factor analysis, and regression. The more model specification is automated and data

12. See Cooper 1995 and Spirtes, Meek, and Richardson 1995.

driven, the more substantive prior assumptions are separated mechanically from inferences made from the data, the more algorithms in the box give out only information justified by explicit assumptions, the less likely is the kind of work *The Bell Curve* represents.

The statistical community, by and large and no doubt with important exceptions, thinks otherwise. Terry Speed once told me that although he agreed that regression is demonstrably unreliable in general, no other methods should be developed or used, even if they are demonstrably more reliable. When a large simulation study showed that the search procedures in their commercial program, LISREL, are wildly unreliable in cases in which there are latent causes of measured variables, and also influences of measured variables on one another, two prominent statisticians, Karl Joreskog and Dag Sorbom (1990), claimed that no such cases are possible: God has arranged that the world can only be as LISREL finds it. These two disparate attitudes have nearly banished from social statistics serious inquiry into methods of search that reliably get to the truth about the questions we most urgently care about from any data we could possibly have. The questions will not go away, and in the absence of answers psychologists and social scientists, whatever their motives, will make use of the methods in which they have been educated. *The Bell Curve*, and innumerable less famous works no more worthy of credence, are the result.

8. Policy. It would be too academic and bloodless to pass by the example *The Bell Curve* offers without saying something about the policy issues at stake. Perhaps surprisingly, Herrnstein and Murray's conclusions about what is happening to American society are substantially the same as those drawn from less evidence by celebrated social democrats (the former Secretary of Labor, Robert Reich, for example), and the same as those of uncountable reports by liberal-minded institutes concerning the state of education and social relations in this country. With liberals, Herrnstein and Murray say the economy increasingly rewards high-level intellectual skills and increasingly penalizes those without such skills; with liberals, they say that unimpeded, the effect is to the disadvantage of nations whose populace is comparatively unskilled in these ways. Within nations those who are talented and skilled at "symbol manipulation" are increasingly segregated from those who are not; public education achieves a level of knowledge and skill that makes Americans stupid compared with citizens of other industrialized nations; America is still socially segregated by race. The principal difference between liberal writers such as Reich and the authors of *The Bell Curve* is that Herrnstein and Murray offer a lot more empirical arguments, some relevant, some not, and conclude against the popular social institutions and interventions—public education, affirmative action, Head Start—that Reich would endorse.

The Bell Curve comes apart exactly when it moves from formulating our social problems to recommending solutions. The book reads like the broadsides of intellectual flacks in the Kennedy and Johnson administrations, the Richard Goodwins of the 1960s and 1970s, who reported the evils of the conflict in Vietnam but always ended with an illogical plea for continuing that war. Sensibly read, much of the data of *The Bell Curve*, as well as other data the book does not report, demands a revived and rational liberal welfare state, but instead the book ends with an incoherent, anti-egalitarian plea for the program of right-wing Republicans.

Consider the decline of the two-parent family. Illegitimacy and divorce come just after atheism in the right wing catalog of the causes of the moral decay of our nation. According to *The Bell Curve*, the dissolution of the family is caused by stupidity, and for all I know that may indeed be a cause. According to more liberal sociologists, William Julius Wilson for example, a more important cause of single parenthood among urban African Americans is the absence of good jobs for unskilled workers, and that may be a cause, too, and a more important one. But two-parent families are in decline throughout industrialized nations, and past social experimentation has shown that one of the first things poor families do when given a guaranteed income is to divorce (see Murray 1984 for a review of the data).

A household is a business given over to caring for small, temporarily insane people, a business subject to cash-flow problems, endless legal harassments, run by people who expect to have sex with each other, who occupy the same space, and who go nuts when either party has sex with anyone else. Once in marriage, a lot of people try to get out as fast as religious tradition, poverty, or devotion to children permits.

The evil social effects, if any, of illegitimacy and broken homes are best addressed by finding a social structure that will replace whatever benefits to children two parents are supposed to provide, rather than by forcing people who cannot stand each other to live together. We can make some reasonable guesses about what children require: security, discipline, stimulation, care, affection, ideals. Sounds like what a school ought to provide, which brings us to education.

