TEST THEORY

James Lumsden
Department of Psychology, University of Western Australia, Nedlands, 6009, Australia

Test theory has had few major ideas. I can list only five:
1. the decomposition of obtained scores into true and error components (20, 32, 66, 81, 90);
2. the duality of psychophysics and psychometrics (16, 72, 73, 88, 91);
3. the notion of unidimensionality (33, 34, 51, 68);
4. the conception of test validity as theoretical equivalence, usually called construct validity (21);
5. the scaling ideas derived from various item characteristic curve models (3, 4, 54, 75, 91, 93).

Apart from the first, these ideas have been hesitatingly and unconfidently applied.
As both an effect and a cause, there has been a general atmosphere of melancholia and lassitude among latter-day test theorists. Book & Wood (7) point out that the uncritical use of tests gives test theorists little incentive to improve them, and this may account for the discouragement. There is, I believe, a more important reason. The shreds of theory that have been developed and the time-worn true score models are not rich sources of ideas about testing, and the ideas that are generated seem to relate mainly to the internal workings of the models.

A VULGAR ANALOGY—THE FLOGGING WALL TEST

What test theory needs is a tool for thought experiments. I have constructed a test for test theorists which illustrates, and at times illuminates, many of the problems of test theory. Along a long wall at widespread intervals flexible canes are attached at various heights. These canes flog slowly and independently up and down at varying amplitudes. Subjects stand on a cart which is drawn quickly along near the wall. A subject's score is the number of canes which touch him. The scores represent a mapping of the attribute of height into the number system. This follows from the fundamental theorem of axiomatic approaches to measurement:

\[ a \in M \text{ iff } a \in O M \]

1This essay covers, very roughly, the period from 1904–1974. More detailed attention is given to the events, mainly unfortunate, of the period 1970–1974.
which may be translated: \( a \) is a legitimate mapping if, and only if, \( a \) is an obviously legitimate mapping.

How could we make the flogging wall test more reliable? One way would be to make it longer, to use more items. Another would be to reduce the amplitude of the flogging, that is, to make the items more reliable. Suppose that we stopped the flog so that the canes were static. The test would be perfectly reliable and the obtained scores would be identical with the platonic true scores. The platonic true score is respectable. It represents the scores obtained when the most popular item selection procedure is carried to the limit.

When we stop the canes the test is obviously perfectly reliable. What method of estimating reliability would show it to be so? The problem is more difficult than it seems. I will consider only the test-retest method and leave the reader to work out the difficulties with other methods. The test-retest method will yield a reliability coefficient of one if there is no practice effect. But suppose that we send a sample of test theorists through the test and that because their heads are unbowed they become bloody. Let us suppose further that soon after, not during, the test all their heads swell by exactly 3 inches. (The relationship between within-testing practice effects and between-testing effects appears not to have been investigated.) Notice that this is an attribute effect and is extremely unlikely. Long Louis hits many more canes than Big Louis, who is 6 inches shorter, and Short Sam hits only one or two. Will, however unlikely, uniform practice effect on the attribute yield a perfect reliability estimate? Not at all, because the effect in score terms will be different. Suppose that there are three canes within less than an inch around 5 feet and the next is at 5 feet 6 inches. A subject whose height at first testing was just under 5 feet would have his score increased by three, while a subject whose height at first
testing was just above 5 feet would remain at the same score on second testing. What is required to produce a perfect reliability coefficient is that each score will be increased by exactly the same amount. It requires a very watchful and kindly providence to stretch all subjects by precisely the different attribute amounts necessary for this to happen. Yet this is the assumption made by test-retest and parallel form estimates of reliability. It is hard to believe that any test theorist could believe it.

Does a perfectly reliable test perfectly reflect the attribute? No, because in the case we have just discussed subjects who differ in height by almost 6 inches (in the region of 5'0" to 5'6") will have the same score, while subjects in other regions who differ in height by only an inch may differ in score by three or more. Subjects with the same score may have widely different attribute values and subjects with different scores may have very similar (never identical) attribute values. The reader may wonder whether the perfectly reliable test is all that he had been led to expect. If we allow the canes to flog, the probability now is that subjects near 5'6" in height will have higher scores than subjects near 5'0", and so on for other positions. If the gaps are large, it may be necessary to have a large amplitude of flog and to make up for the heavy loss of reliability by increasing the number of items. The consequence of this will be that the obtained scores, now somewhat unreliable, will represent the attribute more accurately and that, therefore, the correlation between the scores and any criterion for which height is relevant will increase. In other words, validity will increase when reliability decreases.

Anyone for attenuation? The attenuation paradox has been noted many times (10, 52, 53, 92). It is significant that the attenuation paradox is mentioned in Lord & Novick (66, p. 344) almost 300 pages after an uncritical advocacy of the correction and that nowhere do they state that the correction almost always produces an overestimate. The attenuation paradox has been attributed by Cronbach and associates (20) to a narrowing of the universe, i.e. to a decrease in the variety of item types included in the test. The example quoted above shows that this is not necessarily the case.

Someone may object that the rather eccentric distribution of item difficulties given in the paradox example is untypical of psychological tests. That is true, but many test constructors attempt to peak difficulties around the 0.50 level and have few easy or difficult items. The result with highly reliable items is a breakdown in discrimination at the high and low levels.

The attenuation paradox cannot occur if the item difficulties have a distribution which is adequately and uniformly dense across the range of attribute values to be assessed. Lord & Novick (66, p. 465) seem to agree. Adequate density means that subjects falling in a given interval between items, in attribute terms, will be treated as having the same attribute value for all relevant purposes. What constitutes adequate density will depend on circumstances.

Distribution of item difficulties has another advantage. Recall the second of our major ideas in test theory: the duality of psychophysics and psychometrics. If there is an item characteristic curve, then there must exist a person characteristic curve. If items need two (maybe three) parameters, then persons do so too. The person
characteristic curve is the plot for a particular individual of proportion passed for items of different difficulty. Two persons with the same location value may have different slopes. Bill, a careful worker, may get all easy items right but no hard ones. Joe gets some easy ones wrong but some hard ones right. Whether we should take this into account is a matter for empirical determination. Certainly there seems no theoretical ground for not taking a look. Thorndike (85) and Furneaux (29) have adopted this procedure. The advocacy of peaked tests seems to have arisen from a preoccupation of test theorists with obtaining the most efficient estimation of the location parameter. The third parameter? Well, Guilford (31, p. 139), in a discussion of the constant methods, suggests that proportion of "doubtful" judgments (unwillingness to guess) may be an important personality characteristic. Swineford (84) and Ziller (102) have made similar suggestions.

The flogging wall is a powerful tool for thought experiments. Some may object that computer simulations of various mathematical models of items do much the same thing more elegantly. The alternatives are not vulgar enough. They do not have a highly visible, distinctive attribute. Even more importantly, the distinction between attribute values and scores is not made palpably obvious. The flogging wall test together with some elementary statistics is almost a test theory in its own right, and we shall see demonstrations of its power throughout this essay.

A MUSEUM FOR MODEL T

The most highly regarded notion in all test theory, and the only one to be seriously developed, has been the venerable: \( O = T + E \). I term this Model T for polemical purposes. The idea has been stated and restated and variously applied and interpreted. Legend has it that it began with Udny Yule, who told Spearman (81). The idea was perpetuated and elaborated by a royal line which included Thurstone (90) and Gulliksen (32) and achieved a memorial in a grand mausoleum constructed by Lord & Novick (66). In the period under review Cronbach et al. (20) have made a gallant attempt to boost the sales of the old Tin Lizzie but the color remains black. The decomposition of obtained scores into true scores and error scores is unfruitful. When true score is defined as an ideal (platonic) score stripped of error, the result is a contradiction. When true score is defined in other ways so as to avoid the contradiction, then the resultant statistics have no useful application.

Consider an eight-item completion test in which items are scored dichotomously. The obtained scores can only take integral values from 0–8. Suppose that we define a true score as an obtained score stripped of error, that is, when both the items and the persons are perfectly reliable. The platonic true scores also can only take on integral values from 0–8. Consider a subject whose true score is 8. His obtained score can never exceed this value so that his error score can only be zero or negative. Similarly, a person whose true score is zero can only have error scores which are zero or positive. For other true scores both positive and negative error scores are possible. Error score is not independent of true score, and in most situations it would appear that \( r_{TE} < 0 \), which is a contradiction of the Model T assumption that \( r_{TE} = 0 \). In the face of this it is necessary to abandon either Model T or the platonic...
true score definition. Most test theorists have taken the latter course, agreeing with the Bock & Wood (7, p. 198) statement: "There seems little to be gained from further prolonging the life of the platonic true score as a concept in test theory." There is some curious fear of the platonic true score as representing an "eye-of-God-reality" (20, p. 19). Stanley (82) and Thorndike (86) seem to be similarly troubled when they say that true score is not recorded in the book of heaven. Yet the very same writers will freely admit that estimates of true score may be in error, i.e. differ from some really true score. If not recorded in the book of heaven, one wonders in what demonic book the Lord & Novick expectations or the Cronbach et al universe scores are recorded.

