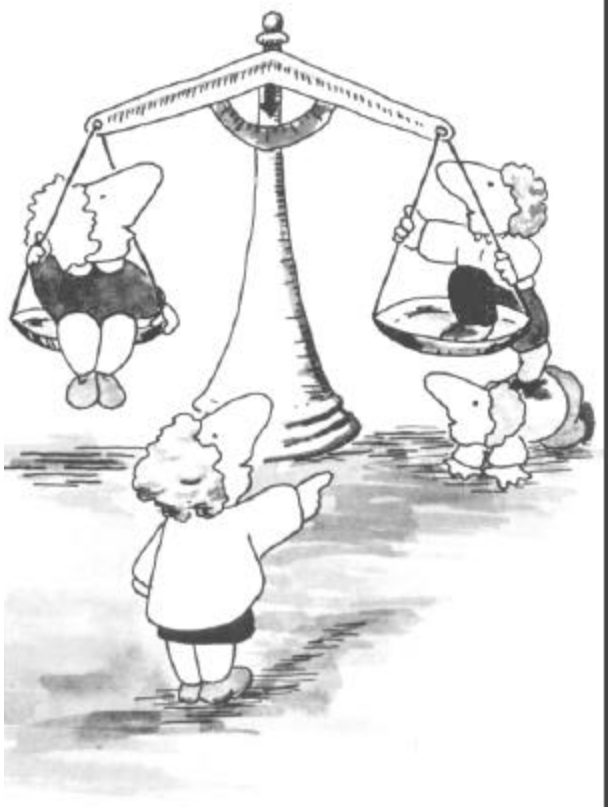
Overcoming Test Anxiety:

Measurement and Statistics for Lawyers and Others



Patricia B. Campbell, Ph.D.

Collaboration for Equity Fairness in Science and Mathematics Education

WHY AM I READING THIS?

Familiar tests such as the SATs, GREs and LSATs have been joined in recent years by state mandated tests, exit exams and even by proposed national tests in reading and mathematics. As tests play larger roles in educational admissions, graduation and licensing as well as in scholarship awards, there is increased controversy and litigation about their appropriate use.

It is essential that lawyers and others concerned with testing issues have an understanding of basic concepts and statistics related to testing. Our purpose here is to provide background information for anyone interested in the role of testing in public policy and/or in the law.

VALIDITY AND RELIABILITY: THE CORE OF TEST DESIGN

VALIDITY means that a test is "valid"-- that it measures that which it is supposed to measure. Forms of validity include:

CONTENT VALIDITY: which asks "Does the test cover the appropriate subject matter?"

PREDICTIVE VALIDITY: which asks "Does the test score predict that which it purports to?"

CONCURRENT VALIDITY: which asks "Do the results of this new measure reflect the results of known validated measures?"

CONSTRUCT VALIDITY: which asks "Do the results of this measure reflect existing theory?"

In much of educational decision-making we are primarily concerned with **CONTENT VALIDITY** and **PREDICTIVE VALIDITY**. Since exit exams are set up to determine if the student has enough knowledge and skills to graduate, these exams need to have good **CONTENT VALIDITY,** that is**,** the content they cover needs to reflect the knowledge and skills students are expected to have**.** Admissions tests are set up to determine how well someone will do in the future. Thus they need good **PREDICTIVE VALIDITY.**

RELIABILITY means that a test is reliable. It is the degree to which tests measure consistently whatever it is they are measuring, be it skills, knowledge or personality. Forms of reliability include:

TEST/RETEST RELIABILITY: the degree of similarity and difference between an individual's score the first and the second time they take a test**.**

ALTERNATE FORM RELIABILITY: the degree of similarity and difference between an individual's score on different forms of the same test.

INTER-RATER RELIABILITY: the degree of similarity and difference of the scores assigned to the same answer or response by different people (or raters).

SPLIT-HALF RELIABILITY: the comparison of scores on half the test with the other half to determine if the test is consistently measuring the intended concept.

Also related to reliability is the **STANDARD ERROR OF MEASUREMENT (SEM**). **SEM** is the range of fluctuation an individual's score can have as a result of irrelevant factors, such as the testing environment. If a test has a SEM of 7 and an individual scores 50, the odds are the person's "true" score is somewhere between 43 and 57.

Production of this material was made possible by a grant from the National Science Foundation. Opinions expressed are those of the authors and not necessarily those of the funders.

A STATISTICS VOCABULARY

WAIT! Your first impulse may be to skip this section, either because you already know this, or because you feel you never will, but READ ON. There are many misconceptions about what statistics are, but basically they are just useful ways to describe and analyze large amounts of information. Knowing even a little about statistics will help you better understand measurement issues in major testing controversies.