Having described a nation Balkanized by race, gender, cognitive ability, income, and occupation, a nation whose only unifying forces are public schools and MTV, *The Bell Curve* nevertheless concludes that education cannot much help to solve our social problems, and we should privatize education in the way Milton Friedman incessantly urges. The consequences are predictable: Ku Klux Klan Schools, Aryan Nation Schools, The Nation of Northern Idaho Schools, Farrakhan Schools, Pure Creation Schools, Scientology Schools, David Koresh Junior High, and a thousand more schools of ignorance, separation, and hatred will bloom like some evil garden, subsidized by taxes. If ever there was a plan to make America into Yugoslavia, that is it.

American schools are run on a complicated, decentralized system in which state and then local school boards have enormous powers, which they mostly use badly. Herrnstein and Murray document some of the results—low standards, an expectation of mediocrity—every intellectually anxious parent knows the phenomena first hand. (Murray and Herrnstein do not document some of the more imaginative practices of local schools. In parts of Utah, religion is not required of high school students. They have elective options, for example, either Mormon Theology or Advanced Quantum Field Theory, choose one.)

The obvious solution is not privatization or more decentralization, but national educational standards of accomplishment with uniform, national tests, and national school funding. More bureaucrats the Republicans say, as though we are better off with real estate brokers running school boards and fighting, as Herrnstein and Murray point out, against more homework. I say a public school that worked as well as the post office would be a godsend.

Schools are the natural place to provide security, stability, and stimulation to children of parents who, for whatever reasons, cannot provide them. Our present schools will not do, not schools that are closed 16 hours a day nine months a year and 24 hours a day 3 months a year and always closed to people under five. What would do is schools that are always open to children from one month to 17 years, always welcoming, always safe, offering meals and fun and learning and, if need be, a place to sleep. Those schools are the sane and comparatively economical way to create and sustain a civil society.

Herrnstein and Murray review data that all say the same thing: early, intensive pre-schooling, as in good Head Start programs, on average improves children's performance on IQ tests by half a standard deviation or more, and the first three years of public schools eliminates the benefit. With genuine perversity, determined to make three of two plus two, they blame the early educational interventions, not the public schools they blame everywhere else. Herrnstein and Murray say that even if intensive early intervention and schooling works, it is too expensive, requiring too many adult teachers per toddler. It may be too expensive if teachers are paid at some urban and suburban rates, often the equivalent of about seventy dollars an hour. But real, round the clock, round the calendar, throughout infancy and childhood schooling is affordable if teachers are paid reasonably. And the opportunity costs of failing to do so are greater than the 100 billion or so such a program would cost each year.

Herrnstein and Murray just hate affirmative action programs that fade from broadening searches to secretly imposing quotas, and especially they hate racial, ethnic, and gender quotas in university admissions. I dislike them as well, chiefly for a reason that Herrnstein and Murray are unwilling to take seriously even when their data compel it: compensatory efforts come at the wrong time in life, too late to make a difference to most, and too concerned with credentials to make a lot of difference other than to position and income. (Herrnstein and Murray rightly complain that we are an overcredentialed society, and here, at least, I agree: The Commonwealth of Pennsylvania requires a Master's Degree to open a spot in a shopping mall to guard and entertain children while parents shop, and a school nurse, who is prohibited from doing anything except calling home and giving emergency first aid, requires a special college degree, and where, no kidding, a superintendent of schools is paid extra for a mail order doctoral degree.) Herrnstein and Murray lay no clear blame and propose no clear remedies, but the remedy is obvious. The blame is with universities and college professors, who profit from every legal restriction that requires or rewards formal education.

Disappointed and disapproving, Herrnstein and Murray predict we will become a custodial society in which the rich and competent support the many more who are poor and incompetent. They entertain no alternatives except vagaries about living together unequally with "valued places" for all. But if a society wants another way to be, there are ready alternatives. Here is an alternative vision, one I claim better warranted by the phenomena Herrnstein and Murray report: nationalized, serious, educational standards, tax-supported day and night care, minimal universal health care, a living minimum wage, capital invested in systems that enable almost anyone with reasonable training to do a job well.[13]

13. Although there is a lot of anxiety that computerization increases wage segregation, the evidence is inconclusive; economic returns to computer use are about the same as economic returns to pencil use. Arguably, where computer systems are widespread and well-designed, we should expect the reverse. The computer is a cognitive prosthesis: it enables people without special gifts, given a little training, to do many tasks—accounting, inventory control, arithmetic, scheduling, sending messages, etc.—as well or better than those with special talents. It is the great equalizer, provided there is adequate software and hardware capital, and adequate education.