There is, however, no joy for Model T pushers in other definitions of true score. Lord & Novick (66) define true score as the expected value of the obtained score over replications. Definitions in terms of limits (Gulliksen 32) amount to the same thing. The contradiction is now avoided since the only way that maximum or minimum true scores can be obtained is by the occurrence of invariant maximum or minimum obtained scores over replications. The expected value of the error score can now be zero for all values of true score, and the assertion $r_{TE} = 0$ is not contradicted. But there remains a problem. Error score is not completely independent of true score. While the expected value of the error score is constant (zero) for all values of $T$, the variance of error scores is not constant but clearly zero at the maximum and minimum scores and rising to a maximum at some middling value of $T$. It should be noted also that the distribution of error scores is likely to be highly skewed for extreme values of true score because of the bounds on obtained score. The breakdown in homoscedasticity for the bivariate distribution of true and error scores has been demonstrated by Lord (53).

It is clear that in the Lord & Novick model the regression of obtained score on true score is linear. A little reflection on the results for a short test will show that the regression of true scores on obtained scores is not linear. Lord & Novick (66) state that the conditions for linearity for the true score on obtained score regression are that the errors should be normally distributed around true scores and that the distribution of the true scores should be normal. The first of these conditions is impossible and the second not very likely. In a a later section Lord & Novick concede that generally the regression is not linear.

The major purpose of the Model T decomposition is to provide a rationale for the reliability coefficient. Why do we need the reliability coefficient? Three major reasons have been suggested: 1. to guide test selection, 2. to support inferences about test scores based on regression estimates of true scores and the standard error of measurement, and 3. to support inferences about the validity of perfectly reliable tests (correction for attenuation). The first of these need not concern us. For most purposes there are better guides (Cronbach 17), and we have already seen that the reliability guide may be misleading. The second reason does not apply. Linear regression estimates of true score may be quite misleading if regression is not linear. Setting up confidence limits using the standard error of measurement will also be misleading since the standard deviation of error scores is not independent of true scores. It is not good enough to scrub around this latter problem as is often done...
in other contexts by asserting that the standard error of measurement represents in
an average sense the standard deviation of the error arrays and that departures from
homogeneity may be tolerated. The departures are systematic and may be too large
to be tolerated. A different line of attack on this problem will be presented in the
consideration of recent work on reliability where it will be shown that even if the
problems of nonlinearity and heteroscedasticity are solved, the true score concept
remains useless.

There remains the third reason: the use of reliability estimates in the correction
for attenuation. This is regarded by Lord & Novick (66, p. 71) as one of the best
justifications of reliability theory. Cronbach et al (20) also strongly advocate the
correction. The correction should never be used. It too often produces corrected
correlations which are greater than one, and it is not sufficient to pass these occa-
sions off with an embarrassed smile and some mutterings about unreliability of
estimates. The problem is simple and unsolvable, at least by classical methods. All
estimates of the reliability coefficient (with the possible exception under certain
conditions of the test-retest estimate) are underestimates, sometimes gross underesti-
mates, of reliability as classically defined. Since the correction for attenuation re-
quires division by the square roots of the reliability coefficients, it follows that the
correction produces estimates which are overestimates, sometimes gross overesti-
mates. Note, too, that the attenuation paradox makes nonsense of the correction.

What has gone wrong? The Yule suggestion that \( O = T + E \) was probably made,
unless it was a cruel joke, in the belief that we were dealing with something like
length measures where the t's and e's are continuous and essentially unbounded. In
this case the assumptions of Model T can hold without contradiction and the
inferences will generally be sound, but now they will refer to attribute values.

A technical criticism of theories based on the true score model and a sketch of
an attribute based model, Model A, may be found in Ross & Lumsden (80). In this
paper it is suggested that test theory followed a will o' the wisp when it set out on
the false trail of true score. Whenever a procedure is set up which estimates the
attribute value, it will be unnecessary and misleading to think of an ideal value of
the estimator.

1970–1974

Few valuable contributions to test theory have appeared during the review period.
There seems to be a degree of desperation among test theorists as if they are groping
for suitable topics to write about. Many papers are of very slight merit and, though
this may simply be a case of reviewer's jaundice, there seem to be more technical
errors than in previous review periods. The standards set by editors and article
reviewers are low, and many papers could be greatly improved by some criticism.
It is appropriate here to suggest that Educational and Psychological Measurement
should review its policy of not publishing comments. These, if prompt and hot, can
have a salutary effect. A battered author and a reviewer who has had to say an abject
mea culpa will be likely to be more careful next time. And so will others.
This review follows Bock & Wood (7) in restricting consideration largely to test theory as concerned with itemized tests of ability and achievement. It is the reviewer's belief that if progress in test theory is to be made at all, and there is not much evidence of it yet, it will be made first in this area. The review is organized under traditional topic titles with the addition of a section on dimensionality. This is partly because your reviewer is conservative, but mainly because it is important to reveal what has not been done as well as what has been done. There are some alarming holes.

Reliability

As usual, there have been many papers in this area. Generally the authors uncritically adopt one form or another of the true score model and reveal little understanding of the problems which the model was designed to solve. There is, with one exception, no imaginative approach to any of the problems but simply a grinding of mathematical symbols (with fantastic assumptions when things get tough) to produce nothing of any worth. It is extraordinary that gifted mathematicians should constantly fail to realize that the obtained scores can do exactly the same task as the estimated true scores. Rarely do they see that the reliability problem is best conceived as part of a wider question: how well does the test represent the attribute? When this question is posed and understood it becomes clear that the true score and the reliability coefficient can play no part in providing an answer and should be abandoned.

What will almost certainly come to be considered the outstanding contribution of the period is the book by Cronbach, Gleser, Nanda & Rajaratnam (20), which is the culmination of a decade of work on reliability seen as generalizability. Cronbach et al treat the obtained score as a sample from a universe of admissible observations. It is the responsibility of a test publisher to state what he considers to be this universe. The statement will consist of specifications of a number of facets or dimensions and of permitted levels of them. Thus he may state that all forms (or only one form) of a particular test are to be considered as equivalent, and in the universe of admissible observations, various examiners trained or certified in certain ways may be used, other conditions such as time and place may vary within certain defined limits. A particular person's score is obtained from one form of the test, with a particular examiner, in a certain place and on a certain occasion. This obtained score will normally differ from his universe score which is defined as the expected score over all admissible observations. The universe score is perfectly analogous to the true score of classical reliability theory.

After defining the universe of admissible observations, Cronbach et al say that the publisher should carry out a G (generalizability) study, which will permit variance components to be estimated for persons, forms, examiners, occasions, and for the interaction terms. Ideally the G study should be completely crossed, but it is recognized that some nesting (e.g. with occasion) is inevitable and that economics, subject fatigue, etc will often require further nesting. From the variance components it is possible to estimate the coefficient of generalizability defined as the ratio of
universe score variance to expected observed score variance. The estimate is made
by calculating an intra class correlation from the ratio of person variance to the sum
of the person variance and the variance components for person interactions, and the
residual variance. The coefficient of generalizability is perfectly analogous to the
classical reliability coefficient. It is recognized that different test users may have
different acceptable modes of forming the universe of admissible observations and
that therefore a particular test may have many coefficients of generalizability. These
may often be calculated from the publisher's G study by fixing one facet at a
particular level and making the appropriate changes in the variance components for
the intra class correlation.

The coefficient of generalizability obtained in the G study may be used in a D
(decision) study to estimate the universe score from a subject's obtained score. The
universe score is said by Cronbach et al (20, p. 15) to be the "ideal datum on which
to base. . . . decision" (s).

The book is in many respects a considerable achievement. It bites firmly on the
sour apple of Guttman's (35) critique of Gulliksen (32), where he points out that
a test will have as many reliabilities as there are acceptable bases for forming parallel
tests. It makes concrete sense of the sampling model (in many places at least) and
points out the importance of often disregarded facets such as examiners. The book
provides the foundation and much of the superstructure of the unified treatment of
reliability via the intra class correlation that was requested by Bock & Wood (7).
A detailed examination with numerous examples is given of the various experimen-
tal designs that may be used for G and D studies. Cronbach et al are frank about
the problems, e.g. with estimation of variance components, and I have no doubt that
the next decade will see many papers attempting to solve these problems (101).
Wiggin's textbook (100) already has an accurate account of the generalizability
approach, and one can confidently predict that many others soon will be attempting
to do the same.

