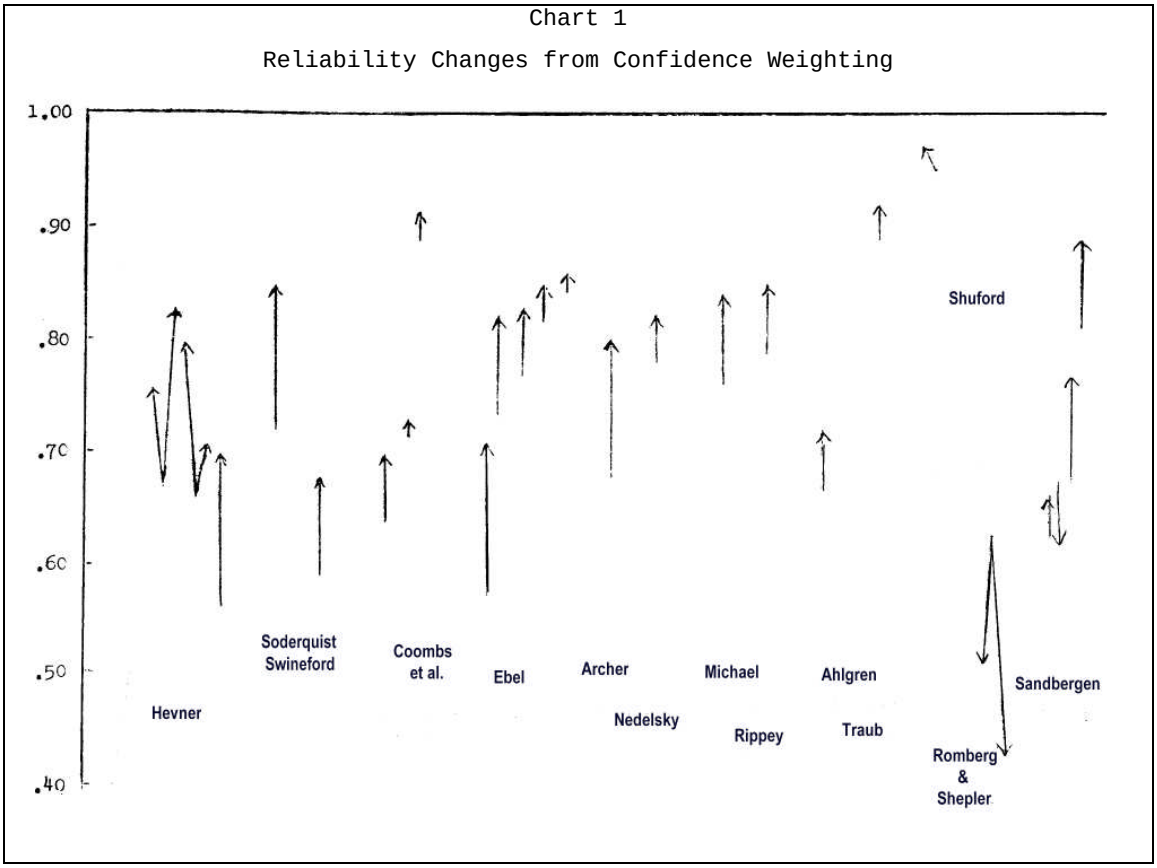
RELIABILITY, PREDICTIVE VALIDITY, AND PERSONALITY BIAS OF CONFIDENCE-WEIGHTED SCORES*

Andrew Ahlgren, Harvard University

*Remarks delivered in the symposium "Confidence on Achievement Tests -- Theory, Applications" at the 1969 meeting of the AERA and NCME. [Ms assembled by A.R. Gardner-Medwin from text & figures at http://www.p-mmm.com/founders/AhlgrenBody.htm, Aug 2005]

