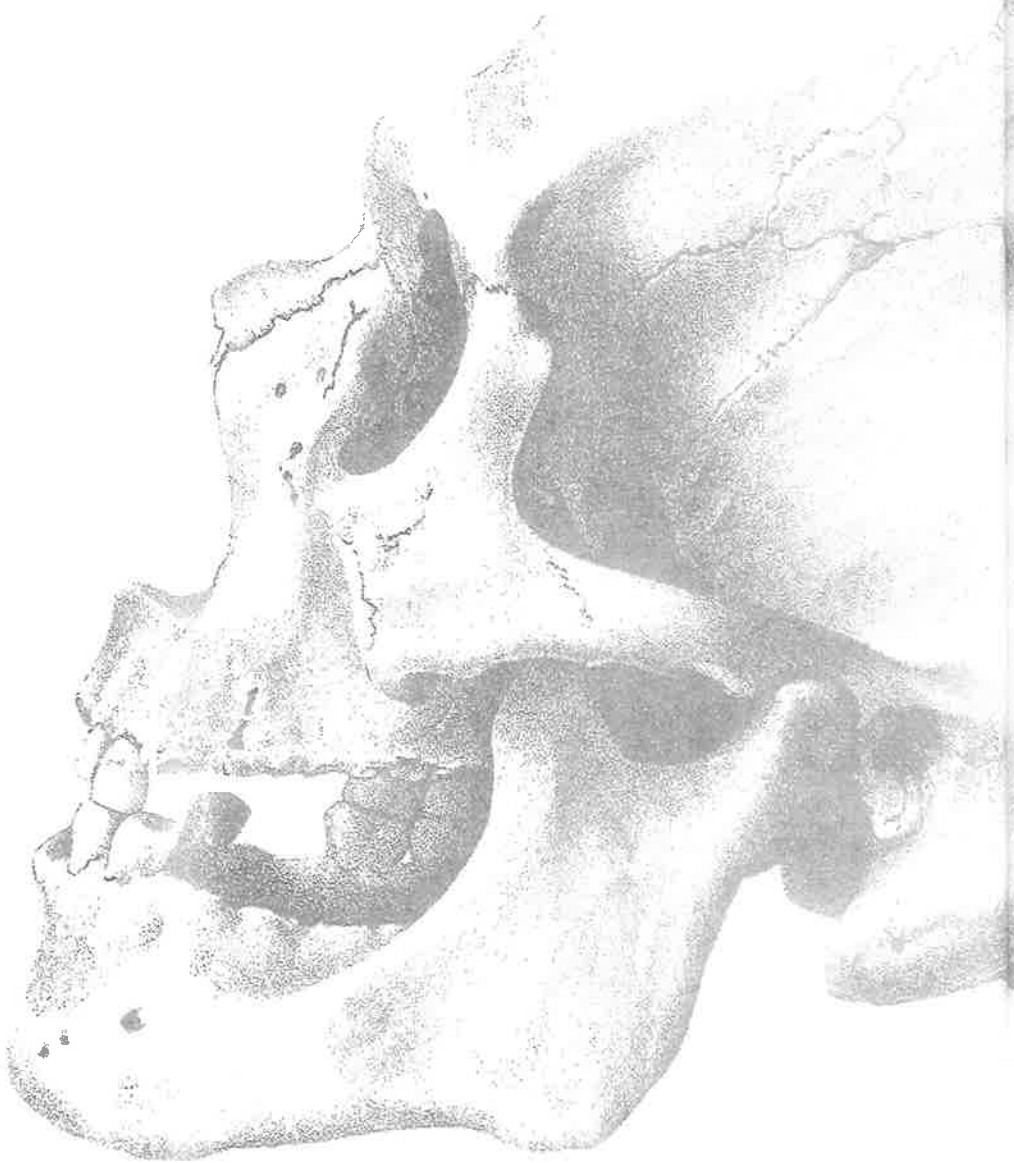


REVISED AND EXPANDED

# THE Mismeasure of Man

*BY* STEPHEN JAY GOULD

*If the misery of our poor be caused not by the laws of nature, but by our institutions, great is our sin.* —CHARLES DARWIN, *Voyage of the Beagle*

W • W • NORTON & COMPANY • *NEW YORK* • *LONDON*

SIX

# The Real Error of Cyril Burt

## Factor Analysis and the Reification of Intelligence

> It has been the signal merit of the English school of psychology, from Sir Francis Galton onwards, that it has, by this very device of mathematical analysis, transformed the mental test from a discredited dodge of the charlatan into a recognized instrument of scientific precision.
>
> —Cyril Burt, 1921, p. 130

### The case of Sir Cyril Burt

If I had any desire to lead a life of indolent ease, I would wish to be an identical twin, separated at birth from my brother and raised in a different social class. We could hire ourselves out to a host of social scientists and practically name our fee. For we would be exceedingly rare representatives of the only really adequate natural experiment for separating genetic from environmental effects in humans—genetically identical individuals raised in disparate environments.

Studies of identical twins raised apart should therefore hold pride of place in literature on the inheritance of IQ. And so it would be but for one problem—the extreme rarity of the animal itself. Few investigators have been able to rustle up more than twenty pairs of twins. Yet, amidst this paltriness, one study seemed to stand out: that of Sir Cyril Burt (1883–1971). Sir Cyril, doyen of mental testers, had pursued two sequential careers that gained him a preeminent role in directing both theory and practice in his field of educational psychology. For twenty years he was the official psychologist of the London County Council, responsible for the



administration and interpretation of mental tests in London's schools. He then succeeded Charles Spearman as professor in the most influential chair of psychology in Britain: University College, London (1932–1950). During his long retirement, Sir Cyril published several papers that buttressed the hereditarian claim by citing very high correlation between IQ scores of identical twins raised apart. Burt's study stood out among all others because he had found fifty-three pairs, more than twice the total of any previous attempt. It is scarcely surprising that Arthur Jensen used Sir Cyril's figures as the most important datum in his notorious article (1969) on supposedly inherited and ineradicable differences in intelligence between whites and blacks in America.

The story of Burt's undoing is now more than a twice-told tale. Princeton psychologist Leon Kamin first noted that, while Burt had increased his sample of twins from fewer than twenty to more than fifty in a series of publications, the average correlation between pairs for IQ remained unchanged to the third decimal place—a statistical situation so unlikely that it matches our vernacular definition of impossible. Then, in 1976, Oliver Gillie, medical correspondent of the London *Sunday Times*, elevated the charge from inexcusable carelessness to conscious fakery. Gillie discovered, among many other things, that Burt's two "collaborators," a Margaret Howard and a J. Conway, the women who supposedly collected and processed his data, either never existed at all, or at least could not have been in contact with Burt while he wrote the papers bearing their names. These charges led to further reassessments of Burt's "evidence" for his rigid hereditarian position. Indeed, other crucial studies were equally fraudulent, particularly his IQ correlations between close relatives (suspiciously too good to be true and apparently constructed from ideal statistical distributions, rather than measured in nature—Dorfman, 1978), and his data for declining levels of intelligence in Britain.

Burt's supporters tended at first to view the charges as a thinly veiled leftist plot to undo the hereditarian position by rhetoric. H. J. Eysenck wrote to Burt's sister: "I think the whole affair is just a determined effort on the part of some very left-wing environmentalists determined to play a political game with scientific facts. I am sure the future will uphold the honor and integrity of Sir Cyril without any question." Arthur Jensen, who had called Burt a

"born nobleman" and "one of the world's great psychologists," had to conclude that the data on identical twins could not be trusted, though he attributed their inaccuracy to carelessness alone.

I think that the splendid "official" biography of Burt recently published by L. S. Hearnshaw (1979) has resolved the issue so far as the data permit (Hearnshaw was commissioned to write his book by Burt's sister before any charges had been leveled). Hearnshaw, who began as an unqualified admirer of Burt and who tends to share his intellectual attitudes, eventually concluded that all allegations are true, and worse. And yet, Hearnshaw has convinced me that the very enormity and bizarreness of Burt's fakery forces us to view it not as the "rational" program of a devious person trying to salvage his hereditarian dogma when he knew the game was up (my original suspicion, I confess), but as the actions of a sick and tortured man. (All this, of course, does not touch the deeper issue of why such patently manufactured data went unchallenged for so long, and what this will to believe implies about the basis of our hereditarian presuppositions.)

Hearnshaw believes that Burt began his fabrications in the early 1940s, and that his earlier work was honest, though marred by rigid a priori conviction and often inexcusably sloppy and superficial, even by the standards of his own time. Burt's world began to collapse during the war, partly by his own doing to be sure. His research data perished in the blitz of London; his marriage failed; he was excluded from his own department when he refused to retire gracefully at the mandatory age and attempted to retain control; he was removed as editor of the journal he had founded, again after declining to cede control at the specified time he himself had set; his hereditarian dogma no longer matched the spirit of an age that had just witnessed the holocaust. In addition, Burt apparently suffered from Ménières disease, a disorder of the organs of balance, with frequent and negative consequences for personality as well.

Hearnshaw cites four instances of fraud in Burt's later career. Three I have already mentioned (fabrication of data on identical twins, kinship correlations in IQ, and declining levels of intelligence in Britain). The fourth is, in many ways, the most bizarre tale of all because Burt's claim was so absurd and his actions so patent and easy to uncover. It could not have been the act of a



rational man. Burt attempted to commit an act of intellectual parricide by declaring himself, rather than his predecessor and mentor Charles Spearman, as the father of a technique called "factor analysis" in psychology. Spearman had essentially invented the technique in a celebrated paper of 1904. Burt never challenged this priority—in fact he constantly affirmed it—while Spearman held the chair that Burt would later occupy at University College. Indeed, in his famous book on factor analysis (1940), Burt states that "Spearman's preeminence is acknowledged by every factorist" (1940, p. x).

Burt's first attempt to rewrite history occurred while Spearman was still alive, and it elicited a sharp rejoinder from the occupant emeritus of Burt's chair. Burt withdrew immediately and wrote a letter to Spearman that may be unmatched for deference and obsequiousness: "Surely you have a prior claim here. . . . I have been wondering where precisely I have gone astray. Would it be simplest for me to number my statements, then like my schoolmaster of old you can put a cross against the points where your pupil has blundered, and a tick where your view is correctly interpreted."

But when Spearman died, Burt launched a campaign that "became increasingly unrestrained, obsessive and extravagant" (Hearnshaw, 1979) throughout the rest of his life. Hearnshaw notes (1979, pp. 286–287): "The whisperings against Spearman that were just audible in the late 1930's swelled into a strident campaign of belittlement, which grew until Burt arrogated to himself the whole of Spearman's fame. Indeed, Burt seemed to be becoming increasingly obsessed with questions of priority, and increasingly touchy and egotistical." Burt's false story was simple enough: Karl Pearson had invented the technique of factor analysis (or something close enough to it) in 1901, three years before Spearman's paper. But Pearson had not applied it to psychological problems. Burt recognized its implications and brought the technique into studies of mental testing, making several crucial modifications and improvements along the way. The line, therefore, runs from Pearson to Burt. Spearman's 1904 paper was merely a diversion.

