Psychometrics

Psychology (Statistics) 484

Statistics, Ethics, and the Social and Behavioral Sciences

June 15, 2013

Beginning Quotations

We must distinguish between Spearman's method of analyzing the intercorrelations of a set of variables for a single common factor and his theory that intelligence is such a common factor which he calls 'g'.
– L. L. Thurstone, *The Theory of Multiple Factors* (1933)

One of the great mistakes is to judge policies and programs by their intentions rather than their results.
– Milton Friedman

I'm thirty years old, but I read at the thirty-four-year-old level.
– Dana Carvey

We've upped our standards. Up yours.
– Pat Paulsen (campaign slogan)

Experience has sufficiently shown that the facts of human nature can be made the material for quantitative science.
– E.L. Thorndike, *An Introduction to the Theory of Mental and Social Measurements* (1904/1912)

Week 7: Psychometrics

— the settlement between the Educational Testing Services and the Golden Rule Insurance Company, and its devastating consequences for carrying out psychometrically defensible "high-stakes" testing

— the delay (because of the American Psychological Association) in the publication of the article, "Is Criminal Behavior a Central Component of Psychopathy?"; how construct validation should be done, and the need to define a construct by more than just the Hare Psychopathy Checklist

— the darker side of psychometrics and statistics: eugenics, forced sterilization, immigration restriction, racial purity laws

— Buck v. Bell (1927): The Court upheld a statute instituting compulsory sterilization of the unfit "for the protection and health of the state."

— Loving v. Virginia (1967): The Court declared Virginia's antimiscegenation statute, the Racial Integrity Act of Virginia (1924), unconstitutional, thereby ending all race-based legal restriction on marriage in the United States.

Required Reading:

SGEP (227-256) —
Traditional True Score Theory Concepts of Reliability and Validity
Test Fairness
Quotidian Psychometric Insights
Psychometrics, Eugenics, and Immigration Restriction
Eugenics
Immigration Act of 1924
Racial Purity Laws

Popular Articles —
Annals of Medicine: The Dictionary of Disorder, Alix Spiegel (*New Yorker*, January 3, 2005)
Personality Plus, Malcolm Gladwell (*New Yorker*, September 20, 2004)

Suggested Reading:
Suggested Reading on Psychometric Issues
Appendix: Excerpts From Brigham's *A Study of American Intelligence*
Appendix: Racial Integrity Act of 1924 (State of Virginia); Loving v. Virginia

Film:

The Loving Story (77 minutes)

Nazi Medicine (54 minutes)

Psychometrics is a branch of psychology that deals with the design, administration, scoring, and interpretation of objective tests developed for the measurement of psychologically relevant variables, such as aptitude, personality, job performance, or academic achievement.

Because of the socially relevant implications of high-stakes testing in areas such as job selection and promotion, licensure, college admission, accreditation, and graduation and other competency certification, the issues of fairness and discrimination surrounding testing are often discussed in the media, and increasingly, in the courts.

Campbell's Law

In discussing possible negative consequences of high-stakes testing in the classroom (such as that mandated by the 2001 "No Child Left Behind" Act), we should keep Campbell's (1975) law in mind:

The more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.

Campbell's law is generally relevant to any area where numbers drive some societal process and where "scamming the numbers" can lead to personal gain or glory.

One such recent instance occurred at the University of Illinois College of Law.

It involved law class profiles and how they were constructed and reported to various agencies by the Admissions Dean, Paul Pless (who resigned in November 2011).

True Score and Item Response Theory

Because of the importance of testing in many of our societal endeavors, it is a relevant statistical literacy skill to have some basic understanding of the concerns of both traditional true score theory (TST), and its emphasis on assessing reliability and validity, and the newer item response theory (IRT), where there is an attempt to scale jointly subjects and assessment items through variants of logistic regression models.