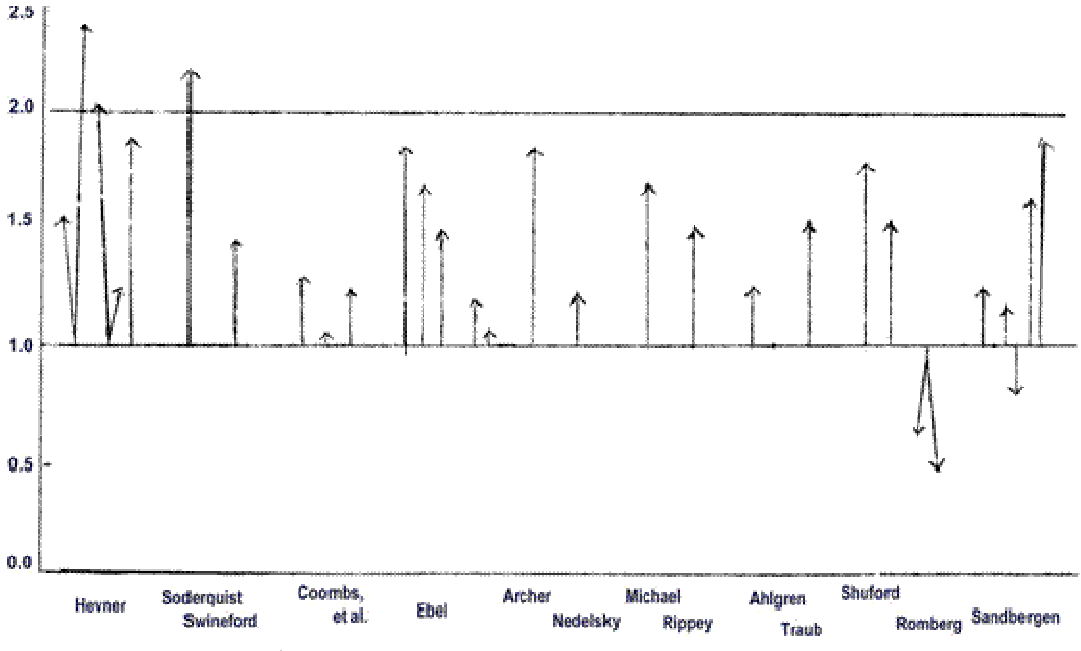
Weighting test scores by appropriateness of confidence has, almost without exception, raised the reliability of test scores. Chart 1 shows typical results, beginning with Kate Hevner's in 1932. The lower end of the arrows are the reliability estimates for the conventional scores, and the arrows indicate the increases produced by weighting the scores. (Some of the arrows are approximate averages over several subtests or subgroups.) A pair of arrows is used where there were alternative scoring systems. An exception to the general increase are the results of Romberg and Shepler, who would seem to have had a disaster using Shuford's "valid confidence score," Shuford has pointed out that there is a shrinkage of variance in this score and he himself uses a more simply-derived score to correlate with other variables.



Greater gains appear to occur for the less reliable tests, but that is at least partly because the more reliable a test is to begin with, the more difficult it is to improve it. The increases can be I displayed in a somewhat standardized fashion by expressing them in terms of how much longer the conventionally-scored test would have to have been to show the same increase in reliability -- by virtue of its greater length alone. Such a calculation can be made directly with the Spearman-Brown formula1 and has been given various names -- let me call it the "effective test length." Chart 2 is the transformed display of reliability increases expressed as effective test lengths. This has been a fairly common way of expressing the merit of a confidence weighting system, and I want to make three points about interpreting it:

Chart 2

Changes in Effective Test Length
for Reliability Changes



First, effective test lengths are not very reliable - small uncertainties in the test reliability estimates can lead to large uncertainties in this ratio.

Second, they are derived on the assumption that the items which would be added to extend the test would have the same characteristics as the original item set. Because it is hard to construct good items, these ratios probably under estimate the advantage of the confidence-weighting. Third, these hypothetical effective test lengths ought to be compared with the actual increases in testing time required for confidence-marking. For example, Joan Michael found an effective test length of 1,7, Her confidence marking procedure was a separate session, for which an additional 18 minutes was allowed on a 35 minute test, so testing time was increased by a factor of only 1,5.

I am tempted to introduce a new variable called, say, "test efficiency," that would be the ratio of hypothetical length to actual testing time. For Joan Michael, that would be 1,5 or about 1,1, which could be taken as an indication of the practical usefulness of her confidence-marking system, Unfortunately, testing time increases have been reported only rarely. Moreover, in many situations the testing time may not be an important consideration. So you are spared another chart on reliability increases.