Burt told his story again and again. He even told it through one of his many aliases in a letter he wrote to his own journal and signed Jacques Lafitte, an unknown French psychologist. With the exception of Voltaire and Binet, M. Lafitte cited only English

sources and stated: "Surely the first formal and adequate statement was Karl Pearson's demonstration of the method of principal axes in 1901." Yet anyone could have exposed Burt's story as fiction after an hour's effort—for Burt never cited Pearson's paper in any of his work before 1947, while all his earlier studies of factor analysis grant credit to Spearman and clearly display the derivative character of Burt's methods.

Factor analysis must have been very important if Burt chose to center his quest for fame upon a rewrite of history that would make him its inventor. Yet, despite all the popular literature on IQ in the history of mental testing, virtually nothing has been written (outside professional circles) on the role, impact, and meaning of factor analysis. I suspect that the main reason for this neglect lies in the abstrusely mathematical nature of the technique. IQ, a linear scale first established as a rough, empirical measure, is easy to understand. Factor analysis, rooted in abstract statistical theory and based on the attempt to discover "underlying" structure in large matrices of data, is, to put it bluntly, a bitch. Yet this inattention to factor analysis is a serious omission for anyone who wishes to understand the history of mental testing in our century, and its continuing rationale today. For as Burt correctly noted (1914, p. 36), the history of mental testing contains two major and related strands: age-scale methods (Binet IQ testing), and correlational methods (factor analysis). Moreover, as Spearman continually stressed throughout his career, the theoretical justification for using a unilinear scale of IQ resides in factor analysis itself. Burt may have been perverse in his campaign, but he was right in his chosen tactic—a permanent and exalted niche in the pantheon of psychology lies reserved for the man who developed factor analysis.

I began my career in biology by using factor analysis to study the evolution of a group of fossil reptiles. I was taught the technique as though it had developed from first principles using pure logic. In fact, virtually all its procedures arose as justifications for particular theories of intelligence. Factor analysis, despite its status as pure deductive mathematics, was invented in a social context, and for definite reasons. And, though its mathematical basis is unassailable, its persistent use as a device for learning about the physical structure of intellect has been mired in deep conceptual errors from the start. The principal error, in fact, has involved a



major theme of this book: reification—in this case, the notion that such a nebulous, socially defined concept as intelligence might be identified as a "thing" with a locus in the brain and a definite degree of heritability—and that it might be measured as a single number, thus permitting a unilinear ranking of people according to the amount of it they possess. By identifying a mathematical factor axis with a concept of "general intelligence," Spearman and Burt provided a theoretical justification for the unilinear scale that Binet had proposed as a rough empirical guide.

The intense debate about Cyril Burt's work has focused exclusively on the fakery of his late career. This perspective has clouded Sir Cyril's greater influence as the most powerful mental tester committed to a factor-analytic model of intelligence as a real and unitary "thing." Burt's commitment was rooted in the error of reification. Later fakery was the afterthought of a defeated man; his earlier, "honest" error has reverberated throughout our century and has affected millions of lives.

## Correlation, cause, and factor analysis

### *Correlation and cause*

The spirit of Plato dies hard. We have been unable to escape the philosophical tradition that what we can see and measure in the world is merely the superficial and imperfect representation of an underlying reality. Much of the fascination of statistics lies embedded in our gut feeling—and never trust a gut feeling—that abstract measures summarizing large tables of data must express something more real and fundamental than the data themselves. (Much professional training in statistics involves a conscious effort to counteract this gut feeling.) The technique of *correlation* has been particularly subject to such misuse because it seems to provide a path for inferences about causality (and indeed it does, sometimes—but only sometimes).

Correlation assesses the tendency of one measure to vary in concert with another. As a child grows, for example, both its arms and legs get longer; this joint tendency to change in the same direction is called a *positive correlation*. Not all parts of the body display such positive correlations during growth. Teeth, for example, do not grow after they erupt. The relationship between first incisor

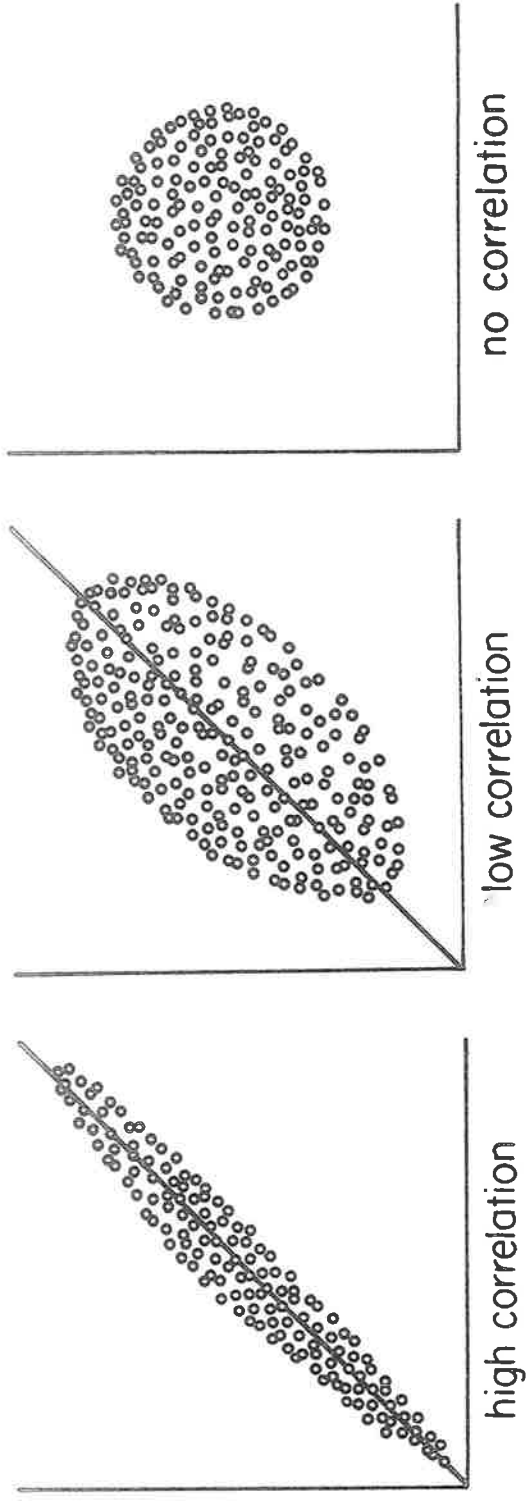
length and leg length from, say, age ten to adulthood would represent *zero correlation*—legs would get longer while teeth changed not at all. Other correlations can be negative—one measure increases while the other decreases. We begin to lose neurons at a distressingly early age, and they are not replaced. Thus, the relationship between leg length and number of neurons after mid-childhood represents *negative correlation*—leg length increases while number of neurons decreases. Notice that I have said nothing about causality. We do not know why these correlations exist or do not exist, only that they are present or not present.

The standard measure of correlation is called Pearson's product moment correlation coefficient or, for short, simply the correlation coefficient, symbolized as $r$. The correlation coefficient ranges from +1 for perfect positive correlation, to 0 for no correlation, to −1 for perfect negative correlation.*

In rough terms, $r$ measures the shape of an ellipse of plotted points (see Fig. 6.1). Very skinny ellipses represent high correlations—the skinniest of all, a straight line, reflects an $r$ of 1.0. Fat ellipses represent lower correlations, and the fattest of all, a circle, reflects zero correlation (increase in one measure permits no prediction about whether the other will increase, decrease, or remain the same).

The correlation coefficient, though easily calculated, has been plagued by errors of interpretation. These can be illustrated by example. Suppose that I plot arm length vs. leg length during the growth of a child. I will obtain a high correlation with two interesting implications. First, I have achieved *simplification*. I began with two dimensions (leg and arm length), which I have now, effectively, reduced to one. Since the correlation is so strong, we may say that the line itself (a single dimension) represents nearly all the information originally supplied as two dimensions. Secondly, I can, in this case, make a reasonable inference about the *cause* of this reduc-

*Pearson's $r$ is not an appropriate measure for all kinds of correlation, for it assesses only what statisticians call the intensity of linear relationship between two measures—the tendency for all points to fall on a single straight line. Other relationships of strict dependence will not achieve a value of 1.0 for $r$. If, for example, each increase of 2 units in one variable were matched by an increase in $2^2$ units in the other variable, $r$ would be less than 1.0, even though the two variables might be perfectly "correlated" in the vernacular sense. Their plot would be a parabola, not a straight line, and Pearson's $r$ measures the intensity of linear relationship.




*6 • 1* Strength of correlation as a function of the shape of an ellipse of points. The more elongate the ellipse, the higher the correlation.

tion to one dimension. Arm and leg length are tightly correlated because they are both partial measures of an underlying biological phenomenon, namely growth itself.

Yet, lest anyone become too hopeful that correlation represents a magic method for the unambiguous identification of cause, consider the relationship between my age and the price of gasoline during the past ten years. The correlation is nearly perfect, but no one would suggest any assignment of cause. The fact of correlation implies nothing about cause. It is not even true that intense correlations are more likely to represent cause than weak ones, for the correlation of my age with the price of gasoline is nearly 1.0. I spoke of cause for arm and leg lengths not because their correlation was high, but because I know something about the biology of the situation. The inference of cause must come from somewhere else, not from the simple fact of correlation—though an unexpected correlation may lead us to search for causes so long as we remember that we may not find them. The vast majority of correlations in our world are, without doubt, noncausal. Anything that has been increasing steadily during the past few years will be strongly correlated with the distance between the earth and Halley's comet (which has also been increasing of late)—but even the most dedicated astrologer would not discern causality in most of these relationships. The invalid assumption that correlation implies cause is probably among the two or three most serious and common errors of human reasoning.

Few people would be fooled by such a reductio ad absurdum as the age-gas correlation. But consider an intermediate case. I am given a table of data showing how far twenty children can hit and throw a baseball. I graph these data and calculate a high $r$. Most people, I think, would share my intuition that this is not a meaningless correlation; yet in the absence of further information, the correlation itself teaches me nothing about underlying causes. For I can suggest at least three different and reasonable causal interpretations for the correlation (and the true reason is probably some combination of them):

1. The children are simply of different ages, and older children can hit and throw farther.

2. The differences represent variation in practice and training. Some children are Little League stars and can tell you the year that



Rogers Hornsby hit .424 (1924—I was a bratty little kid like that); others know Billy Martin only as a figure in Lite beer commercials.

3. The differences represent disparities in native ability that cannot be erased even by intense training. (The situation would be even more complex if the sample included both boys and girls of conventional upbringing. The correlation might then be attributed primarily to a fourth cause—sexual differences; and we would have to worry, in addition, about the cause of the sexual difference: training, inborn constitution, or some combination of nature and nurture).

In summary, most correlations are noncausal; when correlations are causal, the fact and strength of the correlation rarely specifies the nature of the cause.