MEAN, commonly known as the "average," is the sum of a set of scores divided by the number of scores.

STANDARD DEVIATION (SD) is an indication of how varied or spread out a set of scores is. If a set of scores has a mean of 10 and a standard deviation of 1, most of the scores are very close to 10 with about two thirds of them between 9 and 11. If the standard deviation is 5, the scores will be more spread out; with about two thirds of scores between 5 and 15.

CORRELATIONS range from 0 (no relationship) to 1 (a perfect relationship). The relationship between birth and death is a perfect 1, meaning once you are born it can be predicted with certainty that you will die. The closer the relationship is to 1 the better the prediction. A correlation of .6 is high, a correlation of .1 is low. **RELIABILITY** is usually measured using a correlation.

REGRESSION is an equation or set of equations developed to predict an outcome, such as first year college grades, based on variables you already know about individuals such as their high school grades, SAT scores and socioeconomic status. Sometimes in order to get the best predictions, the regression equations may be different for different groups (e.g. women and men; African Americans and whites; private and public school students).

OTHER TERMS TO REMEMBER

RAW SCORES describe how many questions, also known as items, an individual got correct**.**

NORM REFERENCED TESTS are the standardized test equivalent of grading on a curve. The score is not based on how much you know, but on how much more, or less, you know than others who have taken the test. Most admission tests (SAT, LSAT, Miller Analogies) are norm referenced. They report the percent of test takers scoring better (or worse) than you, rather than reporting how many items you got correct or how much you know.

STANDARD SCORES are raw scores that have been transformed into norm referenced scores. They don't describe the number correct, they describe how far an individual is above or below the mean. For example SAT:M and SAT:V are standard scores ranging from 200 to 800 with means of 500. An SAT:M score of 500 indicates the person is at the 50^{th} percentile while a score of 600 indicates a person is in the 85^{th} percentile..

STATISTICAL SIGNIFICANCE is the odds that differences among groups are real and would be found if the study were done again with similar groups. A significance level of $p<.05$ means that the probability (p) or odds that a difference is real are at least 95 out of 100. Knowing that a difference is **STATISTICALLY SIGNIFICANT** does not necessarily mean that it is important or meaningful; it means that the difference is probably real and not an anomaly.

DIFFERENTIAL ITEM FUNCTIONING (DIF) is a process for identifying test items on which some groups (i.e. girls; Hispanics) score disproportionately lower than others. Items where the score differences among different groups are much larger than overall test score differences of those groups are seen as suspect and are usually eliminated or rewritten.

A NEW LOOK AT TESTING CONTROVERSIES[1]

ADMISSIONS TESTING AND AFFIRMATIVE ACTION

There has been a great deal of discussion and litigation about the role of admissions tests in student selection. Often the discussion has assumed that admissions tests are the fairest way to select students and that the higher the admission test score the better the student. Efforts to include other variables, such as rank in class, in admissions decisions have at times been viewed as affirmative action or a lowering of standards.

Admissions tests are designed to have **PREDICTIVE VALIDITY**, to predict success in high school, college or graduate school. They are not designed to measure intelligence, potential or academic achievement. For example, the Scholastic Achievement Test (now called the SAT I) and the admissions test given by American College Test (ACT) were designed to predict success in college as measured by first year college grades.

The correlation between SAT scores and first year college grades is about .5. The correlation between high school grades and first year college grades at .48 is about the same. With a correlation of about .59, the combination of high school grades and test scores is a better predictor of college grades than either are separately (College Board Website, Feb. 17, 1998).

Using test scores alone is not better than using both test scores and grades for admission; it actually is slightly worse. For many years test developers have recommended that test scores alone not be used for admission decisions. Good measurement practice calls for test scores to be used with other variables, including grades, in admissions decisions.

POOR TESTS OR POOR EDUCATION

As exit exams become more common at many educational levels, there is much concern about whether the tests are biased against minority students. If minority students are passing at a much lower rate than white students, the quality of the test is often questioned. The test indeed may be biased. However it may also be that the test is accurately reflecting the results of an educational system in which minority students are not receiving the education they need.

Good exit exams must have strong **CONTENT VALIDITY,** accurately reflecting the academic goals of the school and the skills and content areas covered in that school. They also need to have high **TEST/RETEST RELIABILITY**, to ensure that individual scores are accurate. It is important to see how exit exam scores correlate with grades. If they don't correlate highly, then grades and the exit exam are measuring different things.

Before condemning a test, it is important to check to see if all students taking the test have the opportunity to learn the content and skills needed for the test, and if students scoring poorly on the test have different classes and/or fewer educational resources than students who score well. If students don't have the opportunity to learn that which is covered in the test, the test is not a valid measure for them. The problem may be in the educational system.

