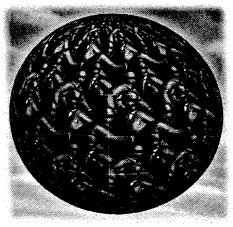

Psychometrics of Intelligence

Peter H. Schonemann
Purdue University, West Lafayette, Indiana, USA

Glossary

common factors Latent variables implied by Spearman's (and later Thurstone's) factor analysis model. On partialling out all common factors of the observed variables, only uncorrelated specific factors are left.
congruence coefficient Cosine.
correlation Measure of linear relationship between two variables varying between −1 and 1 (0, no linear relationship; 1, perfect linear relationship).
criterion A variable of practical interest that a test is intended to predict.
dominant eigenvector The eigenvector associated with the largest (dominant) eigenvalue of a symmetric matrix.
eigenvector A vector mapped into a scalar multiple of itself by a square matrix. The scalar is called (the associated) eigenvalue.
general ability (*g*) An unobserved variable implied by a mathematical model proposed by Spearman to account for positive correlations among intelligence tests. In contrast to a PC1, *g* is not a linear combination of the observed tests.
intelligence Technically undefined, the term refers to a broad spectrum of "cognitive" skills presumed relevant for educational and economic success.
IQ total test score on an IQ test, normed to have a mean of 100 and a standard deviation of 15 or 16.
IQ test A test presumed to measure intelligence.
item A subtest of a test, usually scored 1/0 (pass/fail, binary item). The total test score is the sum of the item scores.
latent Implied but not observed overtly.
linear combination (linear composite) Weighted average.
matrix A rectangular array of real numbers.
partial correlation A correlation that remains after the influence of one or more other variables has been statistically removed (partialled out).
PC1, first principal component A linear combination of the observed tests that has largest variance among all possible linear combinations (subject to the constraint that the defining weight vector be of unit length).
reliability A measure of stability of test scores under repeated application to the same subjects; usually expressed as a correlation.
scalar A real or complex number.
validity A measure of the extent to which a test measures what it is designed to measure; often expressed as a correlation.
vector A linear array of real numbers set out either as a row or as a column.

Psychometrics, which literally means "measurement of the soul," is a subdiscipline of psychology devoted to the development, evaluation, and application of mental tests. It is useful to distinguish between two branches of psychometrics, a theoretical and an applied branch. They do not interact as much as might be expected. Here, both branches will be tracked side by side since it is impossible to gauge the merits of a theory without knowing what benefits it produced in practice.

This article discusses developments in test theory and factor analysis, with emphasis on applications to intelligence. This topic originally spawned interest in psychometrics, spurred on its early growth, inspired most of its lasting achievements, and, in the end, also revealed its limitations. It also has had the most profound social impact.

Introduction

The history of psychometrics spans approximately a century, beginning in earnest in approximately 1904/1905, when Binet and Spearman laid the foundations for future developments. This era was followed by a period of consolidation and growing acceptance of IQ tests and college admission tests in the United States.

Under Thurstone in the 1940s, psychometrics became academically respectable but also progressively more dogmatic and mechanical. The field reached its creative apogee under Cronbach and Guttman in the 1950s. Then, it began to stagnate and eventually regressed back to its racist roots in the 1920s.

In this article, matrices are denoted by capital letters, and scalars are denoted by lowercase letters. Column vectors are denoted by boldface lowercase letters, and row vectors are denoted by boldface lowercase letters followed by a prime denoting transposition.

Auspicious Beginnings: Charles Spearman and Alfred Binet

Galton, Wundt, and Wissler

The applied branch of psychometrics goes back at least to the days of Wundt and Galton (1880s). Inspired by the example of the physical sciences, both scholars attempted to measure basic sensory and physiological aspects of human behavior first, such as reaction time, visual acuity, and the like, to lay the grounds for the investigation of more complex variables later.

Galton, together with some of his younger colleagues, especially Karl Pearson and Udny Yule, also contributed to the foundation of the theoretical branch by developing basic correlational techniques that soon were to play a central role in both branches of psychometrics.

At the turn of the 20th century, Wissler applied these new techniques to school grades and some of the mental tests available at the time. The results proved disappointing. The sensory and physiological measures correlated moderately with each other, as did school grades, as indicators of more complex forms of mental ability. However, the correlations between both sets of variables were virtually zero. At this critical juncture, and almost simultaneously, two events breathed new life into the seemingly moribund new science.