*Correlation in more than two dimensions*

These two-dimensional examples are easy to grasp (however difficult they are to interpret). But what of correlations among more than two measures? A body is composed of many parts, not just arms and legs, and we may want to know how several measures interact during growth. Suppose, for simplicity, that we add just one more measure, head length, to make a three-dimensional system. We may now depict the correlation structure among the three measures in two ways:

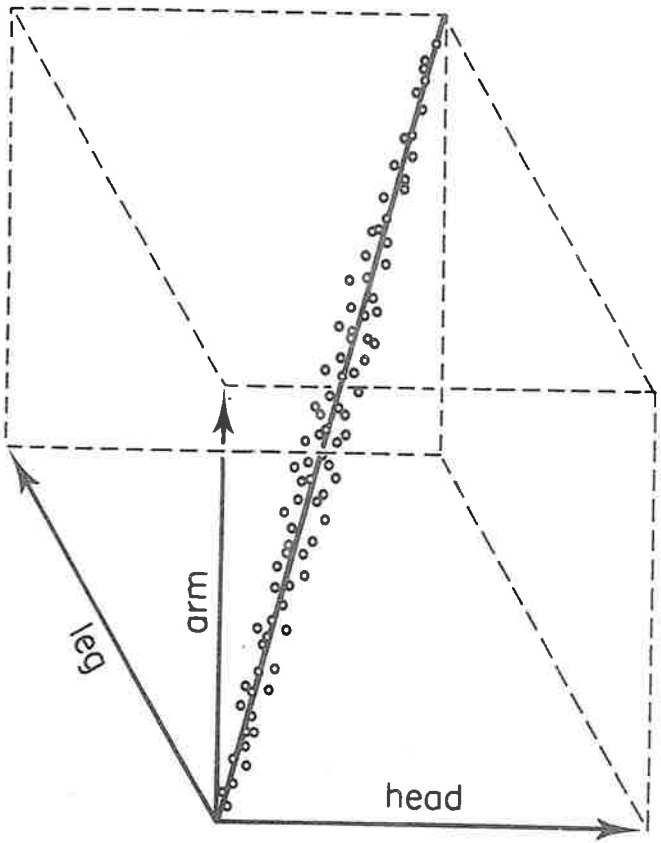
1. We may gather all correlation coefficients between pairs of measures into a single table, or *matrix* of correlation coefficients (Fig. 6.2). The line from upper left to lower right records the necessarily perfect correlation of each variable with itself. It is called the principal diagonal, and all correlations along it are 1.0. The matrix is symmetrical around the principal diagonal, since the correlation of measure 1 with measure 2 is the same as the correlation of 2 with 1. Thus, the three values either above or below the principal diagonal are the correlations we seek: arm with leg, arm with head, and leg with head.

2. We may plot the points for all individuals onto a three-dimensional graph (Fig. 6.3). Since the correlations are all positive, the points are oriented as an ellipsoid (or football). (In two dimensions, they formed an ellipse.) A line running along the major axis of the football expresses the strong positive correlations between all measures.

| | arm | leg | head |
|---|---|---|---|
| arm | 1.0 | 0.91 | 0.72 |
| leg | 0.91 | 1.0 | 0.63 |
| head | 0.72 | 0.63 | 1.0 |

*6•2* A correlation matrix for three measurements.

*6•3* A three-dimensional graph showing the correlations for three measurements.






We can grasp the three-dimensional case, both mentally and pictorially. But what about 20 dimensions, or 100? If we measured 100 parts of a growing body, our correlation matrix would contain 10,000 items. To plot this information, we would have to work in a 100-dimensional space, with 100 mutually perpendicular axes representing the original measures. Although these 100 axes present no mathematical problem (they form, in technical terms, a hyperspace), we cannot plot them in our three-dimensional Euclidian world.

These 100 measures of a growing body probably do not represent 100 different biological phenomena. Just as most of the information in our three-dimensional example could be resolved into a single dimension (the long axis of the football), so might our 100 measures be simplified into fewer dimensions. We will lose some information in the process to be sure—as we did when we collapsed the long and skinny football, still a three-dimensional structure, into the single line representing its long axis. But we may be willing to accept this loss in exchange for simplification and for the possibility of interpreting the dimensions that we do retain in biological terms.

### *Factor analysis and its goals*

With this example, we come to the heart of what *factor analysis* attempts to do. Factor analysis is a mathematical technique for reducing a complex system of correlations into fewer dimensions. It works, literally, by factoring a matrix, usually a matrix of correlation coefficients. (Remember the high-school algebra exercise called "factoring," where you simplified horrendous expressions by removing common multipliers of all terms?) Geometrically, the process of factoring amounts to placing axes through a football of points. In the 100-dimensional case, we are not likely to recover enough information on a single line down the hyperfootball's long axis—a line called the *first principal component.* We will need additional axes. By convention, we represent the second dimension by a line *perpendicular* to the first principal component. This second axis, or *second principal component,* is defined as the line that resolves more of the remaining variation than any other line that could be drawn perpendicular to the first principal component. If, for example, the hyperfootball were squashed flat like a flounder, the

first principal component would run through the middle, from head to tail, and the second also through the middle, but from side to side. Subsequent lines would be perpendicular to all previous axes, and would resolve a steadily decreasing amount of remaining variation. We might find that five principal components resolve almost all the variation in our hyperfootball—that is, the hyperfootball drawn in 5 dimensions looks sufficiently like the original to satisfy us, just as a pizza or a flounder drawn in two dimensions may express all the information we need, even though both original objects contain three dimensions. If we elect to stop at 5 dimensions, we may achieve a considerable simplification at the acceptable price of minimal loss of information. We can grasp the 5 dimensions conceptually; we may even be able to interpret them biologically.

Since factoring is performed on a correlation matrix, I shall use a geometrical representation of the correlation coefficients themselves in order to explain better how the technique operates. The original measures may be represented as vectors of unit length,*

*(Footnote for aficionados—others may safely skip.) Here, I am technically discussing a procedure called "principal components analysis," not quite the same thing as factor analysis. In principal components analysis, we preserve all information in the original measures and fit new axes to them by the same criterion used in factor analysis in principal components orientation—that is, the first axis explains more data than any other axis could and subsequent axes lie at right angles to all other axes and encompass steadily decreasing amounts of information. In true factor analysis, we decide beforehand (by various procedures) not to include all information on our factor axes. But the two techniques—true factor analysis in principal components orientation and principal components analysis—play the same conceptual role and differ only in mode of calculation. In both, the first axis (Spearman's *g* for intelligence tests) is a "best fit" dimension that resolves more information in a set of vectors than any other axis could.

During the past decade or so, semantic confusion has spread in statistical circles through a tendency to restrict the term "factor analysis" only to the rotations of axes usually performed after the calculation of principal components, and to extend the term "principal components analysis" both to true principal components analysis (all information retained) and to factor analysis done in principal components orientation (reduced dimensionality and loss of information). This shift in definition is completely out of keeping with the history of the subject and terms. Spearman, Burt, and hosts of other psychometricians worked for decades in this area before Thurstone and others invented axial rotations. They performed all their calculations in the principal components orientation, and they called themselves "factor analysts." I continue, therefore, to use the term "factor analysis" in its original sense to include any orientation of axes—principal components or rotated, orthogonal or oblique.

I will also use a common, if somewhat sloppy, shorthand in discussing what



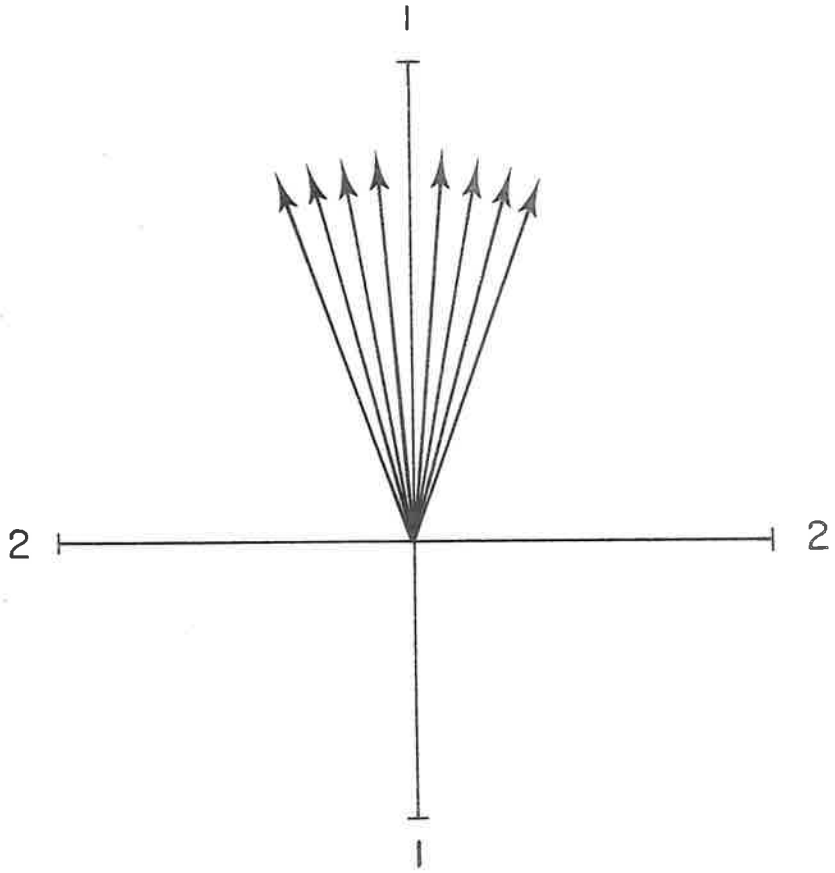
radiating from a common point. If two measures are highly correlated, their vectors lie close to each other. The cosine of the angle between any two vectors records the correlation coefficient between them. If two vectors overlap, their correlation is perfect, or 1.0; the cosine of 0° is 1.0. If two vectors lie at right angles, they are completely independent, with a correlation of zero; the cosine of 90° is zero. If two vectors point in opposite directions, their correlation is perfectly negative, or −1.0; the cosine of 180° is −1.0. A matrix of high positive correlation coefficients will be represented by a cluster of vectors, each separated from each other vector by a small acute angle (Fig. 6.4). When we factor such a cluster into fewer dimensions by computing principal components, we choose as our first component the axis of maximal resolving power, a kind of grand average among all vectors. We assess resolving power by projecting each vector onto the axis. This is done by drawing a line from the tip of the vector to the axis, perpendicular to the axis. The ratio of projected length on the axis to the actual length of the vector itself measures the percentage of a vector's information resolved by the axis. (This is difficult to express verbally, but I think that Figure 6.5 will dispel confusion.) If a vector lies near the axis, it is highly resolved and the axis encompasses most of its information. As a vector moves away from the axis toward a maximal separation of 90°, the axis resolves less and less of it.