In a D study with test theorists as raters, I am sure the Cronbach et al contribu-
tion would receive a very high universe score. But here we are concerned with its
value on the attributes of logical correctness and fruitfulness. On this basis the book
is—unfortunate. The great expertise so often displayed and the highly professional
care in exposition only make things worse. The book shows all the poverty of
imagination and all the elementary logical and even mathematical errors so charac-
teristic of reliability theory. I have several reasons for these strong remarks. The
general ones are set out in the introductory sections of this essay. The occasion
demands, however, a more specific rebuttal.

In all their designs Cronbach et al advocate repeated measures without emphasizing
the obvious dangers that main effects and interactions with persons are inextrica-
ibly confounded with sequence effects and their interactions. This may matter more
in some cases than others, but it is clear that for parallel forms or test-retest
procedures the occurrence of differential practice effects cannot be ruled out and
cannot be considered as properly included, for example, within the between-form
effect or the person-by-form interaction. Much that is of value in the generalizability
approach can be obtained without repeated measures; for example, by simple fully randomized designs main effects for form or for examiners can be readily estimated. I hesitate to suggest it in case someone does it, but some very powerful and efficient designs can be worked out using parallel persons. I am fascinated by the thought that ETS should employ 100 sets of identical twins, all reared together, who do tests all morning and are used for other experiments in the afternoon.

Cronbach et al properly point out that the common practice of setting up confidence limits around the obtained score is not correct. They recommend that universe scores should be estimated via linear regression and confidence limits set out around them using an equivalent of the standard error of measurement. They realize that if all scores are regressed toward the population mean, then the universe score estimates will be perfectly correlated with observed scores and little will be gained. But if there exist subgroups of persons with different means and the regression estimates are toward the subgroup means, then the correlation will not be perfect and the universe score estimates will reflect important information not obtainable from the obtained score alone. This approach can be criticized on several grounds. As we have seen previously, it is known that the regression of true scores (universe scores) on obtained scores is nonlinear. It is true that if the reliability estimate is high, then the departures from linearity cannot be great. But in this case the regression estimate of the universe score will differ only trivially from the obtained score. (Recall, too, that estimates of unreliability are typically underestimates.) When the reliability is not high, then the departures from linearity may be substantial. It follows that when the reliability is high the linear regression estimate is not worth making, and when it is low the linear regression estimate may be misleading. We have already seen that the standard error of measurement is not constant for all values of true score and is, therefore, a dubious base for the formation of confidence limits. But even if by some mathematical magic these problems can be overcome, the approach remains inappropriate. We do not need the true score!

Cronbach et al (20, p. 14) state that measuring procedures are to be used as the basis for decisions about classifications, course success, etc, and that “the accuracy of measurement must in principle be examined separately for each application of the procedure.” In the light of this very wise statement, why not simply examine the scatter diagram relating criterion performance to test performance? For any given obtained score we can observe the scatter of criterion scores and from this obtain the expected criterion score and the required confidence limits. For different subgroups we can if necessary set up different scatter diagrams. Why do we need the estimates of universe scores? But suppose we were foolish enough to calculate the regressed scores as recommended. Would we not, in principle, still need to consult the scatter diagrams in order to arrive at a sensible decision? And wouldn’t these be exactly the same scatter diagrams with the numbers shifted a bit on one axis?

Let me make this matter painfully obvious with some brilliant mathematics. Suppose that you were unhappily calculating $x$ from $y$ using some complicated function such as $x = ay + b$ and that someone told you that a better procedure was to calculate $x$ from $z$ using the function $x = cz + d$. You inquire how to find the
z's and are told that you can estimate these accurately enough from \( y \) using the function \( z = ey + f \). Would you tell this drongo\(^2\) to go jump in a polykay, nick himself with a jackknife, or lose himself in a Bayesian convolution?

Cronbach et al. go further and recommend that profiles be plotted in universe scores with all the information from all the tests used to yield a multiple regression estimate (linear of course) of the universe scores. This is consistent with the recommendation for a single test score. Subjects with similar scores on several of the tests are treated in effect as a subgroup, and the score obtained in a particular test is regressed toward the mean of that group. The advocated procedure amounts to bad, indeed dangerous, practice. Consider the case of a boy who is referred with high scores on an intelligence and an arithmetic test and with a low score on a reading test. We are interested precisely because we expect superior reading performance from high performers on intelligence and arithmetic tests. Now any psychologist worth his salt knows that test scores are not completely reliable and that differences between scores may be misleading. So we take no action until we check the results in some way. We ask the teacher, who says that Joe seems bright, is good at arithmetic, but poor at reading. We proceed with inquiries into Joe's reading problems. Suppose that after regressing to the universe scores Joe's reading score was within the normal range. Would Cronbach et al. tell Joe's teacher to quit worrying about Joe's reading? Of course they would do nothing so crass because the information from the teacher would be taken into account. But often information arrives at different times and may be passed through several hands. How can we protect against possible gross misinterpretation? Finally there is the gruesome thought, which may be termed the erosion nightmare, that through inadvertance or incompetence a profile might be regressed several times. Let us stick to the old-fashioned percentile ranks and a little common sense.

Cronbach et al. (20, p. 309) give as part of the justification of their advocacy of universe score profiles this statement: "The scientist is concerned with relations among constructs which are always universe scores or functions of universe scores." This presumably is also a justification of their advocacy of the correction for attenuation. The first part of the quoted sentence is false and the second misleading. Constructs (in measurement theory) are attributes, traits, latent abilities, etc, and should be sharply distinguished from obtained scores, true scores, or universe scores. If constructs are functions of universe scores (estimated), then they will be the same kind of function of obtained scores. The quoted statement should be compared with Cronbach's (18) well-argued defense of surplus meaning of constructs against ultraoperationists.

But let's look again at the correction for attenuation. If the regression estimates of true score are any good, then if we calculate them for \( x \) and then for \( y \) the correlation between these estimates should approach the correlation between \( x \) and \( y \) corrected for attenuation. Would anyone like to bet on the result of this foolish experiment? Why hasn't it been done?

\( ^2 \)Drongo: an extinct Australian bird that slept with its eyes open and flew with its eyes shut.
As a final example of the looseness of thinking that Model T, by some rule of perversity, seems so often to engender, consider the following statement: "The distribution of \( \mu_p \) (universe scores) is unknown. The most plausible approximation available is a normal distribution with mean \( \mu \) and variance \( \sigma^2(p) \)" (20, pp. 146-47). Suppose that we select two very large samples randomly from a given population and that one sample does an easy and the other a difficult arithmetic test. Will the frequency distributions of the obtained scores have the same shape? It would be very strange if they did. The most likely result is that for the easy test the score would be negatively skewed and for the hard test positively skewed. What would be the distribution shapes of the universe scores for the two tests? From the assumptions of Cronbach et al, they would be almost exactly the same. After all, they go to considerable lengths to try to convince us that linear regression provides a reasonable estimate of universe scores. There is no "plausible assumption" that can be made a priori about the shape of the distribution of universe scores. The distribution of obtained scores and of universe scores is a complex function of the distribution of attribute values for the sample and of the item characteristics of the test.

Let us pause to consider what has been revealed by this tedious critique of Cronbach et al. Simply that Model T in either its classical form or in generalizability trim cannot perform the task for which it was designed. The reliability coefficient and statistics calculated from it have no useful application. They should not be used to select tests, to estimate true scores, to estimate confidence limits for scores either true or obtained, in the correction for attenuation, or for anything else. More fortunately it has been shown that this does not matter. The problems for which these devices were alleged to be the solution can be handled more simply, more sensibly, more fruitfully, though perhaps (to some) less elegantly by the application of simple regression methodology, i.e. by best validity practice.

Lest I have made an accidental friend, it should be made clear that I regard most of the criticisms as applying to reliability theory as a whole. The book by Cronbach et al received special attention because it is the latest, and by all odds one of the greatest, in the Model T tradition. But all reliability theorists from Spearman on have been about equally mistaken.