There is, however, an overall limitation to figures on reliability increases. "Reliability" in this context means "internal consistency" -- estimated in the early studies by boosted split-half correlations and in the more recent studies usually by the Kuder-Richardson formula. In increasing either of these reliability estimates, we are reducing the random error in measuring whatever it is that the test measures. But what if the confidence weighting adds to the test a component that measures something else? To those who have turned to confidence-testing solely to suppress guessing error, an increase in reliability is sufficient and new components are suspicious at best.

I propose to you, however, that there are very few instances in education where we really care about how much knowledge a student has crammed into his head for an achievement test. Most instruction is intended to have long-term effects -- over months at least and perhaps over an entire lifetime. Yet most achievement-testing is done immediately after an instructional sequence. I suspect (in fact, I know very

well) that a substantial part of knowledge measured on an achievement test is ephemeral knowledge, stored fleetingly for the purpose of taking the test. If confidence-testing allows us to weight heavily on well-settled knowledge and weight lightly on transient knowledge, then the weighted score might predict much better the state of knowledge at a later time -- which is usually what we are really after. For this purpose, reliability increases are not sufficient and may not even be necessary. We can well afford to measure something less well if it is something that is more worth knowing. In a word, I am trying to sell validity.

Chart 3

Validity Changes from Confidence Weighting

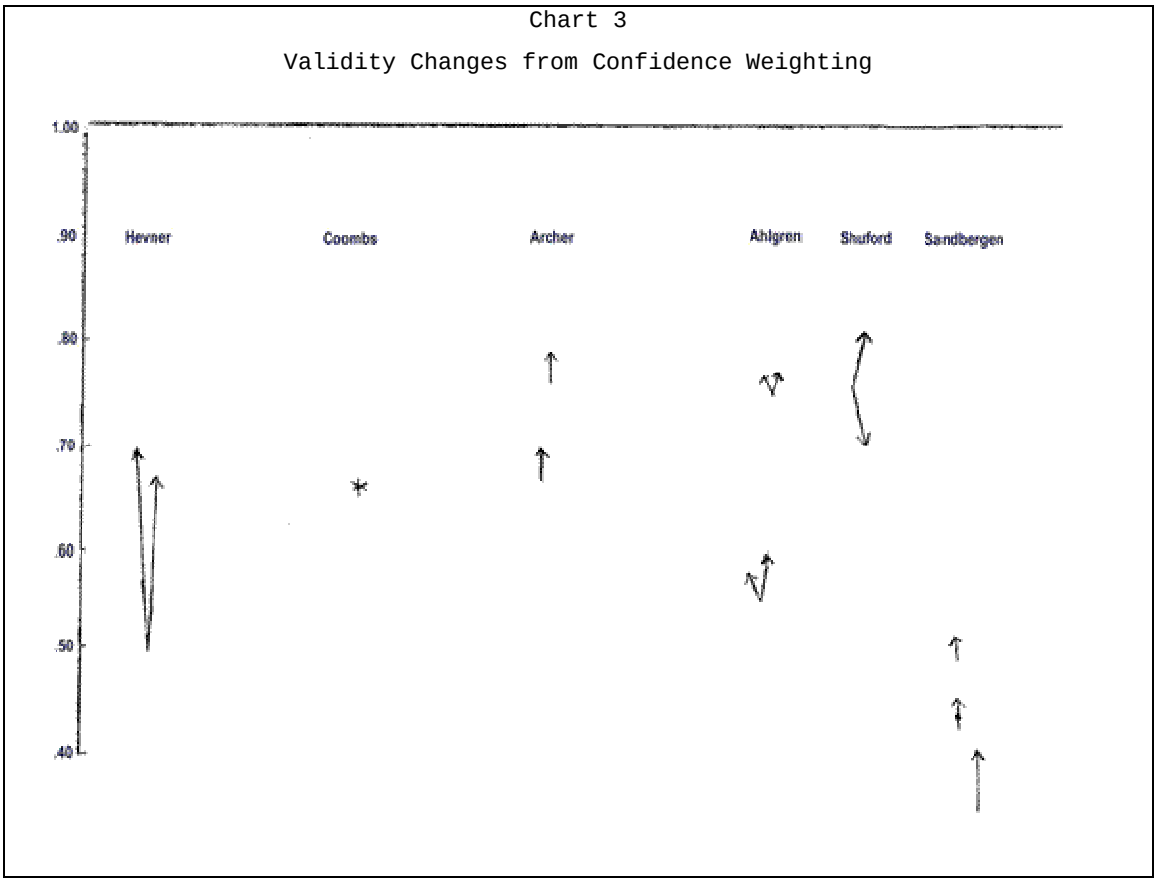


Chart 3 displays changes in validity resulting from confidence- weighting scores. The validities are Pearson correlations with some criterion, and the same convention is used as in Chart 1. Hevner (who as far as I know really began confidence-testing) gave two self-rating instruments to validate her system. Validities have not often been reported since then. Coombs, Milholland, and Womer reported, in 1956, no significant differences in concurrent validity; but they did not publish the actual figures. Archer reported, in 1962 and 1963, mixed changes in concurrent validity.