We position the first principal component (or axis) so that it resolves more information among all the vectors than any other axis could. For our matrix of high positive correlation coefficients, represented by a set of tightly clustered vectors, the first principal component runs through the middle of the set (Fig. 6.4). The second principal component lies at right angles to the first and resolves a maximal amount of remaining information. But if the first component has already resolved most of the information in all the vectors, then the second and subsequent principal axes can only deal with the small amount of information that remains (Fig. 6.4).

factor axes do. Technically, factor axes resolve variance in original measures. I will, as is often done, speak of them as "explaining" or "resolving" information—as they do in the vernacular (though not in the technical) sense of information. That is, when the vector of an original variable projects strongly on a set of factor axes, little of its variance lies unresolved in higher dimensions outside the system of factor axes.

Such systems of high positive correlation are found frequently in nature. In my own first study in factor analysis, for example, I considered fourteen measurements on the bones of twenty-two species of pelycosaurian reptiles (the fossil beasts with the sails on their backs, often confused with dinosaurs, but actually the ancestors of mammals). My first principal component resolved 97.1 per-

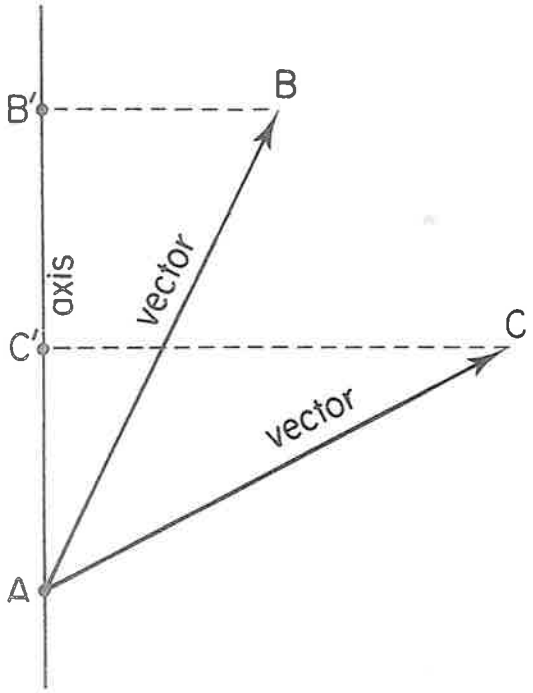
*6•4* Geometric representation of correlations among eight tests when all correlation coefficients are high and positive. The first principal component, labeled 1, lies close to all the vectors, while the second principal component, labeled 2, lies at right angles to the first and does not explain much information in the vectors.






cent of the information in all fourteen vectors, leaving only 2.9 percent for subsequent axes. My fourteen vectors formed an extremely tight swarm (all practically overlapping); the first axis went through the middle of the swarm. My pelycosaurs ranged in body length from less than two to more than eleven feet. They all look pretty much alike, and big animals have larger measures for all fourteen bones. All correlation coefficients of bones with other bones are very high; in fact, the lowest is still a whopping 0.912.

*6•5* Computing the amount of information in a vector explained by an axis. Draw a line from the tip of the vector to the axis, perpendicular to the axis. The amount of information resolved by the axis is the ratio of the projected length on the axis to the true length of the vector. If a vector lies close to the axis, then this ratio is high and most of the information in the vector is resolved by the axis. Vector AB lies close to the axis and the ratio of the projection AB′ to the vector itself, AB, is high. Vector AC lies far from the axis and the ratio of its projected length AC′ to the vector itself, AC, is low.



Scarcely surprising. After all, large animals have large bones, and small animals small bones. I can interpret my first principal component as an abstracted size factor, thus reducing (with minimal loss of information) my fourteen original measurements into a single dimension interpreted as increasing body size. In this case, factor analysis has achieved both *simplification* by reduction of dimensions (from fourteen to effectively one), and *explanation* by reasonable biological interpretation of the first axis as a size factor.

But—and here comes an enormous but—before we rejoice and extol factor analysis as a panacea for understanding complex systems of correlation, we should recognize that it is subject to the same cautions and objections previously examined for the correlation coefficients themselves. I consider two major problems in the following sections.

### *The error of reification*

The first principal component is a mathematical abstraction that can be calculated for any matrix of correlation coefficients; it is not a "thing" with physical reality. Factorists have often fallen prey to a temptation for *reification*—for awarding *physical meaning* to all strong principal components. Sometimes this is justified; I believe that I can make a good case for interpreting my first pelycosaurian axis as a size factor. But such a claim can never arise from the mathematics alone, only from additional knowledge of the physical nature of the measures themselves. For nonsensical systems of correlation have principal components as well, and they may resolve more information than meaningful components do in other systems. A factor analysis for a five-by-five correlation matrix of my age, the population of Mexico, the price of swiss cheese, my pet turtle's weight, and the average distance between galaxies during the past ten years will yield a strong first principal component. This component—since all the correlations are so strongly positive—will probably resolve as high a percentage of information as the first axis in my study of pelycosaurs. It will also have no enlightening physical meaning whatever.

In studies of intelligence, factor analysis has been applied to matrices of correlation among mental tests. Ten tests may, for example, be given to each of one hundred people. Each meaningful entry in the ten-by-ten correlation matrix is a correlation coefficient between scores on two tests taken by each of the one hundred persons. We have known since the early days of mental testing—and it should surprise no one—that most of these correlation coefficients are positive: that is, people who score highly on one kind of test tend, on average, to score highly on others as well. Most correlation matrices for mental tests contain a preponderance of positive entries. This basic observation served as the starting point for factor analysis. Charles Spearman virtually invented the technique in 1904 as a device for inferring causes from correlation matrices of mental tests.



Since most correlation coefficients in the matrix are positive, factor analysis must yield a reasonably strong first principal component. Spearman calculated such a component indirectly in 1904 and then made the cardinal invalid inference that has plagued factor analysis ever since. He reified it as an "entity" and tried to give it an unambiguous causal interpretation. He called it *g*, or general intelligence, and imagined that he had identified a unitary quality underlying all cognitive mental activity—a quality that could be expressed as a single number and used to rank people on a unilinear scale of intellectual worth.

Spearman's *g*—the first principal component of the correlation matrix of mental tests—never attains the predominant role that a first component plays in many growth studies (as in my pelycosaurs). At best, *g* resolves 50 to 60 percent of all information in the matrix of tests. Correlations between tests are usually far weaker than correlations between two parts of a growing body. In most cases, the highest correlation in a matrix of tests does not come close to reaching the *lowest* value in my pelycosaur matrix—0.912.

Although *g* never matches the strength of a first principal component of some growth studies, I do not regard its fair resolving power as accidental. Causal reasons lie behind the positive correlations of most mental tests. But what reasons? We cannot infer the reasons from a strong first principal component any more than we can induce the cause of a single correlation coefficient from its magnitude. We cannot reify *g* as a "thing" unless we have convincing, independent information beyond the fact of correlation itself.

The situation for mental tests resembles the hypothetical case I presented earlier of correlation between throwing and hitting a baseball. The relationship is strong and we have a right to regard

it as nonaccidental. But we cannot infer the cause from the correlation, and the cause is certainly complex.

Spearman's *g* is particularly subject to ambiguity in interpretation, if only because the two most contradictory causal hypotheses are both fully consistent with it: 1) that it reflects an inherited level of mental acuity (some people do well on most tests because they are born smarter); or 2) that it records environmental advantages and deficits (some people do well on most tests because they are well schooled, grew up with enough to eat, books in the home, and loving parents). If the simple existence of *g* can be theoretically interpreted in either a purely hereditarian or purely environmentalist way, then its mere presence—even its reasonable strength—cannot justly lead to any reification at all. The temptation to reify is powerful. The idea that we have detected something "underlying" the externalities of a large set of correlation coefficients, something perhaps more real than the superficial measurements themselves, can be intoxicating. It is Plato's essence, the abstract, eternal reality underlying superficial appearances. But it is a temptation that we must resist, for it reflects an ancient prejudice of thought, not a truth of nature.

### *Rotation and the nonnecessity of principal components*

Another, more technical, argument clearly demonstrates why principal components cannot be automatically reified as causal entities. If principal components represented the only way to simplify a correlation matrix, then some special status for them might be legitimately sought. But they represent only one method among many for inserting axes into a multidimensional space. Principal components have a definite geometric arrangement, specified by the criterion used to construct them—that the first principal component shall resolve a maximal amount of information in a set of vectors and that subsequent components shall all be mutually perpendicular. But there is nothing sacrosanct about this criterion; vectors may be resolved into any set of axes placed within their space. Principal components provide insight in some cases, but other criteria are often more useful.

Consider the following situation, in which another scheme for placing axes might be preferred. In Figure 6.6 I show correlations between four mental tests, two of verbal and two of arithmetical



aptitude. Two "clusters" are evident, even though all tests are positively correlated. Suppose that we wish to identify these clusters by factor analysis. If we use principal components, we may not recognize them at all. The first principal component (Spearman's *g*) goes right up the middle, between the two clusters. It lies close to no vector and resolves an approximately equal amount of each, thereby masking the existence of verbal and arithmetic clusters. Is this component an entity? Does a "general intelligence" exist? Or is *g*, in this case, merely a meaningless average based on the invalid amalgamation of two types of information?

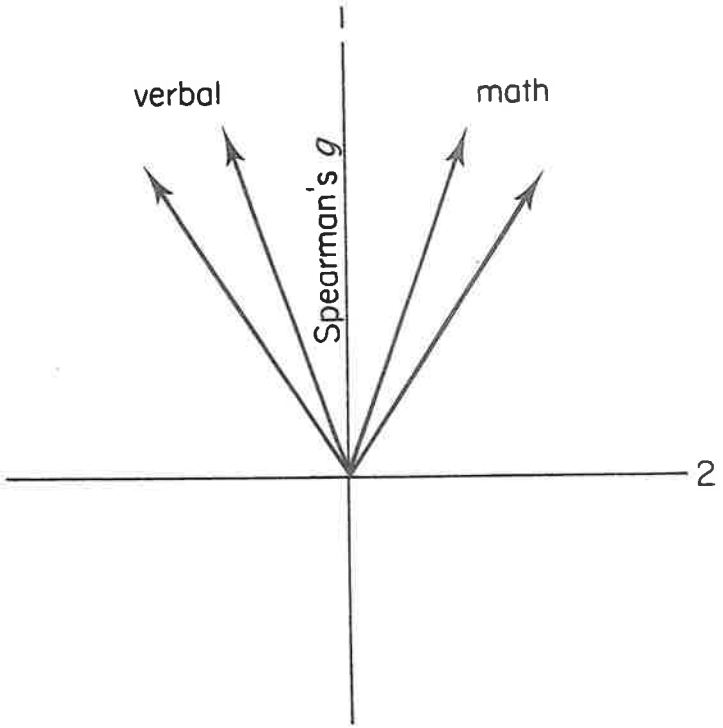
We may pick up verbal and arithmetic clusters on the second principal component (called a "bipolar factor" because some projections upon it will be positive and others negative when vectors lie on both sides of the first principal component). In this case, verbal tests project on the negative side of the second component, and arithmetic tests on the positive side. But we may fail to detect these clusters altogether if the first principal component dominates all vectors. For projections on the second component will then be small, and the pattern can easily be lost (see Fig. 6.6).