Once items are calibrated from some norming sample by estimating parameters characterizing the logistic regressions, the item pool can serve a number of purposes:

(1) subsets of items can be used to construct fixed-length tests with subjects' performance estimated by how they respond to the calibrated items;

(2) psychometrically equivalent tests can be constructed from the item pool, or distinct tests constructed from the pool can be equated, supposedly making it immaterial to a subject as to which test is actually taken;

(3) a computer adaptive testing (CAT) process can be initiated where items are administered according to how a respondent has answered earlier items.

In theory, CAT holds the promise of greater efficiency in testing by allowing instruments to be shorter and/or estimating the trait of interest better.

TST Concepts of Reliability and Validity

A reliable measure is one that measures something consistently.

In theory, as well as sometimes in practice, *test-retest* reliability can be assessed by computing the correlation between repeated administrations of the same instrument over the same subjects.

Or, if two equivalent tests are available, possibly constructed from an IRT-calibrated item pool and given to the same subjects, the correlation between such scores is an *equivalent-forms* reliability.

The homogeneity of a single test form is referred to as *internal consistency* and can be assessed by correlating performance on two halves of the test to give a *split-half* reliability;

this latter value can then be readjusted with the (nonlinear) Spearman–Brown prophecy formula to give the reliability of a full-length test.

More commonly, and as computed by all the commercial statistical software packages (for example, SYSTAT, SPSS, SAS), there is Cronbach's alpha, the mean of all possible split-half coefficients.

If we formally define reliability within a true score plus error model as the ratio of true score variance to observed score variance, alpha can be considered a lower-bound to reliability.

Validity refers to the extent to which the score on an instrument represents what it is supposed to, thus justifying the inferences that are made from that score.

It is usually evaluated by a process. Given other criterion measures that may be available, *concurrent* validity refers to the correlation with measures collected at the same time;

predictive validity to measures collected later;

and *construct* validity to relating the current measures to others that should be part of its theoretical context.

Content validity is more direct and refers to items demonstrably coming from the domain we wish to assess, or that are *bona fide* skills necessary to perform some task.

Construct Validity

As discussed by Cronbach and Meehl in their justifiably well-known 1955 article, "Construct Validity in Psychological Tests" (*Psychological Bulletin*, *52*, 281–282), the most difficult form of validity to establish is construct validity.

In lay terms, a validation process has to be put into place to argue effectively that we are really measuring the construct we think we are measuring.

A recent example of the difficulties inherent in construct validation has appeared in the popular media, and involves the notion of psychopathy—a personality disorder indicated by a pattern of lying, exploitativeness, heedlessness, arrogance, sexual promiscuity, low self-control, and lack of empathy and remorse, all of this combined with an ability to appear normal.

Test Fairness

Some confusions about the fairness of a test continually reappear in the popular culture that could be clarified with an educated understanding of how tests are generally constructed and intended to function.

Because one particular definable subgroup performs less well on a test than another does not automatically imply that the test is "biased" or "unfair" to the subgroup.

We quote part of a passage from Linn and Drasgow (1987) that makes this point particularly well:

[P]sychological tests *do not measure innate abilities or aptitudes*. Instead, they assess a test taker's current repertoire of knowledge and skills. An individual's repertoire of knowledge and skills is certainly affected by his or her "environment"; consequently, we would expect differences in mean test performance for groups whose environments differ. ... The key point is that *unequal environments imply unequal educational achievements and a well-constructed test should reflect this fact*. (latter italics added)

As another way of phrasing this argument, if definable subgroups differ in background environments that should be related to what a test is measuring, similarity of performance for subgroups does not necessarily indicate a fairer or less biased instrument; to the contrary, it is more reflective of a test that may not be very good.

Race-norming

A somewhat different take on the issue of fairness in the use of tests is the practice of "race-norming," the adjustment of scores on a standardized test through separate transformations for different racial groups.

This type of within-group norming was promulgated in the 1980s as a way of reaching federal equal employment opportunity and affirmative action goals;

it was particularly encouraged by the Department of Labor and the United States Employment Services in its promotion of the General Aptitude Test Battery (GATB) to state employment services.