Alfred Binet

On the applied side, in 1905 Alfred Binet (with Simon) published a new mental test that violated all canons of the Galton/Wundt school. Instead of trying to synthesize measures of more complex behaviors from simpler, more easily measured sensory variables, Binet set out to measure a highly complex mental trait of undisputed practical importance—intelligence—directly. The test he devised comprised a large number of items designed to sample various aspects of the implied target variable. This variable soon acquired the technical-sounding acronym IQ. It was defined, eventually, in terms of the total item score (which skipping some of the intervening mental age arithmetic that never applied to adults anyway). Binet's test became the prototype of most IQ tests and also of the scholastic aptitude tests still in use today. The problems with this approach begin after one constructs a second IQ test that does not correlate perfectly with the first, since then it is no longer clear which one is the true measure of intelligence.

Charles Spearman's General Ability (*g*)

To cope with this problem, in 1904 Spearman developed an entirely new theory aimed at supplanting the widely used but murky term intelligence with a clear-cut operational definition. Starting with the observation that most measures thought to relate to intelligence, such as school grades, tended to correlate positively, he reasoned this means they all measure the same latent variable, namely intelligence. To account for the fact that the correlations, although generally positive, were not perfect, he appealed to the recently developed machinery of partial correlations: If all positive correlations in a test battery are due to a single underlying variable—which he called *g*, short for general ability, to steer clear of the term intelligence—then its statistical removal should produce a matrix of partial correlations that are all zero. If this should turn out to be true for all batteries of "intelligence tests" worthy of the name, then this latent variable *g* could serve as an unequivocal definition of intelligence, which thus would have been "objectively determined and measured" (the title of his 1904 paper). The globality clause is usually omitted in modern accounts of Spearman's work, although it is absolutely critical for the stringency of this argument. After testing his theory on 12 previously published small data sets, he found it consistently confirmed: Dodd noted that "it seemed to be the most striking quantitative fact in the history of psychology."

Spearman left no doubt about how he felt about the social relevance of his presumed discovery: "Citizens, instead of choosing their careers at almost blind hazard, will undertake just the professions suited to their capacities. One can even conceive the establishment of a minimum index to qualify for parliamentary vote, and above all, for the right to have offspring" (Hart and Spearman, 1912).

Early Criticisms

This initial phase of exuberance soon gave way to sobering reappraisals. In 1916, Thomson noticed that the same type of correlation matrix that Spearman's theory predicts on postulating one common factor is also implied by a diametrically opposed theory that postulates infinitely many common factors (Thomson's sampling theory). To make matters worse, in 1928 E. B. Wilson, a polymath whose stature Spearman instantly acknowledged, wryly observed in an otherwise favorable review of Spearman's

"Ability of Man" (1927) that Spearman's theory did not suffice to define *g* uniquely because it postulated more independent factors than observed tests. As a result, many widely different "intelligence scores" can be assigned to the same subject on the basis of his or her observed test scores. This "factor indeterminacy" issue further undermined Spearman's claim of having defined intelligence objectively as *g*. For a while, it became the focus of lively debate, which soon died down after Thurstone entered the stage.

Eventually, it also became apparent empirically that the early accolades bestowed on Spearman's presumed discovery had been premature. As it turned out, the number of common factors needed to account for the observed test correlations did not stop at one for larger batteries, as the theory required; rather, it increased with the number of subtests in the battery. Typically, approximately one-third as many common factors as tests are needed to reduce the partial correlations to zero.

The fact that his claim of having empirically defined and measured intelligence turned out to be untenable in no way diminishes Spearman's stature for having been first to recognize and address this fundamental challenge and for dealing with it in the spirit of an empirical science.

1910–1930: Consolidation: World War I, Louis Terman, and Carl Brigham

Classical True Score Theory

Spearman's closely knit theory contained a theory of test scores as a special case that, for a considerable period of time, provided the needed definitions for constructing and evaluating new tests. This so-called classical true score theory (CTT) results on applying Spearman's factor model to only two tests and assuming that both contain the underlying common factor—now termed "true score"—in equal amounts (perfectly parallel tests). Under these assumptions, it is not difficult to derive plausible definitions of test reliability and test validity in correlational terms. Since both concepts are derived from the same underlying model, they are closely related. For example, CTT permits to predict how test reliability increases as one increases the number of items comprising the test (Spearman–Brown prophecy formula) and how test validity varies with the reliability of a predictor test, the criterion measure, or both (correction for attenuation). Kuder and Richardson used CTT to derive one of the most popular reliability estimates still widely used today. Cronbach later renamed it coefficient alpha.