During the 1930s factorists developed methods to treat this dilemma and to recognize clusters of vectors that principal components often obscured. They did this by rotating factor axes from the principal components orientation to new positions. The rotations, established by several criteria, had as their common aim the positioning of axes near clusters. In Figure 6.7, for example, we use the criterion: place axes near vectors occupying extreme or outlying positions in the total set. If we now resolve all vectors into these rotated axes, we detect the clusters easily; for arithmetic tests project high on rotated axis 1 and low on rotated axis 2, while verbal tests project high on 2 and low on 1. Moreover, *g has disappeared.* We no longer find a "general factor" of intelligence, nothing that can be reified as a single number expresssing overall ability. Yet we have lost no information. The two rotated axes resolve as much information in the four vectors as did the two principal components. They simply distribute the same information differently upon the resolving axes. How can we argue that *g* has any claim to reified status as an entity if it represents but one of numerous possible ways to position axes within a set of vectors?

In short, factor analysis simplifies large sets of data by reducing

dimensionality and trading some loss of information for the recognition of ordered structure in fewer dimensions. As a tool for simplification, it has proved its great value in many disciplines. But many factorists have gone beyond simplification, and tried to define factors as causal entities. This error of reification has plagued the technique since its inception. It was "present at the creation" since Spearman invented factor analysis to study the correlation matrix of mental tests and then reified his principal component as *g* or innate, general intelligence. Factor analysis may help us to understand causes by directing us to information beyond the

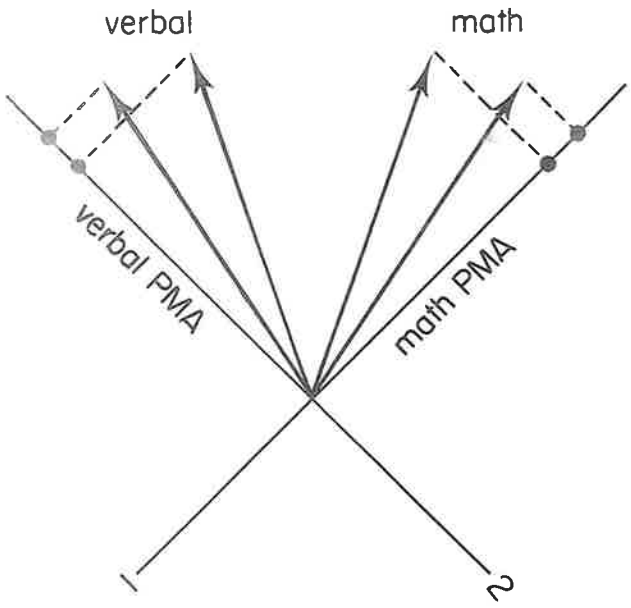
*6•6* A principal components analysis of four mental tests. All correlations are high and the first principal component, Spearman's *g*, expresses the overall correlation. But the group factors for verbal and mathematical aptitude are not well resolved in this style of analysis.






mathematics of correlation. But factors, by themselves, are neither things nor causes; they are mathematical abstractions. Since the same set of vectors (see Figs. 6.6, 6.7) can be partitioned into *g* and a small residual axis, or into two axes of equal strength that identify verbal and arithmetical clusters and dispense with *g* entirely, we cannot claim that Spearman's "general intelligence" is an ineluctable entity necessarily underlying and causing the correlations among mental tests. Even if we choose to defend *g* as a nonaccidental result, neither its strength nor its geometric position can specify what it means in causal terms—if only because its features are equally consistent with extreme hereditarian and extreme environmentalist views of intelligence.

*6•7* Rotated factor axes for the same four mental tests depicted in Fig. 6.6. Axes are now placed near vectors lying at the periphery of the cluster. The group factors for verbal and mathematical aptitude are now well identified (see high projections on the axes indicated by dots), but *g* has disappeared.



## Charles Spearman and general intelligence

### *The two-factor theory*

Correlation coefficients are now about as ubiquitous and unsurprising as cockroaches in New York City. Even the cheapest pocket calculators produce correlation coefficients with the press of a button. However indispensable, they are taken for granted as automatic accouterments of any statistical analysis that deals with more than one measure. In such a context, we easily forget that they were once hailed as a breakthrough in research, as a new and exciting tool for discovering underlying structure in tables of raw measures. We can sense this excitement in reading early papers of the great American biologist and statistician Raymond Pearl (see Pearl, 1905 and 1906, and Pearl and Fuller, 1905). Pearl completed his doctorate at the turn of the century and then proceeded, like a happy boy with a gleaming new toy, to correlate everything in sight, from the lengths of earth worms vs. the number of their body segments (where he found no correlation and assumed that increasing length reflects larger, rather than more, segments), to size of the human head vs. intelligence (where he found a very small correlation, but attributed it to the indirect effect of better nutrition).

Charles Spearman, an eminent psychologist and fine statistician as well* began to study correlations between mental tests during these heady times. If two mental tests are given to a large number of people, Spearman noted, the correlation coefficient between them is nearly always positive. Spearman pondered this result and wondered what higher generality it implied. The positive correlations clearly indicated that each test did not measure an independent attribute of mental functioning. Some simpler structure lay behind the pervasive positive correlations; but what structure? Spearman imagined two alternatives. First, the positive correlations might reduce to a small set of independent attributes—the "faculties" of the phrenologists and other schools of early psychology. Perhaps the mind had separate "compartments" for arithmetic, verbal, and spatial aptitudes, for example. Spearman called such

*Spearman took a special interest in problems of correlation and invented a measure that probably ranks second in use to Pearson's $r$ as a measure of association between two variables—the so-called Spearman's rank-correlation coefficient.



theories of intelligence "oligarchic." Second, the positive correlations might reduce to a single, underlying general factor—a notion that Spearman called "monarchic." In either case, Spearman recognized that the underlying factors—be they few (oligarchic) or single (monarchic)—would not encompass all information in a matrix of positive correlation coefficients for a large number of mental tests. A "residual variance" would remain—information peculiar to each test and not related to any other. In other words, each test would have its "anarchic" component. Spearman called the residual variance of each test its *s*, or specific information. Thus, Spearman reasoned, a study of underlying structure might lead to a "two-factor theory" in which each test contained some specific information (its *s*) and also reflected the operation of a single, underlying factor, which Spearman called *g*, or general intelligence. Or each test might include its specific information and also record one or several among a set of independent, underlying faculties—a many-factor theory. If the simplest two-factor theory held, then all common attributes of intelligence would reduce to a single underlying entity—a true "general intelligence" that might be measured for each person and might afford an unambiguous criterion for ranking in terms of mental worth.

Charles Spearman developed factor analysis—still the most important technique in modern multivariate statistics—as a procedure for deciding between the two- vs. the many-factor theory by determining whether the common variance in a matrix of correlation coefficients could be reduced to a single "general" factor, or only to several independent "group" factors. He found but a single "intelligence," opted for the two-factor theory, and, in 1904, published a paper that later won this assessment from a man who opposed its major result: "No single event in the history of mental testing has proved to be of such momentous importance as Spearman's proposal of his famous two-factor theory" (Guilford, 1936, p. 155). Elated, and with characteristic immodesty, Spearman gave his 1904 paper a heroic title: "General Intelligence Objectively Measured and Determined." Ten years later (1914, p. 237), he exulted: "The future of research into the inheritance of ability must center on the theory of 'two factors.' This alone seems capable of reducing the bewildering chaos of facts to a perspicuous orderliness. By its means, the problems are rendered clear; in many

respects, their answers are already foreshadowed; and everywhere, they are rendered susceptible of eventual decisive solution."

*The method of tetrad differences*

In his original work, Spearman did not use the method of principal components described on pp. 275–278. Instead, he developed a simpler, though tedious, procedure better suited for a precomputer age when all calculations had to be performed by hand.* He computed the entire matrix of correlation coefficients between all pairs of tests, took all possible groupings of four measures and computed for each a number that he called the "tetrad difference." Consider the following example as an attempt to define the tetrad difference and to explain how Spearman used it to test whether the common variance of his matrix could be reduced to a single general factor, or only to several group factors.

Suppose that we wish to compute the tetrad difference for four measures taken on a series of mice ranging in age from babies to adults—leg length, leg width, tail length, and tail width. We compute all correlation coefficients between pairs of variables and find, unsurprisingly, that all are positive—as mice grow, their parts get larger. But we would like to know whether the common variance in the positive correlations all reflects a single general factor—growth itself—or whether two separate components of growth must be identified—in this case, a leg factor and a tail factor, or a length factor and a width factor. Spearman gives the following formula for the tetrad difference

$$r_{13} \times r_{24} - r_{23} \times r_{14}$$

where $r$ is the correlation coefficient and the two subscripts represent the two measures being correlated (in this case, 1 is leg length, 2 is leg width, 3 is tail length and 4 is tail width—so that $r_{13}$ is the correlation coefficient between the first and the third measure, or between leg length and tail length). In our example, the tetrad difference is

(leg length and tail length) × (leg width and tail width) −
(leg width and tail length) × (leg length and tail width)

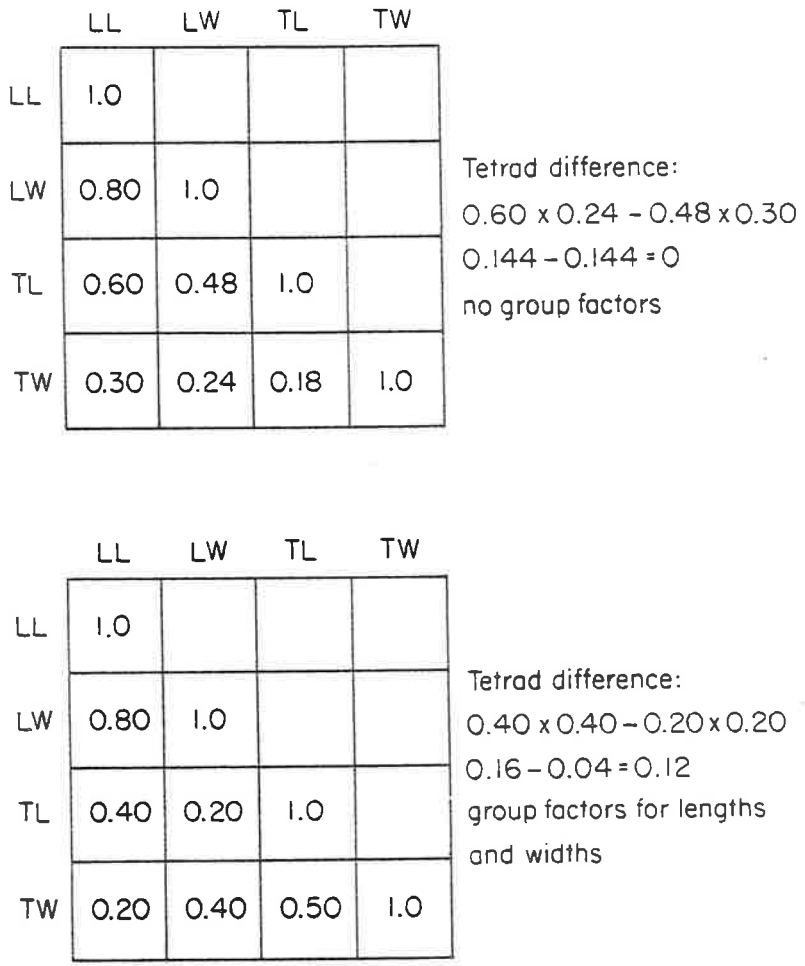
*The $g$ calculated by the tetrad formula is conceptually equivalent and mathematically almost equivalent to the first principal component described on pp. 275–278 and used in modern factor analysis.