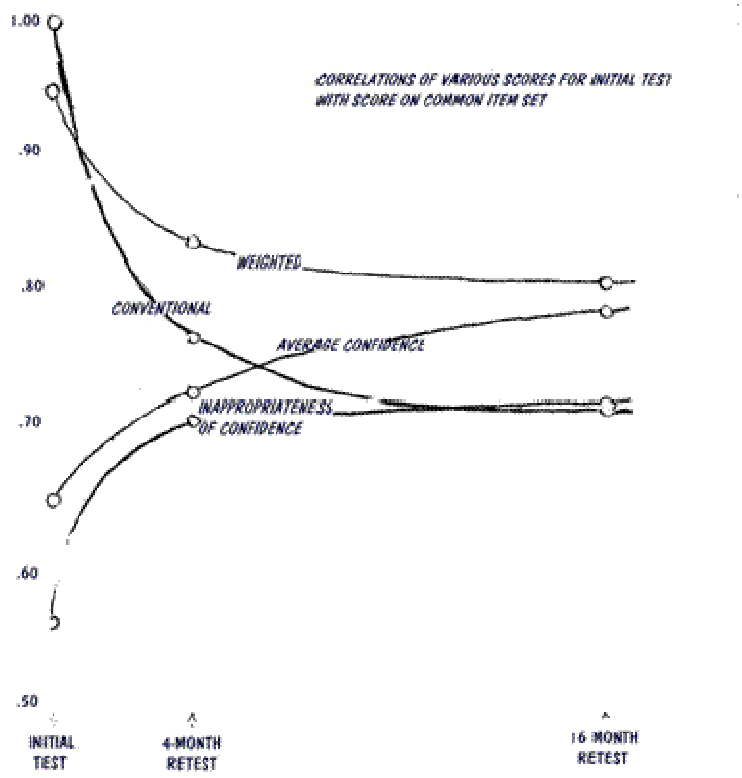
The trouble with concurrent validity is that it implies you already have some better means of measuring what you want to know. At the 1968 AAA meeting I had a conversation with Prof. Ebel about this and he expressed the opinion that usually there is no better validity criterion for a test than the test itself. There is a class of situation, however, in which the test is admittedly not the best criterion. Where external validity becomes important is where you know of a better measure but it is too difficult or too costly to use: you would like to give individual IQ tests, but it is much easier to give group tests; you would like to measure a student's grasp of problems of democracy five years after he has left high school, but it is practical to do that only for very small samples.

In my own research, I was concerned with retention I used two kinds of predictive validity for a three-level confidence-marking scheme, In one study of 160 high school physics students, I used a four-chapter

test to predict the semester grade assigned by the teacher several months later, The weighted scores had higher reliabilities and correlated significantly higher with the semester grades. In a second study, of 320 high school physics students, I again gave a four-chapter test and followed it with a parallel-form retest some four months later. This time the reliability of the weighted score did not go up, but the weighted initial score did a significantly better job anyway of predicting the ordinary retest score, (A summary of this study appears in the attached abstract.)

In a follow-up study, 32 of the students were retested again, sixteen months after the initial test. Again, the weighted initial scores correlated better with the retest score. (It is interesting, by the way, that the best predictor of the score on the unfamiliar retest items was just the average confidence level on the initial test.) A set of graphs prepared for this follow-up study appear in Chart 4. As imprecise as so small a study is, it is tempting to infer from these graphs that conventional scores become less relevant to retention with the passage of time and confidence scores become more relevant.