Irrespective of whatever admirable social goals were envisioned with the use of race norming, the Civil Rights Act of 1991 prohibits discriminatory use of test scores. The new Title VII provision, section 106, reads in full:

It shall be an unlawful employment practice for a respondent, in connection with the selection or referral of applicants or candidates for employment or promotion, to adjust the scores of, use different cutoff scores for, or otherwise alter the results of employment related tests on the basis of race, color, religion, sex or national origin.

Quotidian Psychometric Insights

First, difference or gain scores tend to be unreliable and more so the higher the correlation between the constituent scores.

Given the attenuating characteristics of unreliability and the general reduction in power, it may be generally problematic to use gain scores in empirical investigations.

This is a major hurdle for teacher evaluation programs that propose basing merit increases on student gain scores, using "value-added models".

Second, the standard error of a correlation is about $\frac{1}{\sqrt{n}}$; thus, correlations based on small sample sizes need to be fairly large to argue that they reflect more than random variability for a population where the true correlation might, plausibly, be zero.

Third, constructing a test by item selection to optimize an index such as Cronbach's alpha will tend to produce a homogeneous test satisfying a single common factor model.

This is because items having correlations of similar size with other items will be culled, and alpha increases with the average correlation between items.

However, the implication in the other direction is not correct, that a high value for alpha reflects unidimensionality;

unidimensionality depends on the pattern of inter-item covariances, whereas the value of alpha depends on their size.

Fourth, as discussed at some length by Thurstone in *The Theory of Multiple Factors* (1933), the computations suggested by Spearman for the single common factor model need separation from Spearman's theory that intelligence is such a common factor, referred to as "g."

In fact, one might go on to observe that "g" may just be an artifact of correlations generally being positive among intellective-type tests;

here, as stated by the Perron–Frobenius Theorem, weights defining the first principal component must all be positive, providing a simple weighting of the tests.

To reify this weighted composite as "g," without additional supporting evidence, seems a bit of a stretch.

Stephen Jay Gould's popular book, *The Mismeasure of Man* (1996), takes on "the argument that intelligence can be meaningfully abstracted as a single number capable of ranking all people on a linear scale of intrinsic and unalterable mental worth"

Along the way, several observations about factor analysis are made that deserve reporting:

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.

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.

From the midst of an economic depression that reduced many of its intellectual elite to poverty, an America with egalitarian ideas (however rarely practiced) challenged Britain's traditional equation of social class with innate worth. Spearman's g had been rotated away, and general mental worth evaporated with it.

Fifth, the unreliability of a measure has a number of statistical consequences: it attenuates the correlation with other tests so any study of a measure's validity needs to take this into consideration;

it reduces the power of common statistical analyses, such as analysis of variance;

it biases regression coefficients in regression models.

In all cases, the more unreliable a measure, the more difficult it will be for it to have value in any scientific or practical investigation.

Sixth, the standard TST idea of reliability can be put into the framework of random effects analysis of variance.

These components-of-variance models are commonly discussed in the year-long graduate sequence in statistics, and extensions are now being touted as the "next big thing" in hierarchical linear modeling (HLM).

In turn, components-of-variance models lead to the estimation of heritability coefficients in behavioral genetics.

Here, the "observed score" variance being additively decomposed as "true score" variance plus uncorrelated "error" variance, is reinterpreted as "phenotypic variance," being the sum of "genotypic variance" and uncorrelated "environmental variance."

Thus, some basic proficiency in how reliability is defined and approached generally should also help in reasoning with the sometimes controversial topic of heritability (because the various heritability coefficients are just ratios of true-score to observed-score variances), and thinking through how HLM studies are reported, with the inclusion of random effects to account for the type of hierarchical sampling that has occurred.

A more extensive discussion of heritability appeared earlier in emphasizing intraclass correlations in Week 4 on Correlations.

Finally, although personality tests may be subject to the same TST ideas of reliability and validity, the status of IRT modeling in personality is somewhat different than for intellective measurement.