Spearman argued that tetrad differences of zero imply the existence of a single general factor while either positive or negative values indicate that group factors must be recognized. Suppose, for example, that group factors for general body length and general body width govern the growth of mice. In this case, we would get a high positive value for the tetrad difference because the correlation coefficients of a length with another length or a width with another width would tend to be higher than correlation coefficients of a width with a length. (Note that the left-hand side of the tetrad equation includes only lengths with lengths or widths with widths, while the right-hand side includes only lengths with widths.) But if only a single, general growth factor regulates the size of mice, then lengths with widths should show as high a correlation as lengths with lengths or widths with widths—and the tetrad difference should be zero. Fig. 6.8 shows a hypothetical correlation matrix for the four measures that yields a tetrad difference of zero (values taken from Spearman's example in another context, 1927, p. 74). Fig. 6.8 also shows a different hypothetical matrix yielding a positive tetrad difference and a conclusion (if other tetrads show the same pattern) that group factors for length and width must be recognized.

The top matrix of Fig. 6.8 illustrates another important point that reverberates throughout the history of factor analysis in psychology. Note that, although the tetrad difference is zero, the correlation coefficients need not be (and almost invariably are not) equal. In this case, leg width with leg length gives a correlation of 0.80, while tail width with tail length yields only 0.18. These differences reflect varying "saturations" with $g$, the single general factor when the tetrad differences are zero. Leg measures have higher saturations than tail measures—that is, they are closer to $g$, or reflect it better (in modern terms, they lie closer to the first principal component in geometric representations like Fig. 6.6). Tail measures do not load strongly on $g$.* They contain little common variance and must be explained primarily by their $s$—the information unique to each measure. Moving now to mental tests: if $g$ represents general intelligence, then mental tests most saturated with

*The terms "saturation" and "loading" refer to the correlation between a test and a factor axis. If a test "loads" strongly on a factor then most of its information is explained by the factor.




*6•8* Tetrad differences of zero (above) and a positive value (below) from hypothetical correlation matrices for four measurements: LL = leg length, LW = leg width, TL = tail length, and TW = tail width. The positive tetrad difference indicates the existence of group factors for lengths and widths.



*g* are the best surrogates for general intelligence, while tests with low *g*-loadings (and high *s* values) cannot serve as good measures of general mental worth. Strength of *g*-loading becomes the criterion for determining whether or not a particular mental test (IQ, for example) is a good measure of general intelligence.

Spearman's tetrad procedure is very laborious when the correlation matrix includes a large number of tests. Each tetrad difference must be calculated separately. If the common variance reflects but a single general factor, then the tetrads should equal zero. But, as in any statistical procedure, not all cases meet the expected value (half heads and half tails is the expectation in coin flipping, but you will flip six heads in a row about once in sixty-four series of six flips). Some calculated tetrad differences will be positive or negative even when a single *g* exists and the expected value is zero. Thus, Spearman computed all tetrad differences and looked for normal frequency distributions with a mean tetrad difference of zero as his test for the existence of *g*.

### *Spearman's* g *and the great instauration of psychology*

Charles Spearman computed all his tetrads, found a distribution close enough to normal with a mean close enough to zero, and proclaimed that the common variance in mental tests recorded but a single underlying factor—Spearman's *g*, or general intelligence. Spearman did not hide his pleasure, for he felt that he had discovered the elusive entity that would make psychology a true science. He had found the innate essence of intelligence, the reality underlying all the superficial and inadequate measures devised to search for it. Spearman's *g* would be the philosopher's stone of psychology, its hard, quantifiable "thing"—a fundamental particle that would pave the way for an exact science as firm and as basic as physics.

In his 1904 paper, Spearman proclaimed the ubiquity of *g* in all processes deemed intellectual: "All branches of intellectual activity have in common one fundamental function . . . whereas the remaining or specific elements seem in every case to be wholly different from that in all the others. . . . This *g*, far from being confined to some small set of abilities whose intercorrelations have actually been measured and drawn up in some particular table, may enter into all abilities whatsoever."

The conventional school subjects, insofar as they reflect aptitude rather than the simple acquisition of information, merely peer through a dark glass at the single essence inside: "All examination in the different sensory, school, and other specific faculties may be considered as so many independently obtained estimates of the one great common Intellective Function" (1904, p. 273). Thus Spearman tried to resolve a traditional dilemma of conventional education for the British elite: why should training in the classics make a better soldier or a statesman? "Instead of continuing ineffectively to protest that high marks in Greek syntax are no test as to the capacity of men to command troops or to administer provinces, we shall at last actually determine the precise accuracy of the various means of measuring General Intelligence" (1904, p. 277). In place of fruitless argument, one has simply to determine the *g*-loading of Latin grammar and military acuity. If both lie close to *g*, then skill in conjugation may be a good estimate of future ability to command.

There are different styles of doing science, all legitimate and partially valid. The beetle taxonomist who delights in noting the peculiarities of each new species may have little interest in reduction, synthesis, or in probing for the essence of "beetleness"—if such exists! At an opposite extreme, occupied by Spearman, the externalities of this world are only superficial guides to a simpler, underlying reality. In a popular image (though some professionals would abjure it), physics is the ultimate science of reduction to basic and quantifiable causes that generate the apparent complexity of our material world. Reductionists like Spearman, who work in the so-called soft sciences of organismic biology, psychology, or sociology, have often suffered from "physics envy." They have strived to practice their science according to their clouded vision of physics—to search for simplifying laws and basic particles. Spearman described his deepest hopes for a science of cognition (1923, p. 30):

> Deeper than the uniformities of occurrence which are noticeable even without its aid, it [science] discovers others more abstruse, but correspondingly more comprehensive, upon which the name of laws is bestowed. . . . When we look around for any approach to this ideal, something of the sort can actually be found in the science of physics as based on the three primary laws of motion. Coordinate with this *physica corporis* [physics of bodies], then, we are today in search of a *physica animae* [physics of the soul].



With *g* as a quantified, fundamental particle, psychology could take its rightful place among the real sciences. "In these principles," he wrote in 1923 (p. 355), "we must venture to hope that the so long missing genuinely scientific foundation for psychology has at last been supplied, so that it can henceforward take its due place along with the other solidly founded sciences, even physics itself." Spearman called his work "a Copernican revolution in point of view" (1927, p. 411) and rejoiced that "this Cinderella among the sciences has made a bold bid for the level of triumphant physics itself" (1937, p. 21).

### *Spearman's* g *and the theoretical justification of IQ*

Spearman, the theorist, the searcher for unity by reduction to underlying causes, often spoke in most unflattering terms about the stated intentions of IQ testers. He referred to IQ (1931) as "the mere average of sub-tests picked up and put together without rhyme or reason." He decried the dignification of this "gallimaufry of tests" with the name intelligence. In fact, though he had described his *g* as general intelligence in 1904, he later abandoned the word intelligence because endless arguments and inconsistent procedures of mental testers had plunged it into irremediable ambiguity (1927, p. 412; 1950, p. 67).

Yet it would be incorrect—indeed it would be precisely contrary to Spearman's view—to regard him as an opponent of IQ testing. He had contempt for the atheoretical empiricism of the testers, their tendency to construct tests by throwing apparently unrelated items together and then offering no justification for such a curious procedure beyond the claim that it yielded good results. Yet he did not deny that the Binet tests worked, and he rejoiced in the resuscitation of the subject thus produced: "By this one great investigation [the Binet scale] the whole scene was transformed. The recently despised tests were now introduced into every country with enthusiasm. And everywhere their practical application was brilliantly successful" (1914, p. 312).

What galled Spearman was his conviction that IQ testers were doing the right thing in amalgamating an array of disparate items into a single scale, but that they refused to recognize the theory behind such a procedure and continued to regard their work as rough-and-ready empiricism.

Spearman argued passionately that the justification for Binet

testing lay with his own theory of a single *g* underlying all cognitive activity. IQ tests worked because, unbeknownst to their makers, they measured *g* with fair accuracy. Each individual test has a *g*-loading and its own specific information (or *s*), but *g*-loading varies from nearly zero to nearly 100 percent. Ironically, the most accurate measure of *g* will be the average score for a large collection of individual tests of the most diverse kind. Each measures *g* to some extent. The variety guarantees that *s*-factors of the individual tests will vary in all possible directions and cancel each other out. Only *g* will be left as the factor common to all tests. IQ works because it measures *g*.

> An explanation is at once supplied for the success of their extraordinary procedure of . . . pooling together tests of the most miscellaneous description. For if every performance depends on two factors, the one always varying randomly, while the other is constantly the same, it is clear that in the average the random variations will tend to neutralize one another, leaving the other, or constant factor, alone dominant (1914, p. 313; see also, 1923, p. 6, and 1927, p. 77).

Binet's "hotchpot of multitudinous measurements" was a correct theoretical decision, not only the intuitive guess of a skilled practitioner: "In such wise this principle of making a hotchpot, which might seem to be the most arbitrary and meaningless procedure imaginable, had really a profound theoretical basis and a supremely practical utility" (Spearman quoted in Tuddenham, 1962, p. 503).

Spearman's *g*, and its attendant claim that intelligence is a single, measurable entity, provided the only promising theoretical justification that hereditarian theories of IQ have ever had. As mental testing rose to prominence during the early twentieth century, it developed two traditions of research that Cyril Burt correctly identified in 1914 (p. 36) as correlational methods (factor analysis) and age-scale methods (IQ testing). Hearnshaw has recently made the same point in his biography of Burt (1979, p. 47): "The novelty of the 1900's was not in the concept of intelligence itself, but in its operational definition in terms of correlational techniques, and in the devising of practicable methods of measurement."

No one recognized better than Spearman the intimate connection between his model of factor analysis and hereditarian interpretations of IQ testing. In his 1914 *Eugenics Review* article, he



prophesied the union of these two great traditions in mental testing: "Each of these two lines of investigation furnishes a peculiarly happy and indispensable support to the other. . . . Great as has been the value of the Simon-Binet tests, even when worked in theoretical darkness, their efficiency will be multiplied a thousand-fold when employed with a full light upon their essential nature and mechanism." When Spearman's style of factor analysis came under attack late in his career (see pp. 326–332), he defended *g* by citing it as the rationale for IQ: "Statistically, this determination is grounded on its extreme simpleness. Psychologically, it is credited with affording the sole base for such useful concepts as those of 'general ability,' or 'IQ' " (1939, p. 79).