Louis Terman

On the applied side, Terman promoted Binet's version of intelligence that is tied to a particular test. Availing himself of the concepts and techniques of CTT, he adapted it to an English-speaking clientele. He shared with many others of his generation, but in marked contrast to Binet, the eugenic prejudices of the early English pioneers (Galton, Pearson, Spearman, and others). This thought collective took for granted that intelligence—whatever it might be—(i) exists and can be measured and (ii) is primarily genetically predetermined (the figure usually given, until very recently, was 80%). Therefore, Terman and his disciples strove to purge their "intelligence tests" as much as possible of environmentally induced impurities (such as educational background) that might mask the underlying, presumed immutable genetic contribution. Note that the first premise conflicts with the outcome of Spearman's efforts.

World War I Army Tests

World War I provided an opportunity to put psychometric theories into practice when the need arose to assess, on short notice, the military qualifications of millions of inductees with widely different social and educational backgrounds. Under the stewardship of a former American Psychological Association president, Robert Yerkes, a committee was charged with the development of intelligence tests suited to the particular needs of the U.S. Army. These efforts resulted in two different tests, the Army Alpha and Army Beta tests. By necessity, both were conceived as group tests (in contrast to the classical Binet test and Terman's American adaptation that had to be administered individually). The Army Alpha was essentially a group version of the Stanford–Binet that relied on the assumption that the testee was fluent in English. The Army Beta was intended as a "nonverbal" substitute for linguistically handicapped inductees whose innate intellectual potential might be masked by traditional verbal tests.

Charles Brigham

World War I offered an opportunity for both branches of psychometrics to promote the fruits of their labors. It also produced a huge database waiting to be mined after the war for new findings to further improve the new technology. This task fell to a young assistant professor, Charles Brigham, who published his findings in a book titled "A Study of American Intelligence," with a foreword by Yerkes.

Brigham's conclusions were in tune with the zeitgeist and seemed perfectly timed to answer steadily mounting concerns over untoward consequences of unrestricted

immigration. He found that the average intelligence of American soldiers was frightfully low. On stratifying his data by country of origin, he further found that "according to all evidence available . . . the American intelligence is declining, and will proceed with an accelerating rate as the racial admixture [having shifted, as he observed, from 'Nordic' to 'Alpine'] becomes more and more extensive" (Brigham, 1923, p. 210). Partly in response to these ominous tidings, the U.S. Congress enacted more restrictive immigration laws in 1924.

These laws were not revoked when Brigham later, to his credit, recanted his earlier prophecies (Brigham, 1930):

> *This review has summarized some of the more recent test findings which show that comparative studies of various national and racial groups may not be made with existing tests, and which show, in particular, that one of the most pretentious of these comparative racial studies—the writer's own—was without foundation.* (p. 165)

The Scholastic Aptitude Test

Brigham went on to develop the first standardized college admissions test [the Scholastic Aptitude Test (SAT)], which in essence was an IQ test tailored to the needs of the decentralized education system in the United States, again stressing the need to tap into innate potential uncontaminated by educational experience. Just as Spearman would have wished, the SAT soon became virtually mandatory for access to higher education in the United States.

Predictive Validities of College Admission Tests

Over the decades, these tests have been refined by infusing ever more sophisticated theoretical and computational advances. However, this did not help improve them in terms of traditional measures of test efficiency. For example, it has been well-known virtually since its inception that the SAT has validities for First Year College GPA (GPA1) that hover around 0.4, approximately 10 correlation points below those of High School Rank (HSR). A tabulation published by the College Board shows that over the time span from 1964 through 1982, the HSR validities varied little (between 0.46 and 0.54, thus explaining approximately 25% of the criterion variance). For the SAT, they range between 0.37 and 0.46 (17%). To make matters worse, Humphreys showed in 1968 that the ACT validities drop into the 0.20s (4%) once the prediction interval is extended to the eighth semester GPA (GPA8). Similarly, it has repeatedly been shown (e.g., by Horn and Hofer and by Sternberg and Williams) that for long-range criteria of practical interest, such as graduation, the validities of the GRE are virtually zero. These findings cannot be dismissed as being due to random error. In contrast to the dubious figures reported for commercial IQ tests, such as the Wechsler and the Stanford–Binet, the sample sizes for the SAT, ACT, and GRE often range in the millions.