Stanley (82), in his chapter on reliability, gives an insightful account of the sources of variance in test scores and goes on to a highly technical consideration of recent work. The rule of perversity still operates. Stanley manages to produce this gem: Differences in test reliability among several tests or different ways of scoring tests frequently appear negligible unless expressed in signal-to-noise ratios. An improvement of only .01 in a high reliability coefficient is equivalent to the increase in reliability obtained by lengthening the test 10 per cent or more. One can readily appreciate the practical significance of the improvement in these terms (82, p. 375).

I have news for Professor Stanley. The practical significance of an increase in a reliability coefficient of .01 is negligible whether it represents an increase of 10, 100, or even 1000 per cent in the length of the less reliable test. Reflect on what use might
be made of the reliability coefficient. The signal-to-noise ratio is as useless as any other reliability statistic. Another example of the rule is provided by Cureton (23), who calculates what he calls a coefficient of stability by correcting a between-forms correlation for attenuation using KR20 reliability estimates. The KR20 estimates are directly affected by the dispersion of item difficulties and by item heterogeneity. The between-forms correlation is not so affected. The correction seems likely to produce in many cases a gross overestimate. Cureton obligingly provides one example where the resultant correlation is greater than 1.0. There is only one word to describe coefficients like this. The word is: Baloney! (Cureton 22).

Cureton and five others (24) compare the standard errors of measurement for six tests (one each?) and find confirmation for Lord’s (55, 56) assertion that tests of the same length have the same standard error of measurement. The Lord assertion is based on the consideration that when the item intercorrelations increase the reliability increases but so does the test variance, and these effects compensate to give a constant standard error of measurement. Consideration of a flogging wall with static canes or very little flog will show that the effect is not general across the possible range of item intercorrelations. The Cureton et al study considers only tests where the mean item intercorrelations are very similar, and therefore it is not a very interesting confirmation.

Loyal members of the Model T club will find some pretty algebra in Kristoff’s (46, 48) papers on the reliability of reliability estimates. Joreskog (44) continues his saga of the statistical treatment of congeneric tests. Zimmerman (103) gives the umpteenth rederivation of KR20 with “relaxed” assumptions. The derivation is extremely rigorous but, as always, it turns out that the only feasible way that the relaxed assumptions can be met is if the original (49) stringent ones are met. Kristoff (47) shows that for tests differing only in length signal-to-noise ratios are additive. Lord (63) considers partial correlations corrected for attenuation. He finds, sadly, that the sampling error is “overwhelming” but gives a more powerful test of the hypothesis that the sign of the partial is positive or negative! Ebel (28) gives a very neat demonstration of why a longer test is usually more reliable than a shorter test. Bay (2) shows that the sampling distributions and standard errors of reliability estimates under an analysis of variance model are substantially affected if the distribution of true scores is not normal. Jackson (40) attempts to estimate true score variance and error variance. In order to keep the algebra tractable, or for some other reason about which he does not care to inform us, Jackson assumes that true scores are normally distributed and that the scores for repeated measures on parallel forms are independently and identically distributed.

A different and promising approach to the problems of reliability (and some others as well) is given by the information measure developed by Birnbaum (4). This measure I(θ,x) is considered by Birnbaum to be a kind of index of precision which for a given test and scoring formula reflects the information provided by the test in the vicinity of a given value of the attribute θ. It should be noted that this measure is not based on any of the Model T assumptions. This is why it is described as promising. Birnbaum provides a formula for this which is based on the slope of the score on ability regression line in the vicinity of a particular value of θ. A little
reflection on a flogging wall test with static canes will provide an intuitive justification of the procedure and illustrate some of the determinants of the measure. The flogging wall test will give much information at ability levels where the canes are dense and little information where there are few canes. It is clear that the slope of score on ability will be steep at points where the canes are dense and that at these points \( I(\theta, x) \) will be large; where the slope is flat (few canes) \( I(\theta, x) \) will be small. When the canes flog, nothing much will be changed except that it will not simply be the local density of the canes that will determine the slope of score on ability.

Birnbaum\(^3\) gives a mathematical justification of his information measure by demonstrating that it is inversely proportional to the width of the confidence interval we would construct in estimating ability \( \theta \) from test score \( x \). This demonstration requires three highly questionable assumptions about \( F(x/\theta) \), the cumulative distribution function of the test scores. These assumptions are: 1. \( F(x/\theta) \) is normal; 2. \( F(x/\theta) \) can be approximated by \( 1\frac{1}{2} \) terms of a Taylor expansion; and 3. \( F(x/\theta) \) is \ldots “continuously strictly increasing in \( x \) and decreasing in \( \theta \)” (4, pp. 417–18).

The first assumption is palpably false for any test unless we avoid the edges of the test. It may be false elsewhere if the attenuation paradox region is approached. Consider a flogging wall test with little flog where items are irregularly dispersed in difficulty. Birnbaum stops the Taylor expansion and rejects terms involving the derivative of the variance, justifying this by assuming that changes in the mean with respect to \( \theta \) dominate changes in the conditional variance:

\[
\frac{dF(x/\theta)}{d\theta} \quad \frac{dV(x/\theta)}{d\theta}
\]

For many tests this second assumption is grossly violated (Ross 79). Consider a flogging wall test as we pass from a region of high density. The mean will change slowly but the variance will decrease sharply and then bounce back as we exit the region of low density.

The third assumption is used to define the estimates of the upper and lower bounds of the confidence interval. Following Birnbaum (4, p. 409), we will consider the case for the lower bound; the case for the upper bound is the same. Birnbaum defines the lower \( \alpha \)-level estimator of \( \theta \) based on the test score \( x^* \), \( t(x^*) \) as that value of \( \theta \) for which \( F(x^*/\theta) = \alpha \) and calls this value \( \theta^* \). In order for \( t \) to be well defined there must be only one value \( \theta^* \) which fulfils this condition for each \( x^* \) and Birnbaum’s assumption is sufficient to guarantee this. This assumption is also sufficient to guarantee that if \( x < y \) then \( t(x) < t(y) \). That the estimator is monotonic with test score allows Birnbaum to relate the distribution of estimators to the distribution of test scores and so define the confidence interval. Note that it is not sufficient to guarantee the existence of a bound.

The third assumption is sufficient to permit definition of the confidence limits if the bounds exist. It turns out to be much more restrictive than appears on the surface. If, for example, \( F(x/\theta) \) takes the form of the logistic (not very different from

\(^3\)I am indebted to Dan Milech, University of Western Australia, for the mathematical argument to follow.
the normal), it can be shown that the assumption implies that the mean test score \( \mu(x/\theta) \) is strictly increasing in \( \theta \) and that the variance \( V(x/\theta) \) is strictly nondecreasing in \( \theta \). This latter restriction is likely to be violated in many tests (79) and in particular cannot be met by a peaked test. For other distribution functions the restriction on variance may not be so simply describable but seems likely to be equally severe.

It is possible that the assumptions of the information measure can be met, at least approximately, by a test with items of equivalent discriminatory power dispersed uniformly across an ability range that is not too broad. In this case the slope of score on ability will be approximately constant and it will be sufficient to take as the information measure the inverse of the conditional variance \( V(x/\theta) \). It is not known just what effect violation of the assumptions will have on the Birnbaum information measure. The measure should be used cautiously or not at all until this matter has been investigated.

Lord (65) considers the Birnbaum relative efficiency index defined as:

\[
RE(x, y) = \frac{I(\theta, x)}{I(\theta, y)}
\]

and shows that by rescaling \( \theta \) to true score values for \( x \) and then for \( y \), \( RE(x, y) \) is equal to the ratio of the variance of obtained scores on the two tests for fixed corresponding true scores multiplied by a scaling factor which varies with \( \theta \), the attribute value. Thus Lord very cleverly makes the slopes constant at unity but pays for this with the variable scaling factor. The scaling factor requires the computation of the density functions of the distribution of true scores for the two tests and the formation of the ratio of the ordinates at corresponding (equipercentile) points. Lord finds good agreement with relative efficiencies computed by the Birnbaum method and at a considerable saving of computer time. Granted the shaky Birnbaum foundation, the Lord derivation is a mathematical tour de force, but one wonders whether the scaling factor ratios will always be as well behaved as they appear to have been in the study. With tests of unequal difficulty, for example, a small change in true score for one test may correspond to a large change in true score for another and the ratio of the ordinates may be unstable.