Chart 4



ADDENDUM:

The one class of students who were in the 11th grade when the test was first given was given the test again at the end of the 12th grade (16 months later). Weighted scores on the initial test were better than conventional initial scores in predicting late retest scores: r = .81 vs. r = .72 for familiar items, and r = .41 vs. r = .37 for unfamiliar items. But for N = 32 the differences were not significant at the .05 level. However, average confidence scores on the initial test did contribute significantly to prediction of late retest scores on both familiar and unfamiliar items. (In fact, initial average confidence scores were the best single predictor of late retest scores on the unfamiliar items r = .50).The curves drawn between data points for unfamiliar items on the graph below, although admittedly unlicensed interpolations, suggest how various scores may relate to long-term retention: the relevance of the weighted score decays less

rapidly than that of the conventional score, and the relevance of the average confidence and appropriateness of confidence scores increases with time.

I have just recently received from the Netherlands a paper by Sandbergen in which he reports several studies of predictive validity. Briefly, a test in psychology was marked with a two-level confidence system, and then used to predict average examination scores in the psychology curriculum at the university. The confidence-weighted scores had higher reliability and higher predictive validity.

It might seem reasonable to extend the idea of effective test length to validity, calling it, say, "effective predictive test length." Such a figure would indicate how much longer a test would have to be in order to give the same improvement in prediction, by virtue of increased reliability alone. A derivation of such a value follows directly from the Spearman-Brown formula and the correction-for-attenuation.2 I found, for example, for unfamiliar items on the retest after four months, an effective predictive test length of about 1.8. (The confidence-marking added about five minutes to a 40-minute test, so testing time was increased by a factor of only about 1.1.) I had planned to prepare a chart of effective predictive test lengths, for the few validity studies we have. This proved not to be feasible, however, because the increases in prediction accuracy were larger than could possibly be accounted for by reliability increase alone. This indicates rather clearly, I think, that some new component appears in the weighted scores. Such a new component is both encouraging and worrisome.

Unquestionably, personality has an effect on confidence-marking. Most researchers have adopted weighting systems that give students the greatest promise of reward for honest marking. But betting against the house is a persistent human trait, and there is ample evidence that characteristic levels of confidence and risk-taking play a large part in confidence-marking. Indeed, confidence-marking has been used (by Swineford, by Gritten & Johnson, and by Ziller) to derive personality measures. It does not necessarily follow, however, that the existence of a personality effect invalidates confidence-testing. Whatever the complex relations between confidence and risk-taking, the question is whether personality differences result in weighted scores having an unfair bias. Now, what is fair and what is not depends on why you are testing. If it is for feedback to students for their own information, then perhaps fairness is not an important issue. Indeed, a personality bias in results could have worthwhile educational effects. If the purpose of testing is, on the other hand, to judge students, or to judge an instructional procedure, then possible unfairness becomes a central problem.

There is very little evidence on this problem. Joan Michael did separate reliability calculations for subgroups cross-classified by sex and IQ: she found no important differences. My own work is the only other I know of along this line, so let me narrow the question still further: "do personality differences result in biased predictions of scores on delayed retests?" I did separate analysis for subgroups classified by sex, initial test score, average confidence, appropriateness of confidence, general test anxiety,3 and general defensiveness,4 and then for subgroups cross-classified on pairs of these variables. A computer program generated plots of systematic and random prediction errors for the subgroups.5 The criterion, you will recall, was a parallel form test given four months afterwards. The conclusion: for many subgroups there was prediction bias when using the weighted scores; for these same subgroups there was also bias when using the conventional scores; the bias for the weighted scores was never more, and was often less, than the bias for the conventional score. My results certainly are not definitive, but they do suggest that, at least with retention as a criterion, it is conventional test scores that may be more susceptible to bias against some personality subgroups.