From Model to Method: 1930s–1940s—Louis Thurstone

Multiple Factor Analysis

Thurstone's reign of psychometrics extends back into the late 1920's when he published a number of papers devoted to applications of testing models to psychophysics and social psychology, especially attitude measurement. However, he scored his greatest triumph when he extended Spearman's failed factor theory of intelligence from one common factor to more than one common factor (multiple factor analysis). In the process, he transformed factor analysis from a substantive theory into a "general scientific method" susceptible to widespread abuse.

The underlying idea seems straightforward. If, after partialling out one common factor, one finds the resulting partial correlations are still not zero, one might consider partialling out a second common factor, and perhaps a third, until all remaining partial correlations are deemed close enough to zero (multiple factor analysis). In hindsight, this idea seems so obvious that it may not surprise to learn that it had already occurred to others (e.g., Garnett in 1919) long before Thurstone took credit for it in 1947.

Simple Structure

Some additional technical work was required before this simple idea could become useful in practice. Virtually single-handedly, and with considerable ingenuity, Thurstone developed both the theoretical and the technical refinements needed to transform what Spearman had originally intended as a substantive theory of intelligence into a general scientific method.

His most important theoretical contribution was his concept of simple structure intended to resolve the so-called rotation problem. This problem does not arise with Spearman's model because it invokes only one common factor. If there is more than one common factor (e.g., two), then the same observed correlations can be described in infinitely many different ways so that a choice has to be made.

This problem is most easily understood geometrically. If there are two common factors, both can be viewed as the orthogonal axes of a two-dimensional coordinate system used to describe a swarm of test points in a plane. This

swarm of points can be described by many other coordinate systems, even if the origin is kept fixed and the two axes are kept orthogonal to each other. If one chooses a new pair of orthogonal axes by rotating the old system, then the coordinates of the test points—that is, their numerical representation relative to the chosen coordinate system—will change but not the relations among the points, and it is these that constitute the empirical observations. Thus, the question arises which coordinate system to select so as to maximize the scientific utility of the resulting numerical representation of the empirically observed relationships.

Thurstone solved this problem by appealing to the principle of parsimony that is usually attributed to Occam: Explanatory causes should not be multiplied beyond necessity. This means, in the present context, that each test should require as few nonzero coordinates ("loadings") as possible so as to explain it with the smallest possible number of common factors. Hence, after starting with an arbitrary coordinate system, an attempt should be made to rotate it in such a way that each test has as many near zero coordinates as possible. Thurstone called this ideal position "simple structure". If there are only two common factors, such a coordinate system is easily found by visual inspection. However, as the number of common factors increases, this task becomes more difficult. Only with the help of modern computers was this "orthogonal rotation problem" eventually solved to everyone's satisfaction. Thurstone later generalized this rotation problem to correlated factors. The between-factor correlations could then be analyzed for "second-order factors" (hierarchical factor analysis).

Using this methodology, Thurstone arrived at a comprehensive theory of mental tests that dominated American psychometrics during the 1940s and 1950s. Most psychometricians agreed with Thurstone that Spearman had been on the wrong track when he postulated a single common factor of intelligence. Of course intelligence is multidimensional, they argued. All one had to do to see this was apply Thurstone's general scientific method to any battery of intelligence tests. Proceeding in this way, Thurstone derived between 5 and 7 primary mental abilities (PMAs), Eysenck derived 4, R. B. Cattell derived 2 (crystallized and fluid ability), and Guilford derived no less than 120 intelligence factors. However, in practice, Thurstone's test, the PMA, did not perform markedly better than other tests that had been constructed without the benefit of factor analysis.