It is a little puzzling that test theorists have neglected the other regression line, the regression of ability on score. If the scatter diagram can be formed, it would be simple and very informative to examine the variance of ability for particular values of the score. An information measure based on the inverse of the conditional variance would seem to do everything useful that the Birnbaum measure does and to require no elaborate justification. The required scatter diagram can be computed from the conditional variance of score on ability if we are willing to make an assumption about the distribution of ability. This would be mild compared with the assumptions of the Birnbaum argument and, in any case, could always be perfectly met with the artificial data so frequently used for scaling studies. A statistic of this type would have many advantages over the reliability statistics currently misused. In the first place, the statistic is attribute based and gives a direct answer to an important question: how well does the test score represent the attribute? In the second place, it will often vary with score and be known to vary. Finally, it is not a regression coefficient and test theorists will be less likely to attempt foolish things with it. The measure will be affected not only by noise (item flog) but also by item
difficulty dispersion. For most practical purposes it will be unnecessary to attempt to separate the components (Ross & Lumsden 80).

I can find only three propositions about reliability worth remembering by test users or constructors:

1. Test scores are unreliable.
2. All other things being equal (item type, dispersion of item difficulties, item correlations) and up to a certain point a longer test is better than a shorter test. By 'better' it is meant that the test score represents the attribute more accurately. The 'certain point' is a function of subject fatigue or boredom.
3. All other things being equal (item type, dispersion of item difficulties, test length) and up to a certain point a test with higher item intercorrelations is better than a test with lower item intercorrelations. The certain point is reached when the item intercorrelations become high enough for the attenuation paradox to occur. The attenuation paradox will not occur if the item difficulties are adequately dispersed.

It should be noted that these propositions do not relate only to problems of reliability and that all can be derived from a contemplation of the flogging wall test without using the concept of true score. Yule's evil suggestion has yielded a meager fruit and has devoured the energy and skill of some of the finest minds in psychology. It is tempting to believe that this is an illustration of the Nimzovitch dictum: "When there is no good move a botch will come along to fill the breach." But there has always been a better move, and wise test users—some of them, curiously, Model T buffs—have often played it. Thus they have chosen tests in terms of their superior validity. They have advocated the use of expectancy tables (e.g. 1, 18) and have derived their expectations and confidence limit information directly from these rather than indirectly and uncertainly from reliability theory statistics.

The conclusion is obvious but I will state it anyway. Reliability theory in its present form should be abandoned. To echo Bock & Wood (7), there seems no point in further prolonging the lives of the true score or the reliability coefficient, classically defined, as concepts in test theory. The problems of reliability should be assimilated into validity and scaling theory where they have already been partly solved. Nothing in this recommendation should be taken as advising students of practice, transfer, or training effects on test scores against calculating test-retest or between form coefficients. These coefficients should not, in my opinion, be called reliability coefficients, but if they are, care should be taken that they are not confused with reliability coefficients as classically defined.

Macbeth: ".............. this is a tale
Told by an idiot, full of sound and fury,
Signifying nothing."

Unidimensionality

The topic of unidimensionality continues to be neglected. Most of the articles considered in the section on test scores pay lip service to the notion, usually in the form of a bland statement about the assumption of local independence, a single
latent trait or the like. Nowhere in the period under review is there any evidence of a serious attempt to construct a unidimensional test, and there is only one attempt (38) to suitably test an assumption of unidimensionality. Yet it is clear that most of test theory, not merely esoteric scaling procedures, depends vitally on the assumption. A simple example is the much admired KR$_{20}$ which may be a gross underestimate of reliability unless the test is unidimensional.

The importance of unidimensionality does not depend only on its function as a quasi-mathematical assumption for reliability and scaling theory. The whole conception of psychological testing as measurement depends on it. This has never been more clearly put than in an early, and much neglected, paper by McNemar (71, p. 268):

Measurement implies that one characteristic at a time is being quantified. The scores on an attitude scale are most meaningful when it is known that only one continuum is involved. Only then can it be claimed that two individuals with the same score or rank can be quantitatively and, within limits, qualitatively similar in their attitude towards a given issue. As an example suppose a test of liberalism consists of two general sorts of items, one concerned with economic and the other with religious issues. Two individuals could thus arrive at the same numerical score by quite different routes. Now it may be true that economic and religious liberalism are correlated but unless highly correlated the meaning of scores based on such a composite is questionable.

Jones (43) makes it clear that measurement is always of an attribute or property of an object or event. He goes on to state that: “no attribute can be observed unless man has arrived at some concept of it” (p. 337). We may paraphrase this as: the beginning of measurement is the conception of the measurable attribute. How can we make any claims to measure if our measuring instrument has a number of different sorts of items based presumably on different attribute conceptions? Jones (pp. 350–57) makes clear his support for the use of unidimensional tests in a statement that echoes (with some irritating imprecisions) McNemar’s.

Foolish things continue to be said about unidimensionality by people who should know better. Thus Guttman (36) asserts that unidimensional scales should be discovered not constructed, and that item culling procedures to obtain a unidimensional set are illegitimate, representing an unworthy capitalization on chance and item diversity. Guttman is right only if the test constructor fails to repeat his work with other samples and if he fails to write a specification for the items and to test that by constructing new items that fit the specifications. Item culling procedures will always be necessary because any specification is likely to be incomplete and because of human imperfections in carrying out the specifications. Suppose that a set of 20 items was constructed and that 19 of them were perfectly scalable and the other was a maverick. Most test constructors would arrange a stealthy murder of the culprit and burn all the records of its birth. What would you do, Professor Guttman? Cronbach (18), in an otherwise excellent paper on validity, asserts that a test constructor should not attempt to maximize item intercorrelations because this may mean that important criterion relevant items will be omitted. Cronbach et al (20) make the same point when they suggest that narrowing of the universe
may lead to a reduction of validity. But nowhere have the advocates of unidimensionality suggested that only one unidimensional test should be constructed in a given area or for a particular purpose.

There seems to be, despite Jones' remarks, a popular belief that the construction of unidimensional tests is difficult. Hambleton & Traub (38) assert that construction of a unidimensional test is very difficult and that the factor analytic procedure is dubious because of difficulty factors. With modern computers there should be no problem. The factor analytic procedure (27, 68) usually converges rapidly if the items are constructed according to a strict specification. Lumsden (Unpublished Masters thesis, University of Western Australia, 1959) found that with four number series tests item rejections were few and with one test a Guttman coup was achieved, all items being retained. Normally it should not be necessary to use tetrachorics. McDonald & Ahlawat (70) have shown that difficulty factors will not arise with phi-coefficients if regression is linear and that generally difficulty factors will be negligible unless item difficulties are widely different or the item discriminatory power is high. For these latter cases it usually should be sufficient to carry out what may be termed cascade analysis. Thus we could divide the items into three difficulty levels and analyze each level in turn, starting with the easiest and passing the more difficult selected items from each analysis into the analysis for the next level.

It is possible that the worry about the difficulty of constructing unidimensional tests arises from a confusion between unidimensionality and theoretical singularity. A unidimensional test does have a single attribute but the attribute is complex. Consider the following sample item:

1 2 4 7 11 ...  

A test with items of this sort is unidimensional and maps the ability of the subject to do items like that (see the validity section for a specification of these items). The ability is complex, involving number and reasoning abilities and probably others as well. It does not reflect a single theoretical attribute or construct in the Cronbach & Meehl (21) terminology. It may be thought of as a compound with constructs as elements. The construction of theoretically singular tests is probably impossible (69).

Special problems in constructing unidimensional tests are likely to be encountered in the personality domain. Vernon (97), in a very early paper which may have been the first to use the term "unidimensional," pointed out that it is usually not feasible to ask the same question many times. This has not deterred all test constructors. Jackson (39) discusses his attempts to produce unidimensional tests of personality by maximizing item-test correlations. He employs an additional constraint that social desirability correlations should be low for each item. The method is not the most effective and in most of the cases reported does not appear to have succeeded very well. Neill & Jackson (74) examine different item selection techniques in the development of an Evaluation Sensitivity test. They found that different methods of selecting items produced tests with about the same K-R20 values (n = 40 K R20 = 93). The method of forming the item pool from a detailed specification is admira-
ble, but the item culling procedures are not as effective as the residual minimization procedure (68).

Kirkham (University of Western Australia, personal communication) made the interesting observation that groups of second year university students were remarkably accurate at estimating the relative difficulty of items from a unidimensional number series test. Both paired comparison and magnitude estimates were used, and the correlations of judged difficulty with proportion passing for a large sample of children were of the order of 0.9. This finding is important because it could lead to a useful saving of time in item tryouts.

**Validity**

Most test theorists have fled from the field of validity, apparently considering that it is adequately handled by regression procedures or that it is too difficult for simple model makers. On the first consideration they are partly, and on the second wholly, justified. Validity theory has been kept alive by the efforts of Cronbach and various associates.