I would like to conclude with three suggestions for research. A first priority is to establish how simple a marking and scoring system can be and still give worthwhile results. Surely one of the major barriers to widespread trial of confidence-testing is complexity of marking and scoring. Shuford's system has, I believe, 26 levels of confidence. Yet Archer, Ebel, and Sandbergen have gotten satisfactory results with a two-level system. (I have a simple computer program for scoring the three-level system with ordinary IBM or DIGITEK answer Sheets. I also have some sheets showing a simple hand-scoring technique.)

What systems are "worthwhile" relates directly to the matter of personality bias, and concern about personality bias is another major obstacle to the spread of confidence-testing. The immediate problem is not to explore endless psychological ramifications of who responds how, but to find out whether confidence-weighted scores are "unfair" to some personality subgroups. This means parallel analyses for subgroups in which bias might be suspected.

"Unfairness" has meaning only in the context of some purpose of testing, and that leads directly to my third suggestion, which is to abandon the fascination with reliability as an end in itself. Suggesting that to an audience of psychometric folk may be a little like knocking aspirin at a medical convention. The analogy may be apt: aspirin is terribly useful, especially when we don't know what else to do -- but it does little to advance research.

I would like to see the main concern focused instead on validity, and validity of a particularly basic kind. First you have to ask a simple (but perhaps embarrassing) question: Why are we testing? In answering this, criteria for validity should appear. With specific criteria, it becomes possible to consider "fairness" operationally, and then to consider how elaborate a marking-system is needed to be "worthwhile."

In brief, I propose that confidence-testing be pursued less in a purely psychometric or pyschological context, and more in a context of education,

NOTES

1. Solving the Spearman-Brown equation for n, the length-increase factor, yields:

$$n = \frac{r'_{xx} - r_{xx}r'_{xx}}{r_{xx} - r_{xx}\, r'_{xx}}$$

2. Combining the Spearman-Brown equation with the correction for attenuation (Guilford, Psychometric Methods, p. 408), and solving for n, the 1ength-increase factor, yields:

$$n = \frac{r'^2 - r'^2 r_{xx}}{r^2 - r'^2 r_{xx}}$$

3. The anxiety instrument was the Alpert-Haber Achievement Anxiety Test. (See Alpert, under Selected References.)
4. The defensiveness instrument was the Marlowe-Crowe Social Desirability Scale which is generally interpreted to measure defensiveness. (See Crowne and also Kogan, under Selected References.)
5. An examp1e of the computer-generated charts is attached as Chart 5.

Selected References

Ahlgren, A. Appropriateness of confidence on achievement tests. Qualifying paper on file at the library of the Harvard Graduate School of Education, 1967.

Alpert, R., & Haber, R.N. Anxiety in academic achievement situations. J. abnorm. & soc. Psychol., 1960, 61, 207-215.

Archer, N.S. A comparison of the conventional and two modified procedures for responding to multiple-choice items with respect to test reliability, validity, and item characteristics, Unpublished doctoral dissertation, Syracuse University, 1962.

Archer. N.S. Effects of confidence weighting by fifth and sixth grade students on objective test scores. Abstract of a paper read at annual meeting of AERA, Chicago, 1963.

Coombs, C.H., Milholland, J.E. & Womer, F.B. The assessment of partial knowledge. Educ. Psychol.Measmt. 1956, 16, 13-37.

Crowne, D.P., & Marlowe, D. A new scale for social desirability independent of psychopathology. J . consult. Psychol., 1960, 24, 349-354.

Ebel, R.L. Some effects of credit for confidence and penalty for error on true-false test scores. Abstract of a paper read at annual meeting of AERA, Chicago, 1961.

Ebel, R.L. Confidence weighting and test reliability. J Educ. Measmt., 1965, 2, 49-57.

Gritten, F., & Johnson, D.M. Individual-differences in judging multiple-choice questions. J. Educ. Psychol., 1941, 32, 423-430.

Hevner, K. Method for correcting for guessing and empirical evidence to support. J. Soc. Psych., 1932, 3, 359-362.

Johnson, D.M. Confidence and the expression of opinion. J. Soc. Psychol., 1940, 12, 213-220.

Kogan, N., & Wallach, M.A. Risk-taking: a study in cognition and personality. New York: Holt, Rinehart & Winston, 1964.

Michael, J.J. An experimental analysis of the reliability of a multiple-choice examination under various test-taking instructions. Paper presented at the annual meeting of the NCME, Chicago, 1968.