Consequences

In retrospect, it seems clear that Thurstone's approach was actually a step backward compared to that of Spearman. One reason why Thurstone, Cattell, Eysenck, Guilford, and numerous other psychometricians gave widely differing answers to Spearman's question—what do we mean when we say we "measure intelligence"?—was that they could not even agree on the number of intelligence factors. Yet, whatever the phrase meant, both Spearman and elementary logic tell us that it cannot possibly refer to a multidimensional concept of intelligence. Only unidimensional variables can be "measured" in the sense that exactly one real number is assigned to each subject so that, at a minimum, the empirical order relations (implied by statements of the type "A is more intelligent than B") are preserved. If there were two different intelligences, then all such statements would have to be qualified with a reference to the particular intelligence that is being measured.

In hindsight, it is not surprising that the Thurstone era of exploratory factor analysis, imposing as it seemed at the time, left few empirical imprints still remembered today. Where Spearman had set himself a clearly defined substantive problem, Thurstone promised a research technique uncommitted to any particular subject area, a technique that, moreover, could never fail. No correlation matrix can ever falsify what amounted to a tautological claim that a given number of observed variables can be reasonably well approximated by a smaller number of common factors. Statistical tests of the simple structure hypothesis, although available, were carefully avoided.

Although Thurstone's empirical results are virtually forgotten today, his gospel of a research automaton capable of dispensing scientific progress without requiring any ingenuity, technical skill, or even familiarity with a substantive area proved hugely popular and subsequently resurfaced in various statistical disguises.

1950s–1960s: Apogee—Louis Guttman and Lee Cronbach

Psychometrics reached its apogee in the 1950s under the leadership of Guttman and Cronbach. Both had strong commitments to both branches of psychometrics and were heavily engaged in empirical research.

Louis Guttman

Guttman was without question technically the most accomplished and creative psychometrician in the history of psychometrics. Early in his career he addressed fundamental problems in (unidimensional) scaling. He is most widely known for the joint scale that bears his name. In test theory, he subjected the seemingly innocuous notion of reliability inherited from CTT to searing logical criticism, anticipating later developments by Cronbach and others.

The Importance of Falsifiability

In marked contrast to Thurstone, Guttman never tired of stressing the need to test strong assumptions (e.g., simple structure). Turning his attention to factor analysis, he emphasized that neither parsimony pillar supporting Thurstone's edifice, small rank and simple structure, can be taken for granted. These hypotheses are not just strong but most likely false. A recurrent theme in Guttman's papers is the need to replicate empirical findings instead of simply relying on facile significance tests ("star gazers").

With his radex theory, he proposed a family of alternative structural models unconstrained by these assumptions. In 1955, he also revived interest in the dormant factor indeterminacy issue, recognizing it as a fundamental problem that undermines the very purpose of factor analysis. Thurstone, in contrast, had never faced up to it.

Facet Theory

Later in his career, Guttman returned to Spearman's question, What is intelligence?, which he sought to answer with his facet theory. This research strategy starts with a definition of the domain of the intended variable. Only after this has been done does it become feasible to determine empirically, with a minimum of untested auxiliary assumptions (such as linearity, normality, and the like), whether the staked out domain is indeed unidimensional as Spearman had claimed. If it is not, then one has to lower one's aim and concede that intelligence, whatever else it may be, cannot be measured.

Lee Cronbach

Cronbach also went his own way. As author of a popular text on testing, he was familiar with the problems practicing psychometricians had to face and not easily dazzled by mathematical pyrotechnics devoid of empirical substance. Just as Guttman before him, he also questioned the simplistic assumptions of CTT. Searching for alternatives, he developed a reliability theory that recognized several different types of measurement error—thus yielding several different types of reliability—to be analyzed within the framework of analysis of variance (generalizability theory).

Mental Tests and Personnel Decisions

Most important, in a short book he wrote with Goldine Gleser in 1957, the authors radically departed from the traditional correlational approach for gauging the merits of a test. Instead of asking the conventional questions in correlational terms, they asked a different question in probabilistic terms: How well does the test perform in terms of misclassification rates?

It is surprising that this elementary question had not received more attention earlier. In hindsight, it seems rather obvious that, since use of a test always entails a decision problem, its merit cannot be judged solely in terms of its validity. How useful a test will be in practice also depends on prior knowledge, including the percentage of qualified applicants in the total pool of testees (the base rate) and the stringency of the admission criterion used (the admission quota).

Base Rate Problem

That knowledge of the correlation between the test and the criterion alone cannot possibly suffice to judge the worth of a test is most easily seen in the context of clinical decisions, which often involve very skewed base rates, with the preponderance of subjects being assessed as "normal."