Cronbach's (18) chapter gives a comprehensive statement of what has been achieved in validity theory. The concentration throughout is on validation as a process of coming to some understanding of the meaning of test scores for various applications of the test procedure. All the standard approaches to validation and the relationships between them are discussed and illustrated. There is a great deal of practical wisdom in the chapter, no doubt derived from Cronbach's concern with problems of using tests in educational settings. I could find only two flaws worth remarking. His advice on test construction should not be accepted blindly. His treatment of his own work on utility interpretations of validity is trenchantly brief and quite unsatisfactory.

Cronbach & Gleser (19) performed a great service in incorporating predictive validity into the general area of decision processes and in emphasizing the benefits to be derived from sequential procedures. The book is not the easiest to read, and it is good that simpler but adequate descriptions of the approach are beginning to be incorporated in textbooks (100). It is curious that no chapter in Thorndike's edition of *Educational Measurement* (87) makes more than a glancing reference to sequential procedures. Test users should not be too much dismayed by the formidable problems of estimating utilities. It is true that absolute utilities are usually difficult and sometimes impossible to estimate. Relative utilities are usually much simpler to estimate and are adequate to support rational choice of procedures.

Test publishers could assist greatly in the application of the Cronbach & Gleser procedures by including suggestions for sequential strategies in their manuals. It would also be helpful if they would adopt suggestions (e.g. 1, 18) to provide validity evidence in the form of scatter diagrams (expectancy tables) rather than via the usual correlation coefficients. Inferences concerning the effect, for example, of changing cutting points can be made more directly and more quickly than by the use of Taylor-Russell tables or the like. Where validity data are available for only a relatively small sample, as is often the case for quite legitimate reasons, a theoretical expectancy table should be constructed from the correlation coefficient and the
usual assumptions. But the generating scatter diagram should be given too so that the test user can add his cases as they arise, eventually correcting any errors in the original estimate. This procedure has the additional merit of allowing for any local variations in the testing or criterion estimation situation.

An interesting approach to the solution of the Cronbach & Gleser equations for the two-stage selection process is given by Rock, Barone & Boldt (76). They ignore costs and propose that the number of applicants undergoing second stage testing should be proportional to the ratio of the squared part correlation of the second stage test with the criterion to the squared zero order correlation of the first stage with the criterion. They justify this with the argument that if the second stage contributes nothing additional to the first stage, then nobody should do it. This is unexceptionable, but they go on to argue that if the second stage adds as much (or more) as the contribution of the first stage, then everyone should do the second stage. But in this case it would normally mean that the second stage would have a higher zero order correlation with the criterion than the first stage. Then surely the test of the second stage should be given first; and it would be unless the first stage test was very much cheaper than the second stage test. But this would mean that fewer subjects should be given the second stage test. It seems that no rational solution to the Cronbach & Gleser problem can be obtained without taking some account of costs.

Sequential testing, where after each item a decision is made to accept, reject, or continue testing, was examined by Linn, Rock & Cleary (50). They compared sequential and conventional short tests on accuracy of assignment to high and low groups on full College Board College Level Examination Program tests. They found that sequential tests required only about half the number of items as conventional tests for the same accuracy of classification.

A comparison of predictor selection techniques using Monte Carlo methods was provided by Rock et al (77). They found, not unexpectedly, that criterion-dependent methods were superior to criterion-independent methods. They also found that forward selection methods were superior to backward elimination methods. Campbell & Ignizio (12) used linear programming methods for predicting student performance and found the prediction less biased for extreme cases than the least squares procedure. Jackson & Novick (41) examined the problem of optimizing the time allocation between various battery components in order to maximize battery validity with a fixed total testing time. They worked out an example but provided no test of the efficacy of the procedure.

Two papers (13, 14) discuss the problems of suppressor variables and conclude that the search for suppressor variables probably had better be suppressed. There is usually more to be gained from finding a predictor for unaccounted criterion variance than in attempting to suppress irrelevant predictor variance. Conger considers the charming possibility of mutual suppression where two variables each have positive correlations with the criterion and negative correlations with each other. Such pairs are likely to be as rare as unicorns.

A number of papers consider the multi-trait multi-method approach of Campbell & Fiske (11), a procedure that is an elaboration of congruent validity. There are
considerable practical and theoretical problems in its application. It places great strain on the art of the test constructor to produce tests of the same trait with different methods. The problems are analogous to those of parallel form reliability where a test is held to be reliable only if it has an identical twin. The multi-trait multi-method procedure seems to require that a test have an identical twin of the opposite sex. Krause (45) points out that convergent confirmations may simply reflect unconscious bias in test construction by the inclusion of some common sources of variance not related to the trait in question or to the method. Similarly discriminant confirmations may represent a failure of test construction rather than trait divergence. There are problems of additivity of method and trait variance, and with factor analytic procedures the possibility of correlation between content and method factors must be considered (8, 9). Fair agreement was found in two cases between analysis of variance, factor analytic, and inspectional procedures. It is clear that the multi-trait multi-method approach should not be employed mechanically. The production of $9 \times 9$ matrices with little thought given to the selection of traits and methods should cease. It seems likely, however, that where great care has been given to test construction, where much is known a priori about traits, methods, and the relationship between them, the validation procedure will more and more resemble construct validation and the explicit setting up of the multi-trait multi-method matrix will be rare. It is clear that the method is not worth using badly. Is it worth the trouble of using it well?

Lumsden & Ross (69) examine the construct validation program of Cronbach & Meehl (21). They argue that the program requires: (a) test unidimensionality and theoretical singularity, (b) operational criteria for all the theoretical terms used to describe tests, and (c) multiple theoretical linkages for the terms. They point out that individual difference measures based on performance are intractably complex and that theories of individual differences are not sufficiently developed to meet the third requirement. Construct validation of tests is, therefore, impossible. It is possible that a refutation of this argument, or a way out of the difficulties, will be discovered. It does seem very unlikely, however, that in any foreseeable time psychology will be able to construct and validate the equivalent of 100 or more thermometers.

A lower keyed approach to the validity problem seems necessary. One possibility is via an extension of some content validity notions. The suggested procedure is not my creation. For reasons of delicacy I withhold the mother's name, but possible fathers of the idea are Ebel, Guttman, Guilford, and Cronbach. We begin by constructing a unidimensional test of, say, 20 items. We characterize the test by saying that it measures the ability to do items like that, pointing to any item of the set. Notice that it is impossible to do this with a heterogeneous test since no single item can be considered as characteristic of the set. We now attempt to write specification sentences for the item set. Suppose that the items in the set are all number series where the subject is required to write the next two numbers in the series and that they are all of this type:

\[
1 \ 2 \ 4 \ 7 \ 11 \ \cdots \ \cdots
\]
The specification sentences for the items might be as follows:

1. They are increasing number series.
2. The differences are in increasing A.P.
3. The first difference is greater than zero.
4. The numbers in the series are all positive integers less than 100.

Suppose that we now write many items from the specifications and find that they all meet the unidimensionality criterion with the original set. We now have confidence that we know what we are doing and can make judgments about set membership without empirical evidence. In the language of automata theory (Rogers 78), the original set was recursively enumerable but is now also recursive. If counterexamples occur during the testing of the items generated from the specifications, then the specifications will be altered. Of course our confidence in the specifications can never be absolutely complete since there remains always the possibility of a counterexample appearing. It is a species of theory confirmation: a hundred or even a thousand confirmations cannot produce logical certainty, but most of us will proceed as if it did.

So far this is simply content analysis carried out rather carefully under conditions which permit confirmation or disconfirmation of the specifications. Suppose we now ask the questions: are the specifications too stringent? Are there other types of items which belong in the set? We can treat each specification in turn as a facet for possible variation. Will letter series (with appropriate variations to specifications 2 and 4) be found to belong in the set? Will decreasing number series (i.e. reversals of original items) belong in the set? If the answer is yes to any of these questions, then the specification can be broadened. If the answer is no, then the next question is: do the new set of items themselves form a unidimensional set? If not can a sizable group of items which meet a unidimensionality criterion be found in the set? Proceeding in this fashion it seems likely that a well-constructed test will frequently give rise to a family of tests differing only in one or two clearly described facets. We can now look at the intercorrelations between tests in the family and even more importantly at the correlations with exterior measures. From these results and with a mode of thinking characteristic of factor analysts we may come to some understanding of the meaning in ability terms of the facets and the exterior measures.

The approach is low keyed and does not pretend to meet the challenge posed by Cronbach & Meehl. It does seem appropriate for factor analysts who wish to have some basis for factor naming and in any event provides some guide for test constructors. The method seems far superior to the much trouble much muddle procedures currently in vogue.