To be sure, the professional testers did not always heed Spearman's plea for an adoption of *g* as the rationale for their work. Many testers abjured theory and continued to insist on practical utility as the justification for their efforts. But silence about theory does not connote an absence of theory. The reification of IQ as a biological entity has depended upon the conviction that Spearman's *g* measures a single, scalable, fundamental "thing" residing in the human brain. Many of the more theoretically inclined mental testers have taken this view (see Terman et al., 1917, p. 152). C. C. Brigham did not base his famous recantation solely upon a belated recognition that the army mental tests had considered patent measures of culture as inborn properties (pp. 262–263). He also pointed out that no strong, single *g* could be extracted from the combined tests, which, therefore, could not have been measures of intelligence after all (Brigham, 1930). And I will at least say this for Arthur Jensen: he recognizes that his hereditarian theory of IQ depends upon the validity of *g*, and he devotes much of his major book (1979) to a defense of Spearman's argument in its original form, as do Richard Herrnstein and Charles Murray in *The Bell Curve* (1994)—see essays at end of this book. A proper understanding of the conceptual errors in Spearman's formulation is a prerequisite for criticizing hereditarian claims about IQ at their fundamental level, not merely in the tangled minutiae of statistical procedures.

### *Spearman's reification of* g

Spearman could not rest content with the idea that he had probed deeply under the empirical results of mental tests and

found a single abstract factor underlying all performance. Nor could he achieve adequate satisfaction by identifying that factor with what we call intelligence itself.* Spearman felt compelled to ask more of his *g:* it must measure some physical property of the brain; it must be a "thing" in the most direct, material sense. Even if neurology had found no substance to identify with *g*, the brain's performance on mental tests proved that such a physical substrate must exist. Thus, caught up in physics envy again, Spearman described his own "adventurous step of deserting all actually observable phenomena of the mind and proceeding instead to invent an underlying something which—by analogy with physics—has been called mental energy" (1927, p. 89).

Spearman looked to the basic property of *g*—its influence in varying degree, upon mental operations—and tried to imagine what physical entity best fitted such behavior. What else, he argued, but a form of energy pervading the entire brain and activating a set of specific "engines," each with a definite locus. The more energy, the more general activation, the more intelligence. Spearman wrote (1923, p. 5):

> This continued tendency to success of the same person throughout all variations of both form and subject matter—that is to say, throughout all conscious aspects of cognition whatever—appears only explicable by some factor lying deeper than the phenomena of consciousness. And thus there emerges the concept of a hypothetical general and purely quantitative factor underlying all cognitive performances of any kind. . . . The factor was taken, pending further information, to consist in something of the nature of an "energy" or "power" which serves in common the whole cortex (or possibly, even, the whole nervous system)."

If *g* pervades the entire cortex as a general energy, then the *s*-factors for each test must have more definite locations. They must represent specific groups of neurons, activated in different ways by the energy identified with *g*. The *s*-factors, Spearman wrote (and not merely in metaphor), are engines fueled by a circulating *g*.

> Each different operation must necessarily be further served by some specific factor peculiar to it. For this factor also, a physiological substrate has been suggested, namely the particular group of neurons specially serving the particular kind of operation. These neural groups would thus function as alternative "engines" into which the common supply of "energy" could be alternatively distributed. Successful action would always depend, partly on the potential of energy developed in the whole cortex, and partly on the efficiency of the specific group of neurons involved. The relative influence of these two factors could vary greatly according to the kind of operation; some kinds would depend more on the potential of the energy, others more on the efficiency of the engine (1923, pp. 5–6).

*At least in his early work. Later, as we have seen, he abandoned the word intelligence as a result of its maddening ambiguity in common usage. But he did not cease to regard *g* as the single cognitive essence that should be called intelligence, had not vernacular (and technical) confusion made such a mockery of the term.



The differing *g*-loadings of tests had been provisionally explained: one mental operation might depend primarily upon the character of its engine (high *s* and low *g*-loading), another might owe its status to the amount of general energy involved in activating its engine (high *g*-loading).

Spearman felt sure that he had discovered the basis of intelligence, so sure that he proclaimed his concept impervious to disproof. He expected that a physical energy corresponding with *g* would be found by physiologists: "There seem to be grounds for hoping that a material energy of the kind required by psychologists will some day actually be discovered" (1927, p. 407). In this discovery, Spearman proclaimed, "physiology will achieve the greatest of its triumphs" (1927, p. 408). But should no physical energy be found, still an energy there must be—but of a different sort:

> And should the worst arrive and the required physiological explanation remain to the end undiscoverable, the mental facts will none the less remain facts still. If they are such as to be best explained by the concept of an underlying energy, then this concept will have to undergo that which after all is only what has long been demanded by many of the best psychologists—it will have to be regarded as purely mental (1927, p. 408).

Spearman, in 1927 at least, never considered the obvious alternative: that his attempt to reify *g* might be invalid in the first place.

Throughout his career, Spearman tried to find other regularities of mental functioning that would validate his theory of general energy and specific engines. He enunciated (1927, p. 133) a "law of constant output" proclaiming that the cessation of any mental activity causes others of equal intensity to commence. Thus, he reasoned, general energy remains intact and must always be activating something. He found, on the other hand, that fatigue is "selectively transferred"—that is, tiring in one mental activity entails fatigue in some related areas, but not in others (1927, p. 318). Thus, fatigue

cannot be attributed to "decrease in the supply of the general psycho-physiological energy," but must represent a build up of toxins that act selectively upon certain kinds of neurons. Fatigue, Spearman proclaimed, "primarily concerns not the energy but the engines" (1927, p. 318).

Yet, as we find so often in the history of mental testing, Spearman's doubts began to grow until he finally recanted in his last (posthumously published) book of 1950. He seemed to pass off the theory of energy and engines as a folly of youth (though he had defended it staunchly in middle age). He even abandoned the attempt to reify factors, recognizing belatedly that a mathematical abstraction need not correspond with a physical reality. The great theorist had entered the camp of his enemies and recast himself as a cautious empiricist (1950, p. 25):

> We are under no obligation to answer such questions as: whether "factors" have any "real" existence? do they admit of genuine "measurement"? does the notion of "ability" involve at bottom any kind of cause, or power? Or is it only intended for the purpose of bare description? . . . At their time and in their place such themes are doubtless well enough. The senior writer himself has indulged in them not a little. *Dulce est desipere in loco* [it is pleasant to act foolishly from time to time—a line from Horace]. But for the present purposes he has felt himself constrained to keep within the limits of barest empirical science. These he takes to be at bottom nothing but description and prediction. . . . The rest is mostly illumination by way of metaphor and similes.

The history of factor analysis is strewn with the wreckage of misguided attempts at reification. I do not deny that patterns of causality may have identifiable and underlying, physical reasons, and I do agree with Eysenck when he states (1953, p. 113): "Under certain circumstances, factors may be regarded as hypothetical causal influences underlying and determining the observed relationships between a set of variables. It is only when regarded in this light that they have interest and significance for psychology." My complaint lies with the practice of assuming that the mere existence of a factor, in itself, provides a license for causal speculation. Factorists have consistently warned against such an assumption, but our Platonic urges to discover underlying essences continue to prevail over proper caution. We can chuckle, with the beneficence of hindsight, at psychiatrist T. V. Moore who, in 1933, postulated definite genes for catatonic, deluded, manic, cognitive, and constitutional depression because his factor analysis grouped the supposed measures of these syndromes on separate axes (in Wolfle, 1940). Yet in 1972 two authors found an association of dairy production with florid vocalization on the tiny thirteenth axis of a nineteen-axis factor analysis for musical habits of various cultures—and then suggested "that this extra source of protein accounts for many cases of energetic vocalizing" (Lomax and Berkowitz, 1972, p. 232).



Automatic reification is invalid for two major reasons. First, as I discussed briefly on pp. 282–285 and will treat in full on pp. 326–347, no set of factors has any claim to exclusive concordance with the real world. Any matrix of positive correlation coefficients can be factored, as Spearman did, into *g* and a set of subsidiary factors or, as Thurstone did, into a set of "simple structure" factors that usually lack a single dominant direction. Since either solution resolves the same amount of information, they are equivalent in mathematical terms. Yet they lead to contrary psychological interpretations. How can we claim that one, or either, is a mirror of reality?

Second, any single set of factors can be interpreted in a variety of ways. Spearman read his strong *g* as evidence for a single reality underlying all cognitive mental activity, a general energy within the brain. Yet Spearman's most celebrated English colleague in factor analysis, Sir Godfrey Thomson, accepted Spearman's mathematical results but consistently chose to interpret them in an opposite manner. Spearman argued that the brain could be divided into a set of specific engines, fueled by a general energy. Thomson, using the same data, inferred that the brain has hardly any specialized structure at all. Nerve cells, he argued, either fire completely or not at all—they are either off or on, with no intermediary state. Every mental test samples a random array of neurons. Tests with high *g*-loadings catch many neurons in the active state; others, with low *g*-loadings, have simply sampled a smaller amount of unstructured brain. Thomson concluded (1939): "Far from being divided up into a few 'unitary factors,' the mind is a rich, comparatively undifferentiated complex of innumerable influences—on the physiological side an intricate network of possibilities of intercommunication." If the same mathematical pattern can yield such disparate interpretations, what claim can either have upon reality?

*Spearman on the inheritance of* g

Two of Spearman's primary claims appear in most hereditarian theories of mental testing: the identification of intelligence as a unitary "thing," and the inference of a physical substrate for it. But these claims do not complete the argument: a single, physical substance may achieve its variable strength through effects of environment and education, not from inborn differences. A more direct argument for the heritability of *g* must be made, and Spearman supplied it.

The identification of *g* and *s* with energy and engines again provided Spearman with his framework. He argued that the *s*-factors record training in education, but that the strength of a person's *g* reflects heredity alone. How can *g* be influenced by education, Spearman argued (1927, p. 392), if *g* ceases to increase by about age sixteen but education may continue indefinitely thereafter? How can *g* be altered by schooling if it measures what Spearman called *eduction* (or the ability to synthesize and draw connections) and not *retention* (the ability to learn facts and remember them)—when schools are in the business of imparting information? The engines can be stuffed full of information and shaped by training, but the brain's general energy is a consequence of its inborn structure:

> The effect of training is confined to the specific factor and does not touch the general one; physiologically speaking, certain neurons become habituated to particular kinds of action, but the free energy of the brain remains unaffected. . . . Though unquestionably the development of specific abilities is in large measure dependent upon environmental influences, that of general ability is almost wholly governed by heredity (1914, pp. 233–234).

IQ, as a measure of *g*, records an innate general intelligence; the marriage of the two great traditions in mental measurement (IQ testing and factor analysis) was consummated with the issue of heredity.