For example, assume the actual incidence (base rate) of normal is 0.90, and that for "pathological" it is 0.10. Suppose further that the test cutoff is adjusted so that 90% of the testees are classified as normal on the basis of their test scores and 10% as pathological. If the joint probability of actually being normal and of being correctly classified as such is 0.85, then one finds that the so-called phi coefficient (as an index of validity) is 0.44. This is quite respectable for a mental test. However, on using it, one finds the following for the probability of total correct classifications (i.e., the joint probability of actually being normal and also being so diagnosed plus the joint probability of actually being pathological and being so diagnosed): $0.85 + 0.05 = 0.90$. This value exactly equals the assumed base rate. Thus, the proportion of total correct classifications achieved on using the test could also have been achieved without it by simply classifying all testees as normal.

The moral of his tale is that the practical utility of a test is a joint function of, among other things, validity, base rate, and admission quota. Validity by itself tells us nothing about the worth of a test. Meehl and Rosen had made much the same point a few years earlier. Cronbach and Gleser expanded on it systematically, tracing out the consequences such a decision-theoretic perspective implies.

To my knowledge, the only currently available complete tables for hit rates (the conditional probability that a qualified testee passes the test) and total percentage correct classifications, as joint functions of validity, base rate, and admission quota, are those published in Schonemann (1997b), in which it is also shown that no test with realistic validity (<0.5) improves over random admission in terms of total percentage correct if either base rate exceeds 0.7.

Notwithstanding the elementary nature of this basic problem and its transparent social relevance, the Social Sciences Citation Index records few, if any, references to it in *Psychometrika*. This is surprising considering that much of modern test theory, with its narrow focus on item analysis, may well become obsolete once one adopts Cronbach's broader perspective. In contrast, some more applied journals did pay attention to the work of Meehl, Rosen, Cronbach, and Gleser.

1970s–1980s: Eugenics Revisited—Arthur Jensen

Some purists may wonder why Jensen is included in this review, since the psychometric elite tends to ignore his work. On the other hand, he also has many admirers. Brandt (1985) calls him "the world's most impressive psychometrician" (p. 222). Whatever one may think of his work, Jensen has clearly had a profound impact on the recent course of psychometrics.

How Much Can We Boost Intelligence and Scholastic Achievement?

Jensen achieved instant notoriety when he challenged the received view that intelligence is primarily a function of environment, not genes. This position had gained ground during World War II, gradually replacing the earlier eugenic thesis to the contrary. To appreciate the progress that the field had made since Spearman, note that logically neither stance makes much sense as long as intelligence still remains undefined.

In his *Harvard Educational Review* paper, Jensen proclaimed that previous attempts to narrow the black/white gap on IQ tests were doomed to failure because, according to him, blacks are deficient in the particular genes required for complex information processing. Although this message initially met with some skepticism, it slowly mutated into the new received view, virtually uncontested by psychometricians and statisticians and deemed worthy of a full-page ad in the *Wall Street Journal* that several prominent psychometricians signed.

Undeterred by his detractors, Jensen set out to back up his unorthodox thesis with extensive empirical evidence. He tabulated IQ scores for various ethnic groups, designed an apparatus for measuring complex reaction times, and contributed suggestions on how best to measure the genetic portion of the IQ test variance. In the end, all these efforts converged on the same conclusion: Spearman had been right all along—g does exist and intelligence can be measured. In addition, it is ubiquitous and of utmost importance in virtually all spheres of life. Just as the early pioneers had claimed, it is primarily genetically predetermined. True, blacks are disadvantaged in this respect. However, this is no cause for alarm, as long as costly but futile efforts to bring them up to the level of whites (such as Head Start) are replaced with more realistic alternatives to guard against "dysgenic trends" toward "genetic enslavement" (Jensen, 1969, p. 91f).

The psychometric and statistical establishment initially reacted with quiet scorn to these heresies, fearing perhaps unwanted public scrutiny of psychometrics more generally. What it did not do, in any case, was to squarely face up to Jensen's challenge and simply show what was wrong with his reasoning.