**Test Scoring**

Test scoring or scaling is a topic close to the heart of the problems of test theory. How can we best use the information given us by the item performances of our subjects? How well do the test scores represent the attribute? These are central questions for test theory. It is pleasing to be able to note that test scaling and related problems have been the subject of some of the most energetic work of the period. There are a number of interesting papers; none are inspired but some are elegant
and a few even suggest procedures that may be usable. Most of the papers on scaling use forms of the relatively well developed logistic models. This is, I believe, wise. Unfortunately, all papers which do not rely on simulated data continue a scaling tradition of using inappropriate test material. If the model requires a unidimensional test, then a unidimensional test, or a near facsimile of one, should be found or constructed. It is useless, and bad publicity, to run a Cadillac scaling method on paraaffin item sets.

Most studies using real data make no serious attempt to evaluate the worth of the suggested procedures. What can it mean to say that one scaling method is better than another? The only answer that makes any sense is to say that it means that one method gives a more accurate representation of the attribute values than the other. If this is the case, then the better method will have higher correlations with any criterion for which the attribute is relevant. This test is sure fire and should be completely convincing to a potential user of the method. The correlation test should be carried out routinely, pitting the new procedure against alternatives, one of which should always be that old trouper the normalized standard score. The test has been used many times (83, 98), particularly in the 1920–1940 period, and the results have usually shown that differential weighting of item scores made no appreciable difference to either the reliability or the validity of the test. I have found no instance of the use of this test on real data during the review period. Why?

Goodness-of-fit procedures have been employed (6, 38) and several writers (5, 37, 59–62, 65) have used forms of the Birnbaum information measure whose possible deficiencies have already been noted. But even if the procedures were not objectionable, they can never be as convincing or as direct as the correlation test, particularly when plausible alternatives are not considered. And in any case the test user is left wondering what it all means. Should he spend his dollars on fancy scoring or on more testing? Users of goodness-of-fit tests should ponder Torgerson's (91) discussion of the trading relation between assumptions which sometimes yield good fits when the assumptions have been clearly violated. It is worthy of note, and to his honor, that Rasch (75) was concerned when the fit was too good.

Lord (58) attempts a direct confrontation of the Birnbaum logistic model by attempting to estimate the form of the item characteristic curve by plotting proportion passing against true score. He finds excellent agreement between item characteristic curves for SAT verbal items obtained by this method and those obtained by Lord (57), under the assumption that all were three parameter logistic curves. Lord finds these results impressive. So they are—but not for the reasons given by Lord. In support of his procedure, Lord (58, p. 42) presents the following argument: "Now the true score on a test and the latent trait measured by the test are precisely the same thing, except for a possible monotonic transformation of the scale of measurement. Thus the regression of item score on true score is a characteristic curve of the item."

This argument is absurd in two ways. Recall that the true scores have to be estimated—only the Devil has access to the book in which the expectations are recorded. If estimated true score is a monotonic transformation of the latent trait, then so is the obtained score. Lord could have saved himself a lot of trouble by
simply plotting proportion passing against normalized obtained score. Ross (79) did exactly this and found excellent fit for the logistic. Secondly, Lord is mistaken in thinking that monotonic transformations, even if strictly increasing, make no difference to the form of the item characteristic curve. Try this simple experiment. Draw two normal ogive item characteristic curves with different location values but the same slope. Now apply some monotonic transformations to the score scale. You will find that the shapes of the curves will differ a great deal from the originals. They will also often differ from each other.

A maximum likelihood solution for the normal ogive model is supplied by Bock & Lieberman (6). They find a good fit between theoretical and empirical score distributions for one section of the Law School Admission Test and a not-so-good fit for another. They suggest that the difference may be the result of the greater heterogeneity of items in the second section. They note that proportion passing and item test correlations give excellent estimates of the maximum likelihood solutions for location and slope parameters. Urry (96) gives a graphical procedure for estimating item parameters. Jensema (42) found, for the three-parameter logistic model, a high correlation between the Urry graphical estimates and maximum likelihood estimates, and suggests that with further development the expensive maximum likelihood process may no longer be needed.

The one-parameter Rasch model is compared with two- and three-parameter logistic models by Hambleton & Traub (37). Using simulated data generated by a three-parameter model, they compare the three models using information curves and relative efficiency estimates. Not unexpectedly they find the three-parameter model best when guessing is a factor but only at the lower levels of ability. Where guessing is unimportant and the range of slope parameters is not great the Rasch model is quite efficient. Hambleton & Traub fail to emphasize one of their most important findings. The Rasch one-parameter model is remarkably efficient except at the very lowest ability level when guessing was important (C = 0.20) or when the range of slope parameters (0.19–0.99) was far beyond that conceivably tolerable in a test alleged to be of a single trait. In the Rasch model the simple number right score is a sufficient statistic for estimating the attribute, so there is still a lot of life left in unit weight scoring.

Hambleton & Traub (38) proceed with a comparison of one- and two-parameter logistic models using data from the SAT Mathematics and Verbal Test. This study would be a model of what not to do except that some others are worse. Remarkably, they do a sensible test of undimensionality by factorizing the items. The test fails; more than one factor is required to account for the item intercorrelations. But they proceed anyway! They use fit between theoretical and empirical score distributions as the criterion and find that the two-parameter model does somewhat better than the one-parameter Rasch model. The difference is greater with the shorter test, which is consistent with Birnbaum's (4) observation that with long tests weighting items makes little difference.

Bock (5) points out that in multiple choice items some wrong answers are better than others. He outlines a very elegant maximum likelihood procedure for weighting the various responses whether right or wrong to obtain an estimate of ability. The
computational problems are very formidable for any large number of items. Bock, using a 20-item vocabulary test, compares the procedure with a procedure using dichotomous scoring and finds a moderate increase in information as measured by the Birnbaum index. No external criterion is used and no comparison is made with the Guttman (33) suggestion that the response options be weighted by the criterion scores of subjects giving a particular option. Dalrymple-Alford (25) tackles the problem of wrong answers in multiple choice tests by requiring the subject to continue giving answers until he is correct. He then calculates an average uncertainty measure for each subject. A useful way of collecting the data for this would be to require the subject to order the options in terms of his preference for them as answers. This procedure is simpler and quicker than the De Finetti (26) procedure of asking the subject to assign probabilities of correctness to the option. It would be interesting if, for example, on an easy item best discrimination was given by the second rather than the first choice.

Lord (64) considers the nasty problem of omitted responses and suggests a method which assumes that if subjects were obliged to respond to omitted items, they would choose randomly. This assumption is probably incorrect but must be closer to reality than the usual assumption that the subject would always choose wrongly. Bock (5), in the paper discussed above, simply treats “no response” as another option. There is a problem here in that “no response” is often given a higher weight than some of the incorrect options, if this became known after the test was published, the proportion of omissions might increase and the weights require drastic modification.

Angoff (1) gives a very clear account of the various types of derived scores and norms commonly used in education. His nonmathematical description of the item characteristic curve approach to scaling is very lucid but unwontedly brief. He goes on to a technical consideration of methods of establishing norms and equating scores. Angoff (1, p. 522) resurrects an old notion of Thurstone’s (89) and suggests that mental age norms could just as logically be developed by finding the mean age of children who get a certain score, i.e. by using the age on score regression line rather than the score on age regression line. This is incorrect. The age on score regression line is the appropriate one if we wish to estimate a subject’s chronological age from his score. But we already know his chronological age. The Angoff suggestion could lead to the situation where a 10-year-old boy who had the average score for his age could be described as bright or dull (67). Cooley (15) looks at techniques for considering multiple measurements. He points out difficulties involved in staring at the parallel stalks of the usual profiles and then goes on to discuss with admirable clarity and a simple example the trait-space model. He illustrates the use of centours, a chi square test for profile conformity, discriminant analysis, multiple regression, and canonical correlation. There is a problem with profiles that is not considered by Cooley and is not completely solved by the centour approach. The problem is that the profile of averages for chemists may not represent the average chemist or indeed any chemist. There is an exactly parallel problem in the determination of learning curves. For this Tucker (94, 95) developed the Eckhart-Young procedure which analyzes the results into families of learning curves, each characteristic of
some learners but not others. There seems no reason why this approach could not
be applied to profiles.