[1] Many of the examples and the research cited are from the Educational Testing Service. Their willingness to share the results of their work with others is greatly appreciated.

HIGH STAKES TESTING AND HOW TESTS ARE CONSTRUCTED

We have known how to create valid measures without gender differences since at least 1942 when the Stanford Binet Intelligence test was revised to "produce a scale which will yield comparable IQs for the sexes." After initial testing found women tended to score higher than men, the authors concluded that "intellect can be defined and measured in such a manner as to make either sex appear superior" (McNemar, 1942).

The choice of item type, item content and even the multiple choice responses used can influence the size of the gender differences in test scores. Girls tend to do better on essay questions; boys on multiple choice. Boys tend to do better than girls when one of the multiple choice options is "not enough information is given to answer the question" or when reading comprehension passages are about science (Bridgeman and Schmidt, 1997). Based on how the tests are constructed, two tests may have equal **CONTENT** or **PREDICTIVE VALIDITY,** but one of the tests may reflect or result in gender differences and the other not.

In response to charges that the Preliminary Scholastic Aptitude Test (PSAT), the basis for initial selection for the National Merit Scholarships, was biased against girls, a multiple choice test on writing was added to the test. The addition of the writing test to the PSAT significantly reduced gender differences favoring males while the test remained valid.

DIFFERENT WAYS TO PREDICT SIMILAR OUTCOMES

When using test scores to decide who to admit or who is most likely to succeed, the obvious thing to do is use the same cut off score for everyone, to judge everyone on the same standard. However, sometimes the obviously "fair" choice is not fair.

For example the SAT:M tends to overpredict men's grades in college math courses and underpredict women's grades. When women and men have the same SAT:M score, the women tend to have higher grades in college math courses than do the men, even when the courses are the same (Wainer and Steinberg, 1992). To have the SAT:M be equally successful in predicting women and men's first year college math grades you could use a lower cut off score for women than for men or you could revise the test so that the same score predicts equally well for women and men.[2]

Currently, grades are the predictor variable, how success in high school, college, or graduate school is defined for admissions testing even though at all education levels, women tend to have better grades than men. There may be better ways of defining success than grades but if definitions of success are changed then so must be the tests used to predict that success. In measurement sometimes being unequal means being more accurate and thus more fair.

[2] The CRA of 1991 prohibits the use of different cut scores based on protected categories , i.e. race, sex, national origin. It does not apply to educational admissions.

MOVING ON

FOR FURTHER INFORMATION:

From a test critics' perspective: Fair Test: The National Center for Fair and Open Testing (342 Broadway, Cambridge, MA 02139; 617 864-4810) has a newsletter and a number of publications on bias in testing.

From a test developers' perspective: both the Educational Testing Service (Princeton, NJ 08541; 609 921-9000) and the American College Testing (Iowa City, Iowa 52243; 319 339-1000) have free newsletters and a number of publications on testing.

Bailey, Susan, Burbidge, Lynn, Campbell, Patricia B., Jackson, Barbara, Marx, Fern and McIntosh, Peggy. (1992*). The AAUW Report: How Schools Shortchange Girls*. Washington, DC: AAUW Educational Foundation and NEA.

Bridgeman, Brent and Wendler, C. (1989). Prediction of Grades in College Mathematics Courses as a Component of the Placement Validity of SAT-Mathematics Scores. New York: College Board Report No. 89-9.

Cole, Nancy. (1981). Bias in Testing. *American Psychologist*. 36, pp. 1067-1077.

College Board, (Feb. 17, 1998). Recentered SAT I Scores as a Predictor of College Grades. Http://www.collegeboard.org/index_this/sat/html/cournselors/stat003.html.

McNemar, Quin. (1942). *The Revision of the Stanford-Binet Scale.* Boston: Houghton Mifflin.

National Commission on Testing and Public Policy. (1990). *From Gate Keeper to Gateway: Transforming Testing in America*. Chestnut Hill, MA: National Commission on Testing and Public Policy, Boston College.

Wainer, Howard and Lisa Steinberg. (1992). Sex differences in performance in the mathematics section of the Scholastic aptitude Test: A Bidirectional Validity study. *Harvard Educational Review*. 62. pp. 323-336.

Willingham, Warren W. and Nancy Cole (eds.) (1997). *Gender and Fair Assessment*. Mahwah, NJ: Lawrence Erlbaum, Associates.

The Collaboration for Equity is a joint project of American Association for the Advancement of Science, Education Development Center, Inc., Campbell-Kibler Associates, Inc., Girls Incorporated, Nancy Kreinberg, founding director of EQUALS and Dr. Beatriz Chu Clewell. For further information contact Yolanda George (ygeorge@aaas.org) or Patricia B. Campbell (campbell@campbell-kibler.com).

Illustrations by Judy Butler.