Jensen's g Ersatz versus Spearman's g

What was wrong with it was that Jensen, who greatly admires Spearman, had quietly abandoned Spearman's original theory by replacing his g factor with the first principal component (PC1): "For example, the g [the first principal component] extracted from the six verbal subtests of the Wechsler Adult Intelligence Scale has a correlation of 0.80 with the g extracted from the five Performance subtests, in the standardization sample" (Jensen, 1979, p. 17).

To understand why this seemingly innocuous change actually amounts to a grotesque perversion of Spearman's theory, one has to know how principal components differ from common factors. The technical difference is most easily demonstrated with recourse to elementary matrix algebra. Since a correlation matrix of p tests is symmetric, its eigenvalues are always real and can be ordered. The dominant eigenvector of R, if used as a weight vector, results in a linear combination, PC1, that has largest variance among all possible linear combinations of the observed tests (subject to the constraint that the weight vector be of unit length). Note that this is a definition of a PC1, not an empirical discovery. The variance of the resulting PC1 is given by the dominant eigenvalue c.

Artifacts

Now let us see what happens when the correlation matrix R is "positive" throughout—that is, it has only positive elements, which was the point of departure for Spearman. For the sake of argument, let us assume all correlations are equal to $r\,(>0)$. This greatly simplifies the algebra without unduly violating reality. In this case, R can be written as a sum, $R = r\mathbf{e}\mathbf{e}' + (1-r)I$, where r is the correlation, $\mathbf{e}$ is a column vector of p ones, and I is the identity matrix of order p. The left-hand summand is a matrix of rank 1. It has dominant eigenvalue $r\mathbf{e}'\mathbf{e} = pr$ [since $R\mathbf{e} = (r\mathbf{e}\mathbf{e}')\mathbf{e} = r(\mathbf{e}'\mathbf{e})\mathbf{e} = pr\mathbf{e}$], while all other eigenvalues are zero. The right-hand summand, $(1-r)I$, leaves all eigenvectors intact and simply adds $1-r$ to all eigenvalues. Hence, the dominant eigenvalue of R is

$pr + (1 - r)$, which equals the variance of the PC1. The remaining $p - 1$ eigenvalues are all equal to $1 - r$.

As a result, the percentage of variance of the observed tests accounted for by the PC1 of R is $100[r + (1 - r)/p]$, which tends to $100r$ as p increases. The ratio of the largest eigenvalue to the next largest eigenvalue is $pr/(1 - r) + 1$. This ratio quickly increases as the number of tests, p, increases. Concretely, if $p = 10$ and $r = 0.5$, then this ratio is $p + 1 = 11$. If $p = 20$ and $r = 0.33$, it also is 11. This means that a PC1 always explains much more variance in a positive data matrix R than any of the remaining $p - 1$ PCs, as soon as p is reasonably large.

This is the reason why Jensen was so successful in convincing the uninitiated that "g" (as he calls the PC1 in gross abuse of language) is a powerful variable wherever he looks. The problem is that a PC1 is not g as Spearman had defined it in 1904. Partialling out the PC1 does not leave zero partial correlations. Rather, every one of Jensen's PC1s is a local description of the data at hand. Spearman, in contrast, was searching for a g that underlies all intelligence test batteries, not just a particular one. Jensen obtains a different g every time he analyzes a new R.

Congruence Coefficients

Jensen finesses the question whether all his g's are the same by pointing to so-called congruence coefficients as "a measure of similarity" between the dominant eigenvectors extracted from different batteries: "*Congruence coefficients (a measure of factor similarity)* are typically above 0.95, indicating virtually identical factors, usually with the highest congruence for the g factor" (Jensen, 1998, p. 363, italics added).

Usually these cosines are high. The problem is that they tell nothing whatsoever about "factor similarity": Two sets of variables can have identical within-set correlations while all between-set correlations are zero. In this case, the congruence coefficient will be 1 and the correlation between both PC1s will be zero.

This is Jensen's second sleight of hand: It is not at all difficult to tabulate the cosines between dominant eigenvectors extracted from randomly generated positive R's and a vector of 1's (**e**) of the same order. When this was done, the cosines varied between 0.995 and 0.999 when the parent distribution was uniform and between 0.949 and 0.998 when it was chi-square 1 (Schonemann, 1997a, p. 806). This means that Jensen's g ersatz is simply the average test score of whatever tests he analyzes. It is not g, but a travesty of Spearman's g.