For the latter, it may be reasonable for a positive response probability to increase consistently with an increase in the latent quantity being estimated.

These "dominance" models have IRT functions that increase monotonically with respect to the trait being assessed.

Personality items, on the other hand, may be better represented by unfolding models characterized by single-peaked and nonmonotonic IRT functions centered at a subject's presumed ideal point;

Psychometrics, Eugenics, and Immigration Restriction

Psychometrics and related statistical analyses and interpretation have played major parts in several ethically questionable episodes in United States history.

This section discusses three such situations:

(a) the eugenics movement from the first half of the 20th century and the enforced sterilization of those deemed "unfit" or "feebleminded," and who therefore had to be prevented from passing along such a trait and related ones to their children (for example, moral depravity, criminality, shiftlessness, insanity, poverty);

(b) the use of psychological test data from Army draftees and recruits during World War I to justify racially restrictive United States immigration policies, particularly the Immigration Act of 1924.

Here, the main purveyor of flawed statistical analyses was Carl Brigham's, *A Study of American Intelligence* (1923).

This same Carl Brigham later developed the Scholastic Aptitude Test (SAT) that helped launch the present-day Educational Testing Services;

(c) the use of test data and associated analyses to justify the continuance of laws against interracial marriage (miscegenation);

the prime example is the Racial Integrity Act of Virginia (1924).

Eugenics

The word *eugenics* was coined by Sir Francis Galton in 1883 using the Greek word "eu" (good or well) with the suffix "genēs" (born).

Galton (1908) characterized eugenics as the "study of agencies under social control that may improve or impair the racial qualities of future generations".

The search was for ways to improve the human gene pool, either through positive eugenics (by having the "best" marry and reproduce with the "best"), or negative eugenics (through, for example, segregation and colonization, sterilization, or euthanasia).

The popular eugenics movement in the early decades of the 20th century, both in the United States and England, was heavily influenced by statisticians and biometricians.

In the United States, Charles Davenport was the most prominent as the director of Cold Spring Harbor Laboratory (1910) and founder of the Eugenics Record Office.

In England, besides Galton there were the inaugural editors of *Biometrika*, W.R.R. Weldon and Karl Pearson (a third inaugural editor was the U.S.-based Davenport).

The subtitle for the journal was "a journal for the statistical study of biological problems," and reflected an obvious emphasis on biometry and the newly rediscovered Mendelian genetics.

Other instances could be given for the role that some of our more famous psychometricians have played in justifying eugenic sterilization.

We give one particularly late (and nasty) quotation from E. L. Thorndike's, *Human Nature and the Social Order* (1940):

By selective breeding supported by a suitable environment we can have a world in which all men will equal the top ten per cent of present men. One sure service of the able and good is to beget and rear offspring. One sure service (about the only one) which the inferior and vicious can perform is to prevent their genes from survival.

Henry Goddard and the Kallikak Family

Henry Goddard was a well-known (clinical) psychologist from the first half of the 20th century.

He was the first to distribute widely an English translation of the Binet intelligence test first developed in France and introduced the term "moron" to label people with IQs from 51 to 70.

The range of 0 to 25 was reserved for "idiots," and 26 to 50 for "imbeciles."

The more ambiguous and widely used term of "feebleminded" referred generally to those mental deficiencies that now might be considered various forms and grades of mental retardation or labeled as learning disabilities.

Feeblemindedness might be assessed by a poor performance on a Binet test, or more commonly, by observation from a trained (always female) field worker, and often just by remembrances from others about individuals long dead.

As noted in the introduction to Goddard's famous study, *The Kallikak Family: A Study in the Heredity of Feeble-mindedness* (1912), the field worker responsible for the assessment of feeblemindedness in the Kallikaks was Elizabeth Kite.

Goddard was best known for postulating that feeblemindedness was a hereditary trait, most likely caused by a single recessive gene, and thereby subject to the laws of Mendelian inheritance that had been rediscovered at the turn of the century.