Nedelsky, L. Ability to avoid gross error as a measure of achievement. Educ. Psychol. Measmt., 1954, 14, 459-472.

Rippey, R. Probabilistic Testing. J. Educ. Measmt., 1968, 5, 211-215

Romberg, T.A, & Shepler, J.L. An Experiment Involving a Probability Measurement Procedure, Paper presented at the annual meeting of AERA, Chicago, 1968.

Shuford, E.H., Jr. & Massengill, H.E., Jr. Airman Qualifying Examination - 66 administered as a confidence-test. Report printed by Shuford-Massengill Corporation, 1968.

Slakter, M.J. The effect of guessing strategy on objective test scores. J. Educ. Measmt., 1968, 5, 217-221.

Soderquist, H.O, New method of weighting scores in a true-false test, J. Educ. Res., 1936, 30, 290-292.

Swineford, Frances, Measurement of a personality trait. J. Educ. Psychol., 1938, 29, 295-300.

Swineford; Frances. Analysis of a personality trait. J. Educ. Psychol., 1941, 32, 438-444.

Torrance, E.P.,& Ziller, R.C. Risk and life experience development of a scale for measuring risk-taking tendencies. Unpublished U.S. Government report (AFPTRC-TN-57-23). Air Force Personnel and Training Center, Lackland Air Force Base, Texas. Feb. 1957.

Traub, R.E., et al. Effects of promised reward and threatened penalty on performance on a multiple-choice vocabulary test. Paper read at annual meeting of AERA Chicago, 1968.

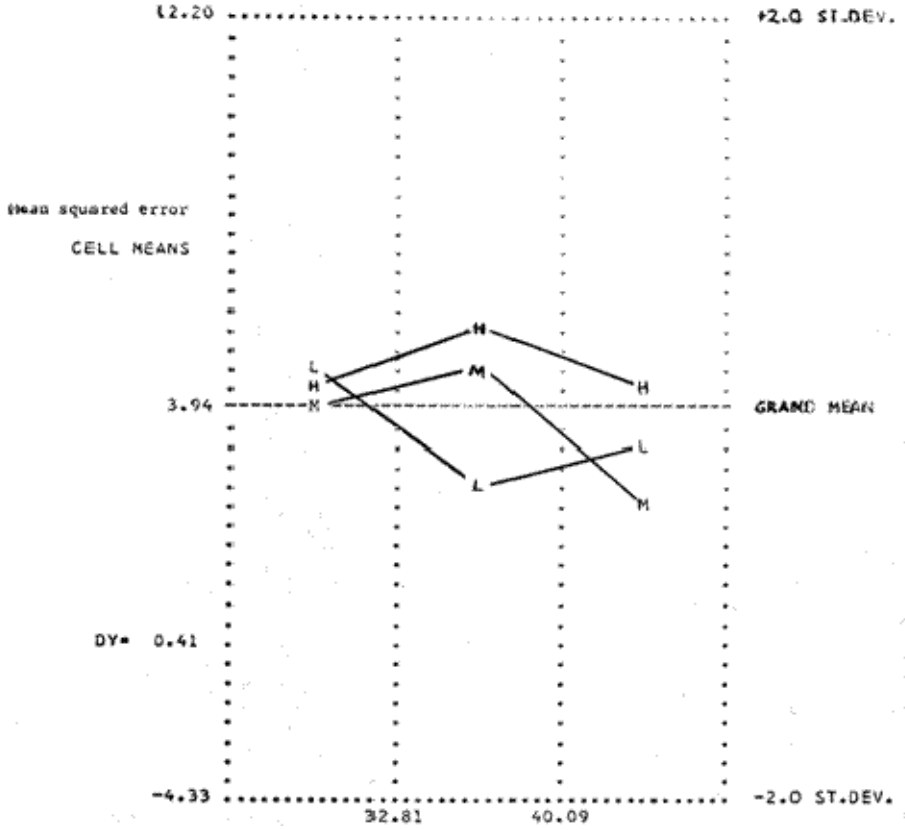
Ziller, R.C. Measure of the gambling response set in objective tests, Psychometrika, 1957, 22, 289-292.