On the vexatious issue of group differences, Spearman's views accorded with the usual beliefs of leading western European male scientists at the time (see Fig. 6.9). Of blacks, he wrote (1927, p. 379), invoking *g* to interpret the army mental tests:



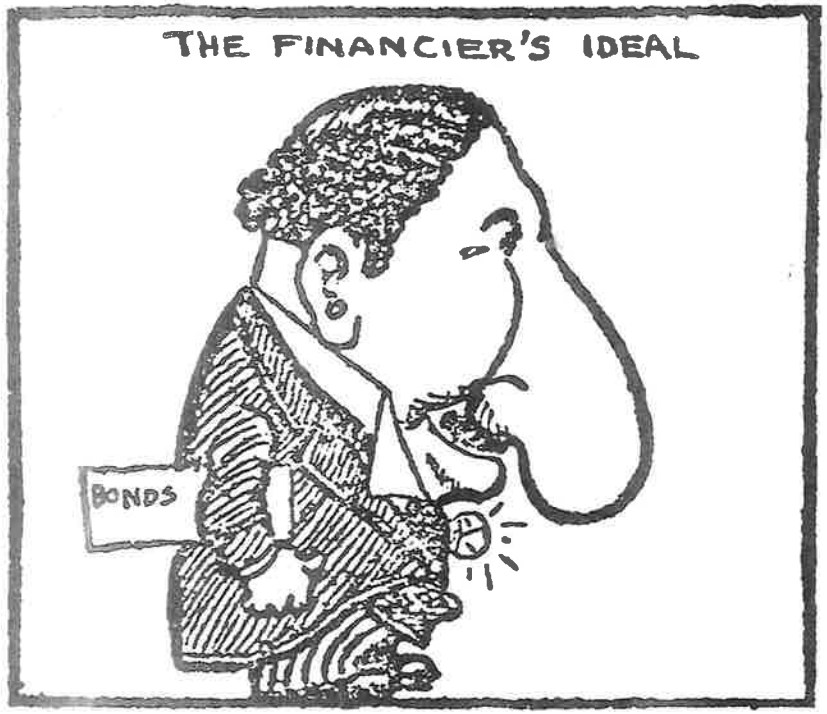
> On the average of all the tests, the colored were about two years behind the white; their inferiority extended through all ten tests, but it was most marked in just those which are known to be most saturated with *g*.

In other words, blacks performed most poorly on tests having strongest correlations with *g*, or innate general intelligence.

Of whites from southern and eastern Europe, Spearman wrote (1927, p. 379), praising the American Immigration Restriction Act of 1924:

> The general conclusion emphasized by nearly every investigator is that, as regards "intelligence," the Germanic stock has on the average a marked advantage over the South European. And this result would seem to have

**6•9** Racist stereotype of a Jewish financier, reproduced from the first page of Spearman's 1914 article (see Bibliography). Spearman used this figure to criticize beliefs in group factors for such particular items of intellect, but its publication illustrates the acceptable attitudes of another age.



had vitally important practical consequences in shaping the recent very stringent American laws as to admission of immigrants.

Yet it would be incorrect to brand Spearman as an architect of the hereditarian theory for differences in intelligence among human groups. He supplied some important components, particularly the argument that intelligence is an innate, single, scorable "thing." He also held conventional views on the source of average differences in intelligence between races and national groups. But he did not stress the ineluctability of differences. In fact, he attributed sexual differences to training and social convention (1927, p. 229) and had rather little to say about social classes. Moreover, when discussing racial differences, he always coupled his hereditarian claim about average scores with an argument that the range of variation within any racial or national group greatly exceeds the small average difference between groups—so that many members of an "inferior" race will surpass the average intelligence of a "superior" group (1927, p. 380, for example).*

Spearman also recognized the political force of hereditarian claims, though he did not abjure either the claim or the politics: "All great efforts to improve human beings by way of training are thwarted through the apathy of those who hold the sole feasible road to be that of stricter breeding" (1927, p. 376).

But, most importantly, Spearman simply didn't seem to take much interest in the subject of hereditary differences among peoples. While the issue swirled about him and buried his profession in printer's ink, and while he himself had supplied a basic argument for the hereditarian school, the inventor of *g* stood aside in apparent apathy. He had studied factor analysis because he wanted to understand the structure of the human brain, not as a guide to measuring differences between groups, or even among individuals. Spearman may have been a reluctant courtier, but the politically potent union of IQ and factor analysis into a hereditarian theory of intelligence was engineered by Spearman's successor in the chair of psychology at University College—Cyril Burt. Spearman may have cared little, but the innate character of intelligence was the idée fixe of Sir Cyril's life.

*Richard Herrnstein and Charles Murray emphasize the same arguments to obviate a charge of racism against *The Bell Curve* (1994)—see first two essays at end of book.



## Cyril Burt and the hereditarian synthesis

### *The source of Burt's uncompromising hereditarianism*

Cyril Burt published his first paper in 1909. In it, he argued that intelligence is innate and that differences between social classes are largely products of heredity; he also cited Spearman's *g* as primary support. Burt's last paper in a major journal appeared posthumously in 1972. It sang the very same tune: intelligence is innate and the existence of Spearman's *g* proves it. For all his more dubious qualities, Cyril Burt certainly had staying power. The 1972 paper proclaims:

> The two main conclusions we have reached seem clear and beyond all question. The hypothesis of a general factor entering into every type of cognitive process, tentatively suggested by speculations derived from neurology and biology, is fully borne out by the statistical evidence; and the contention that differences in this general factor depend largely on the individual's genetic constitution appears incontestable. The concept of an innate, general, cognitive ability, which follows from these two assumptions, though admittedly a sheer abstraction, is thus wholly consistent with the empirical facts (1972, p. 188).

Only the intensity of Sir Cyril's adjectives had changed. In 1912 he had termed this argument "conclusive"; by 1972 it had become "incontestable."

Factor analysis lay at the core of Burt's definition of intelligence as i.g.c. (innate, general, cognitive) ability. In his major work on factor analysis (1940, p. 216), Burt developed his characteristic use of Spearman's thesis. Factor analysis shows that "a *general* factor enters into all *cognitive* processes," and "this general factor appears to be largely, if not wholly, inherited or *innate*"—again, i.g.c. ability. Three years earlier (1937, pp. 10–11) he had tied *g* to an ineluctable heredity even more graphically:

> This general intellectual factor, central and all-pervading, shows a further characteristic, also disclosed by testing and statistics. It appears to be inherited, or at least inborn. Neither knowledge nor practice, neither interest nor industry, will avail to increase it.

Others, including Spearman himself, had drawn the link between *g* and heredity. Yet no one but Sir Cyril ever pursued it with such stubborn, almost obsessive gusto: and no one else

"Please God, I think that you made me in the shape which I now have for reasons best known to Yourselves and that it would be rude to change. If I am to have my choice, I will stay as I am. I will not alter any of the parts which you gave me. . . . I will stay a defenceless embryo all my life, doing my best to make myself a few feeble implements out of the wood, iron, and the other materials which You have seen fit to put before me. . . ." "Well done," exclaimed the Creator in delighted tone. "Here, all you embryos, come here with your beaks and whatnots to look upon Our first Man. He is the only one who has guessed Our riddle. . . . As for you, Man. . . . You will look like an embryo till they bury you, but all the others will be embryos before your might. Eternally undeveloped, you will always remain potential in Our image, able to see some of Our sorrows and to feel some of Our joys. We are partly sorry for you, Man, but partly hopeful. Run along then, and do your best."

# Epilogue

In 1927 Oliver Wendell Holmes, Jr., delivered the Supreme Court's decision upholding the Virginia sterilization law in *Buck v. Bell*. Carrie Buck, a young mother with a child of allegedly feeble mind, had scored a mental age of nine on the Stanford-Binet. Carrie Buck's mother, then fifty-two, had tested at mental age seven. Holmes wrote, in one of the most famous and chilling statements of our century:

> We have seen more than once that the public welfare may call upon the best citizens for their lives. It would be strange if it could not call upon those who already sap the strength of the state for these lesser sacrifices. . . . Three generations of imbeciles are enough.

(The line is often miscited as "three generations of idiots. . . ." But Holmes knew the technical jargon of his time, and the Bucks, though not "normal" by the Stanford-Binet, were one grade above idiots.)

*Buck v. Bell* is a signpost of history, an event linked with the distant past in my mind. The Babe hit his sixty homers in 1927, and legends are all the more wonderful because they seem so distant. I was therefore shocked by an item in the *Washington Post* on 23 February 1980—for few things can be more disconcerting than a juxtaposition of neatly ordered and separated temporal events. "Over 7,500 sterilized in Virginia," the headline read. The law that Holmes upheld had been implemented for forty-eight years, from 1924 to 1972. The operations had been performed in mental-health facilities, primarily upon white men and women considered feeble-minded and antisocial—including "unwed mothers, prostitutes, petty criminals and children with disciplinary problems."

Carrie Buck, then in her seventies, was still living near Charlottesville. Several journalists and scientists visited Carrie Buck and her sister, Doris, during the last years of their lives. Both women, though lacking much formal education, were clearly able and intelligent. Nonetheless, Doris Buck had been sterilized under the same law in 1928. She later married Matthew Figgins, a plumber. But Doris Buck was never informed. "They told me," she recalled, "that the operation was for an appendix and rupture." So she and Matthew Figgins tried to conceive a child. They consulted physicians at three hospitals throughout her child-bearing years; no one recognized that her Fallopian tubes had been severed. Last year, Doris Buck Figgins finally discovered the cause of her lifelong sadness.

One might invoke an unfeeling calculus and say that Doris Buck's disappointment ranks as nothing compared with millions dead in wars to support the designs of madmen or the conceits of rulers. But can one measure the pain of a single dream unfulfilled, the hope of a defenseless woman snatched by public power in the name of an ideology advanced to purify a race. May Doris Buck's simple and eloquent testimony stand for millions of deaths and disappointments and help us to remember that the Sabbath was made for man, not man for the Sabbath: "I broke down and cried. My husband and me wanted children desperately. We were crazy about them. I never knew what they'd done to me."

# Critique of *The Bell Curve*

## *The Bell Curve*

*The Bell Curve* by Richard J. Herrnstein and Charles Murray provides a superb and unusual opportunity for insight into the meaning of experiment as a method in science. Reduction of confusing variables is the primary desideratum in all experiments. We bring all the buzzing and blooming confusion of the external world into our laboratories and, holding all else constant in our artificial simplicity, try to vary just one potential factor at a time. Often, however, we cannot use such an experimental method, particularly for most social phenomena when importation into the laboratory destroys the subject of our investigation—and then we can only yearn for simplifying guides in nature. If the external world therefore obliges and holds some crucial factors constant for us, then we can only offer thanks for such a natural boost to understanding.

When a book garners as much attention as *The Bell Curve* has received, we wish to know the causes. One might suspect content itself—a startling new idea, or an old suspicion now verified by persuasive data—but the reason might well be social acceptability, or just plain hype. *The Bell Curve* contains no new arguments and presents no compelling data to support its anachronistic social Darwinism. I must therefore conclude that its initial success in winning such attention must reflect the depressing temper of our time—a historical moment of unprecedented ungenerosity, when a mood for slashing social programs can be so abetted by an argument that beneficiaries cannot be aided due to inborn cognitive limits expressed as low IQ scores.

*The Bell Curve* rests upon two distinctly different but sequential