A main argument for the hereditary nature of feeblemindedness was the extensive case study of the Kallikak family, and the genealogy of the family's founder, Martin Kallikak.

Martin, a Revolutionary War hero, was on his way home from battle, but stopped to dally once with a "feebleminded" barmaid.

Martin went on to a morally upright life, marrying a Quaker woman and siring a large and prosperous New England family.

The child who was the product of the single dalliance went on to establish another branch of the Kallikak family; this branch consisted mainly of those assessed as feebleminded (apparently, all by Elizabeth Kite).

Goddard argued that the Kallikak study documented a natural (observational) experiment in the heritability of the lack of intelligence and its associated traits of morality and criminality.

The study of the Kallikak family was a popular and widely read cautionary tale about the perils of unfettered reproduction, and the importance of establishing a national eugenics policy so that feeblemindedness could not be passed on to future generations.

Goddard himself argued for segregation into colonies; other eugenicists used the Kallikak study and other similar statistical arguments in the passage of laws involving forced sterilization and restricted immigration.

Once again, we are reminded of the speciousness of naming something without really understanding it, and then being seduced by a proposed but unproven mechanism, this time that feeblemindedness was carried by a single recessive gene operating just like the smoothness or wrinkliness of the peas studied by the Austrian monk, Mendel.

It might be prudent to keep in mind a quotation from John Stuart Mill (1869):

The tendency has always been strong to believe that whatever received a name must be an entity or being, having an independent existence of its own. And if no real entity answering to the name could be found, men did not for that reason suppose that none existed, but imagined that it was something peculiarly abstruse and mysterious.

Harry Laughlin and Buck v. Bell

he Eugenics Record Office founded by Charles Davenport was directed by Harry Laughlin from its inception in 1910 through to its closing in 1939.

Laughlin figures prominently in the eugenics movement, particularly as an advocate of forced sterilization to eliminate the possibility of reproduction for "unfit" members of society.

Laughlin constructed a "model law" for compulsory sterilization that he believed would surpass all constitutional challenges.

Based on Laughlin's model, Virginia enacted such a sterilization law in 1924 that provided for the compulsory sterilization of persons deemed to be "feebleminded" including the "insane, idiotic, imbecilic, or epileptic".

The first person ordered sterilized under Virginia law was Carrie Buck on the grounds that she was the "probable potential parent of socially inadequate offspring."

This carefully chosen test case would go all the way to the Supreme Court as *Buck v. Bell* (1927), and result in one of the most notorious decisions ever handed down by the Court.

The opinion in this 8 to 1 decision was written by the well-known jurist Oliver Wendell Holmes, Jr.

We give two items in appendices to this chapter:

part of the deposition given by Laughlin about Carrie Buck's suitability for sterilization (given in a book by Harry Laughlin entitled *The Legal Status of Eugenical Sterilization* [1930]);

and a redacted Supreme Court opinion in *Buck v. Bell* written by Holmes that contains the famous phrase, "three generations of imbeciles are enough."

Immigration Act of 1924

Harry Laughlin, the head of the Eugenics Record Office met briefly in the discussion of *Buck v. Bell*, had another major success in 1924—the Johnson-Reed Immigration Act.

Laughlin provided extensive statistical testimony to the United States Congress based primarily on the statistical analyses of Carl Brigham that we discuss below.

Brigham's empirical interpretations were phrased within the racial ideology espoused by Madison Grant in *The Passing of the Great Race* (1916).

Laughlin went on to be appointed as an "expert eugenics agent" to the Congressional Committee on Immigration and Naturalization.

The Immigration Act of 1924 set an initial yearly quota on immigration of 165,000, less than 20% of the pre-World War I average.

Furthermore, it based ceilings on the number of immigrants from any particular area by the percentage of each such nationality recorded in the 1890 census.