A third alternative, urged by Newt Gingrich, Phil Gramm, and company, is to remold this country into a nation in which the state does not promise children safety, or nutrition, or education, and does not guarantee adults a living wage, minimal health services, or security against the hazards of industry, a nation pretty much like, well, *Honduras*?

REFERENCES

Akleman, Derya G., David A. Bessler, and Diana M. Burton (preprint), "Modeling Corn Experts and Exchange Rates with Directed Graphs", Texas A&M University, Department of Economics, February 1997.

Blau, Peter M. and Otis Dudley Duncan (1967), *The American Occupational Structure.* New York: Wiley.

Cooper, Gregory F. (1995), "Causal Discovery from Data in the Presence of Selection Bias", Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, FL, pp. 140–150.

Druzdzel, Marek J., and Clark Glymour (1994), "Application of the TETRAD II Program to the Study of Student Retention in U.S. Colleges", Technical report, American Association for Artificial Intelligence. Menlo Park, CA: AAAI Press, pp. 419–430.

Fienberg, Steven et al. (forthcoming) *Chance.*

Fisher, Ronald A. (1958), *The Genetical Theory of Natural Selection.* New York: Dover.

Fraser, Steven (ed.) (1995), *The Bell Curve Wars.* New York: Basic Books.

Glymour, Clark (1980), *Theory and Evidence.* Princeton: Princeton University Press.

Glymour, Clark, Richard Scheines, Peter Spirtes, and Kevin Kelly (1987), *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling.* Orlando: Academic Press.

Hernstein, Richard J. and Charles Murray (1994), *The Bell Curve: Intelligence and Class Structure in American Life.* New York: Free Press.

Holland, Paul (1986), "Statistics and Causal Inference", *Journal of the American Statistical Association* 81: 945–960.

Jones, Ll. Wynn and Charles Spearman (1950), *Human Ability, A Continuation of the "Abilities of Man".* London: Macmillan.

Joreskog, Karl and Dag Sorbom (1990), "Model Search with Tetrad II and LISREL", *Sociological Methods and Research* 19: 93–106.

Junker, Brian W. and Seymour Ellis (preprint), "A Characterization of Monotone Unidimensional Latent Variables", Department of Statistics, Carnegie Mellon University, 1995.

Kiiveri, Harry and Terry Speed (1982), "Structural Analysis of Multivariate Data: A Review", in Samuel Leinhardt (ed.), *Sociological Methodology.* San Francisco: Jossey-Boss.

Kohn, Melvin L. (1967), *Class and Conformity; a Study of Values.* Homewood, Ill.: Dorsey Press.

Murray, Charles (1984), *Losing Ground: American Social Policy 1950–1980.* New York: Basic Books.

Pearl, Judea (1988), *Probabilistic Reasoning Systems: Networks of Plausible Inference.* San Mateo: Morgan Kaufman.

Shipley, Bill (1995), "Structured Interspecific Determinants of Specific Leaf Area in 34 Species of Herbaceous Angiosperms", *Functional Ecology* 9: 312–319.

———. (1997), "Exploratory Path Analysis with Applications in Ecology and Evolution", *The American Naturalist* 149: 1113–1138.

Shipley, Bill and M. F. McKenna (in review; submitted to *Functional Ecology*) "Components of Interspecific Variation in the Relative Growth Rate Between 28 Species of Herbaceous Angiosperms. Part II: Modeling the Determinants of RGR".

Shipley, Bill and M. J. Lechowicz (in review; submitted to *Ecology*), "Variation in the Functional Coordination of Leaf Morphology and Gas Exchange in 40 Wetland Species".

Spearman, Charles (1904), "General Intelligence Objectively Determined and Measured", *American Journal of Psychology* 10: 151–293.

Spirtes, Peter (1993), "Directed Cyclic Graphical Representations of Feedback Models", Proceedings of the 1995 Conference on Uncertainty in Artificial Intelligence. San Mateo: Morgan Kaufman, 491–498.

Spirtes, Peter, Clark Glymour, and Richard Scheines (1993), *Causation, Prediction and Search.* New York: Springer-Verlag.

Spirtes, Peter, Christopher Meek, and Thomas Richardson (1995), "Causal Inference in the Presence of Latent Variables and Selection Bias", Proceedings of the 1995 Conference on Uncertainty and Artificial Intelligence. San Mareo: Morgan Kaufman, 499–506.

Thurstone, Louis L. (1947), *Multiple-factor Analysis; A Development and Expansion of the Vectors of the Mind.* Chicago: University of Chicago Press.