Also, a review of the literature, details of rationale and mathematics of simple scoring schemes, multiple-regression analysis, and interaction plots can be, found in:

Ahlgren, A. Confidence on achievement tests and the prediction of retention. Unpublished doctoral dissertation, Harvard University, 1967.

A limited number of copies of this document is available from the author at the Harvard Project Physics, 8 Prescott St, Cambridge, Mass.

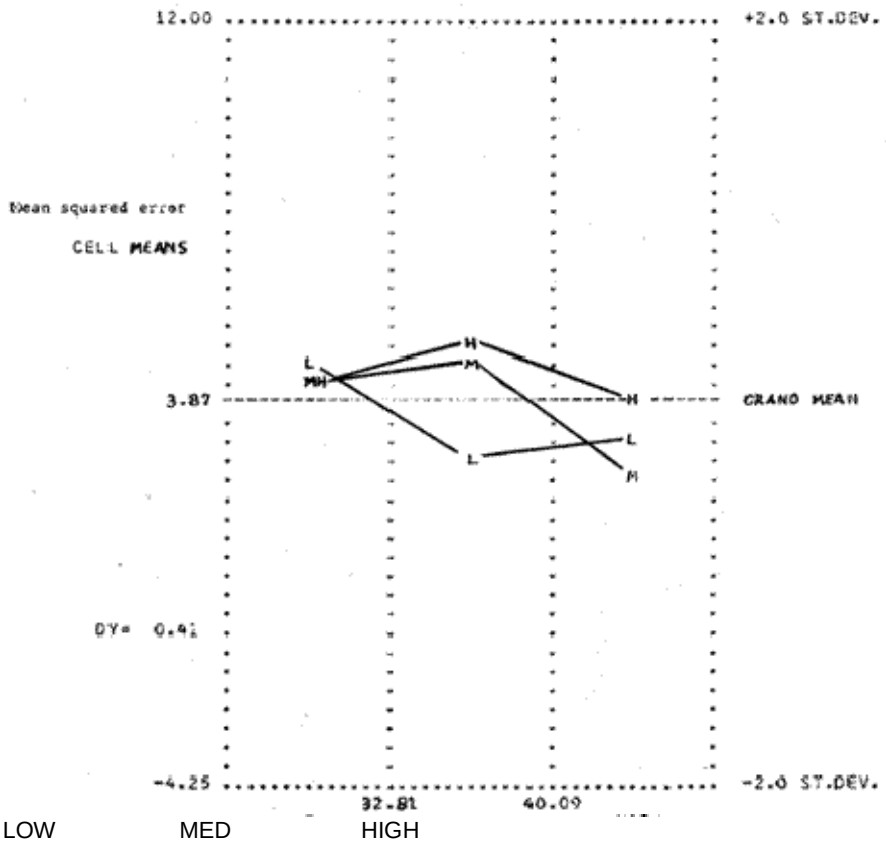
MODERATOR IS DEFENSIVENESS H GROUP IS GREATER THAN 13.70 GROUP B
L GROUP IS LESS THAN 9.90



	LOW	MED	HIGH
			TEST ANXIETY
NUMBER IN CELLS	L= 17 M= 11 H= 21	L= 21 M= 20 H= 21	L= 18 M= 14 H= 15
CELL MEANS	4.91 4.12 4.27	2.31 4.93 5.49	2.94 2.07 4.23
S.E.'S OF MEANS	1.09 1.02 1.00	0.79 1.07 1.04	0.87 0.44 1.11

Chart 5a. Mean squared error in predicting scores on unfamiliar retest items from initial conventional scores.

MODERATOR IS DEFENSIVENESS H GROUP IS GREATER THAN 13.70 GROUP B
L GROUP IS LESS THAN 9.90



	LOW	MED	HIGH
NUMBER IN CELLS	L= 17 M= 11 H= 21	L= 21 M= 20 H= 21	L= 18 M= 14 H= 15
CELL MEANS	4.82 4.12 4.16	2.49 4.75 5.27	3.13 2.14 3.95
S.E.'S OF MEANS	1.11 0.98 1.02	0.80 0.97 1.09	0.89 0.47 0.97

Chart 5b. Mean squared error in predicting scores on unfamiliar retest items from initial weighted scores.