Because most immigration from Southern and Eastern Europe occurred after 1890, it reduced to a trickle those people coming from two of the Caucasoid races (Alpine and Mediterranean according to the categories used by Madison Grant) and encouraged a Nordic influx.

This type of quota system on immigration remained in effect until the 1960s

Carl Brigham

Carl Brigham was a psychologist at Princeton University in the 1920s and 1930s, but earlier had collaborated with Robert Yerkes on the development of the Army Alpha and (the nonverbal) Beta intelligence tests given to well over a million United States Army recruits during World War I.

Based on these data collected on recruits, Brigham published *A Study of American Intelligence* (1923), which quickly become an influential source for justifying the passage of the Immigration Act of 1924 and in popularizing and justifying the eugenics movement in the United States.

The data and results were placed within the context of Nordic theory as espoused by Madison Grant and his hugely successful *The Passing of the Great Race* (1916).

Nordic theory contends that the Caucasoid (European) race should be subdivided further into Nordic (Northern Europe), Alpine (Central Europe), and Mediterranean (Southern Europe).

Based on the Army tests, Brigham concluded that the Nordic race was intellectually superior and argued that immigration should be tightly controlled to prevent Alpine and Mediterranean immigration, and to thereby protect "American Intelligence."

He was particularly concerned with miscegenation between blacks and whites, and viewed "Negroes" as the most inferior intellectually.

The quality of the data with which Brigham had to work was extremely poor and subject to a variety of confoundings based on language and the reliance on current American cultural knowledge.

Carl Brigham in a 1930 *Psychological Review* article, "Intelligence Tests of Immigrant Groups," repudiated his whole study of American intelligence.

Unfortunately, this came rather late; the damage had already been done in its justification for the restrictive Immigration Act of 1924 and other eugenic initiatives.

Here's his conclusion:

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.

Here's another unpublished statement from Brigham in 1934 that is pretty definitive:

The test movement came to this country some twenty-five or thirty years ago accompanied by one of the most glorious fallacies in the history of science, namely that the test measured *native intelligence* purely and simply without regard to training or schooling. I hope nobody believes that now. The test scores very definitely are a composite including schooling, family background, familiarity with English, and everything else, relevant and irrelevant. The *native intelligence* hypothesis is dead.

Sacco and Vanzetti

As apparent from the wide support for immigration restriction in the period after World War I, the thought of a major influx in the Alpine and Mediterrean races truly alarmed many Americans.

This prejudice is manifest in one of the most famous trials of the 20th century, that of Sacco and Vanzetti. Ferdinando Sacco and Bartolomeo Vanzetti were immigrant anarchists convicted of a 1920 murder of two men during an armed robbery in South Braintree, Massachusetts.

The trial and successive appeals were highly politicized and controversial; they were finally executed on August 23, 1927. To give a sense of the continuing historical importance of these trials, we give part (in the end notes of the Chapter) of the Wikipedia entry on Sacco and Vanzetti that deals with the proclamation of Massachusetts Governor Dukakis in 1977.

Racial Purity Laws

Miscegenation laws enforce racial segregation at the level of marriage and intimate relations by criminal sanction, possibly including sex between members of two different races.

In the United States, for example, such laws have been around since the late 17th century.

The most famous was enacted more recently as the Racial Integrity Act of Virginia in 1924 (a particularly good year, it seems, for the passage of racial and eugenic laws).

In *Loving v. Virginia* (1967), the Virginia law was declared unconstitutional by a 9 to 0 vote of the Supreme Court, thus ending all such race-based legal restriction in the United States.

We give excerpts from two items in an appendix to this chapter.

The first is from the Virginia Racial Integrity Act itself that enacts the "one-drop rule," where any amount of African ancestry leads to a classification as "black."

Note, however, the "Pocahontas exception" allowing up to one-sixteenth of American Indian blood with a permissable label of "white" (presumably, because of the genealogy of some prominent Virginians in 1924).

The second set of excerpts provides part of the unanimous opinion in *Loving v. Virginia* written by then Chief Justice Earl Warren.