Unfortunately, there are few examples in the psychometric and, more significantly, the statistical literature of anyone seriously challenging Jensen on methodological grounds (rare exceptions are Kempthorne and Schonemann). Typically, he is challenged on ideological grounds because critics do not like his conclusions.

Retrospective

Looking back, it is difficult not to notice that psychometrics did not live up to the promise of its auspicious beginnings. After Thurstone had shifted the focus from substantive theory to general scientific method, the field progressively lost its moorings and began to drift toward "ideational stagnation" (Sternberg and Williams, 1998, p. 577). On the applied side, steadily declining technical standards eventually empowered prophets of dysgenic doom to revive the eugenic myths of the 1920s, virtually unopposed by psychometricians and statisticians alike. On the theoretical side, Thurstone's implied message "that knowledge is an almost automatic result of gimmickry, an assembly line, a methodology" (Koch, 1969, p. 64) easily won out against Guttman's stern admonition that the essence of science lies in painstaking replication, not in facile significance tests. Numerous "general scientific methods" came and went, ranging from multidimensional scaling ("a picture is worth more than a thousand words") to latent trait theory, item response theory, linear structural models (LISREL), and meta-analysis and log-linear models. None of them has markedly improved the practical utility of mental tests, let alone helped answer any of the basic theoretical questions, such as "What is intelligence?" Just as Spearman noted years ago, "Test results and numerical tables are further accumulated; consequent action affecting the welfare of persons are proposed, and even taken, on the grounds of—nobody knows what!" (Spearman, 1927, p. 15).

See Also the Following Articles

Binet, Alfred • Education, Tests and Measures in • Human Growth and Development • Intelligence Testing • Psychological Testing, Overview • Thurstone's Scales of Primary Abilities • Thurstone, L.L.

Further Reading

Brandt, C. (1985). Jensen's compromise with componentialism. *Behav. Brain Sci.* **8,** 222–223.

Brigham, C. C. (1923). *A Study of American Intelligence.* Princeton University Press, Princeton, NJ.

Brigham, C. C. (1930). Intelligence tests of immigrant groups. *Psychol. Rev.* **37,** 158–165.

Cronbach, L. J., and Gleser, G. C. (1957). *Psychological Tests and Personnel Decisions.* University of Illinois Press, Urbana.

Dodd, S. C. (1928). The theory of factors. *Psychol. Rev.* **35,** 211–234, 261–279.

Donlon, T. F. (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests.* College Entrance Examination Board, New York.

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common factor theory. *Br. J. Statistical Psychol.* **8,** 65–81.

Hart, B., and Spearman, C. (1912). General ability. Its existence and nature. *Br. J. Psychol.* **5,** 51–84.

Humphreys, L. G. (1968). The fleeting nature of the prediction of college academic success. *J. Educational Psychol.* **59,** 375–380.

Jensen, A. (1969). How much can we boost IQ and academic achievement. *Harvard Educational Rev.* **39,** 1–123.

Jensen, A. (1979). g: Outmoded theory or unconquered frontier? *Creative Sci. Technol.* **3,** 16–29.

Jensen, A. (1998). *The g Factor. The Science of Mental Ability.* Praeger, Westport, CT.

Kempthorne, O. (1978). Logical, epistomological and statistical aspects of nature–nurture. *Biometrics* **34,** 1–23.

Koch, S. (1969). Psychology cannot be a coherent science. *Psychol. Today* **14**, 64.

Schonemann, P. H. (1997a). The rise and fall of Spearman's hypothesis. *Cahier Psychol. Cognitive/Curr. Psychol. Cognition* **16,** 788–812.

Schonemann, P. H. (1997b). Some new results on hit-rates and base-rates in mental testing. *Chinese J. Psychol.* **39,** 173–192.

Schonemann, P. H. (1997c). Models and muddles of heritability. *Genetica* **99,** 97–108.

Sternberg, R. J., and Williams, W. M. (1997). Does the Graduate Record Examination predict meaningful success in graduate training of psychologists? A case study. *Am. Psychologist* **52,** 630–641.

Sternberg, R. J., and Williams, W. M. (1998). You proved our point better than we did. A reply to our critics. *Am. Psychologist* **53,** 576–577.

Thurstone, L. L. (1947). *Multiple Factor Analysis.* University of Chicago Press, Chicago.