There are a number of important papers on tailored testing. Lord (59–62) consid-
ners the problem first as a computer application and then, with an eye to practicality,
ways of approximating tailored testing with paper and pencil tests. Lord (61) points
out that shrinking step size procedures are impracticable because they require an
impossibly large number of items even for modest test length. Fixed step size up and
down procedures are about as good as the shrinking step size procedure Hybrid
procedures which begin with a few large steps and then continue with small fixed
steps appear an attractive compromise, analogous to artillery ranging, but in the
Lord simulation were no more effective than simple fixed step size. It should be
noted that in these studies Lord assumed a rather low value for the slope parameter
(Ag = 0.5) and that some of his conclusions may need to be amended with more
discriminating items. Lord (59) considers two-stage testing and concludes that with
no guessing two-stage procedures may be about as effective for all but extreme ability
levels as the best up and down technique. With guessing possible 20% of the time,
up and down procedures were clearly superior. Again Lord only considers peaked
tests with rather low item slope parameters. The flexilevel test (62) is another
attempt to provide tailored testing with paper and pencil presentation. In flexilevel
tests the subject knows whether he was correct or not on each item. Subject answers
first an item of median difficulty. If correct he moves to the easiest item of above
median difficulty; if incorrect he moves to the hardest item of below median diffi-
culty. The examinee attempts only (N+1)/2 items in the set, which has a rectangular
distribution of item difficulty, and it turns out that even though examinees do
different items of different difficulties, the number right score is an excellent estimate
of ability. Using a form of information measure, Lord found that near the middle of
the range for which the test was designed a flexilevel test is slightly less effective than
a conventional peaked test. At other points in the ability range the flexilevel test was
better. It should be noted that the flexilevel test is after a point an increasing step
size up and down procedure and should be inferior to other types of tailored tests.
All of the work described above was done with computer simulated data, and Lord
was able to test a very large variety of tailored procedures. He points out that his
own feet-on-the-desk intuitions about suitable procedures were often shown to be
wrong. Lord's results in this series of studies should be accepted only tentatively
since they depend on a possibly inappropriate information measure.

Waters & Bayroff (99) compared various fixed step size up-down procedures with
conventional tests either peaked or with normal or rectangular distributions of
difficulty. All tests were computer simulated, and the criterion was the correlation
between score and ability. It was found that a peaked conventional test was best at
low item biserials (r_{it} = 0.30) but various branching tests did better, but not much
better, when item biserials were higher. Jensema (42), using computer simulation
and some real data where tailoring was simulated by selection of results, examined
Bayesian procedures for estimating ability. He suggests that tailoring is likely to be
most beneficial when the item discriminating power is fairly high (Ag > 0.8). He
found that the use of prior information did not lead to a substantial reduction in
the number of items required for a given precision or increase in the correlation
between estimated ability and true ability. There is some gain from the prior infor-
mation when the correlation between prior estimate and ability equals 0.9. With a
prior like that the posterior can be given the boot. Test something else instead.

With some trepidation, considering Lord’s experience, I suggest that a useful
two-stage procedure would be to give a short test with items peaked at 0.25 and 0.75
difficulty levels for the population concerned. On the basis of the pretest assign
subjects to three groups to do different tests with items distributed uniformly in
difficulty: (a) from 0.99 to 0.50, (b) from 0.75 to 0.25, and (c) from 0.50 to 0.01.
This procedure with item slopes greater than 1.0 should give excellent discrimina-
tion across the range and permit, if required, the formation of the person characteris-
tic curve. It is probable, however, that it will prove more useful in practice to
assimilate tailored testing into the sequential procedures discussed in the validity
section (Green 30). Thus on the basis of a relatively short test, accept or reject
decisions can be made for some of the subjects and the remainder can then be tested
adequately by a test with items in the range 0.75–0.25 or even by a peaked test.

It is necessary to reconsider the whole approach to test scoring problems. The
item characteristic curve approaches and various estimation procedures are reason-
able well developed, and problems with the information measures probably will be
solved fairly soon. Yet nothing much seems to be happening except that those
obedient monsters, the computers, are now doing almost as many tests as the
disobedient ones. There is need for a more rational and more determined approach
to item writing and item selection. The construction of unidimensional tests seems
possible in many areas, and it is likely that high values of the slope parameter may
often be obtainable (Ross 79). This would make tailored and sequential testing more
attractive. It would make conventional testing with dispersed item difficulties very
attractive too because there will be no need to use tests of more than 15–20 items.
It is important that test theorists should attempt to help themselves here. The tacit
acceptance, amounting at times to advocacy, of poor tests should cease. It is prepos-
terous that test scaling should be the only branch of experimental psychology where
the experimenter does not set up the apparatus and does not care how his data are
collected.

CONCLUSION

The picture revealed is grim. Little of any consequence has been achieved during
the review period; nor can we look with any great pride at the cumulative result for
this century. It is only slightly unfair to say that test theory has failed as theory.
In most areas it does not act as a set of propositions which generates testable
propositions in the content area of interest. It fails in the minimal demand which
can be made of a theory: that it act as an aide memoire. We have seen that reliability
theory has been dominated by an inappropriate and unfruitful model. Other aspects
of test theory have been generally little developed and rarely applied. This is partly
because the problems are difficult and partly because of the obsession of test theorists
with reliability theory.
The popular view is that test theorists sit in corners playing with their t's and e's and minding their p's and q's. It is further believed that they speak seldom, except to each other, and that what they have to say has no or little practical application. The popular view is substantially correct, but I hope that it will not continue to be so. We must change. We need a new kind of test theorist. The new test theorists (happily, some of the old transformed) will be primarily test constructors and validity people who attempt to realize their dreams. They will not accept the current state of the art in testing as the datum but will try to elevate the art. They will not test a new model with a few items from the SAT files (or from a computer), find a mediocre fit to some dubiously relevant criterion, and then go on to the next. Rather they will set out the requirements for the application of the model and sweat to meet those requirements, testing the model against a user-for-blood standard of efficiency. They will not seek salvation in the epicene elegance of elevated algebras but will prefer vulgar analogies. They will not behave like British gentlemen, as if they had been trained not to notice unpleasant things and never to mention them if they do. They will criticize one another incessantly, zestfully, and, I hope, gleefully. But who will train them?

And so let me end it. "I have supped my fill of horrors."

**Literature Cited**

2. Bay, K. S. 1957. Some latent trait models and their use in inferring an examingee's ability. See Ref. 66, 397–472
23. Ibid 1971. The stability coefficient. 31:45–53
43. Jones, L. V. 1971. The nature of measurement. See Ref. 87, 335–55
51. Loevinger, J. 1948. The technic of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. Psychol Bull. 45:507–30
52. Ibid 1954. The attenuation paradox in test theory. 51:493–504
55. Lord, F. M. 1957. Do tests of the same length have the same standard errors of measurement? *Educ. Psychol. Meas.* 17:511–21
56. Ibid 1959. Tests of the same length do have the same standard errors of measurement 19:233–39
59. Lord, F. M. 1970. Some test theory for tailored testing. See Ref. 30
62. Ibid. A theoretical study of the measurement effectiveness of flexilevel tests, 805–13
65. Ibid. The relative efficiency of two tests as a function of ability level, 351–58
73. Ibid 1941. Psychophysics and mental test theory II. The constant process. 48:235–49
80. Ross, J., Lumsden, J. 1968. Attribute reliability. See Ref. 81
93. Tucker, L. R. 1952. A level of proficiency scale for a unidimensional skill. Am. Psychol. 7:408
## CONTENTS

**Personality.** Lee Sechrest  
1

**Cognitive Development.** Herbert Ginsburg and Barbara Koslowski  
29

**Color Vision.** Gerald H. Jacobs  
63

**Biochemistry and Behavior: Some Central Actions of Amphetamine and Antipsychotic Drugs.** Philip M. Groves and George V. Rebec  
91

**Ethology and Comparative Psychology.** William A. Mason and Dale F. Lott  
129

**Models of Learning.** John W. Cotton  
155

**Neurological and Physiological Bases of Psychopathology.** Ralph M. Reitan  
189

**Change Induction in Small Groups.** Morton A. Lieberman  
217

**Test Theory.** James Lumsdon  
251

**Scientific Psychology in France.** Robert Francès  
281

**Engineering Psychology and Human Performance.** Earl A. Alluisi and Ben B. Morgan Jr.  
305

**Consumer Psychology: An Octenium.** Jacob Jacoby  
331

**Psychology and the Law: An Overture.** June Louin Tapp  
359

**Personnel and Human Resources Development.** Frank A. Heller and Alfred W. Clark  
405

**Human Abilities: A Review of Research and Theory in the Early 1970s.** John L. Horn  
437

**Analysis of Qualitative Data.** J. E. Keith Smith  
487

501

**Projective Tests.** Walter G. Klopfen and Earl S. Taulbee  
543

**Program Evaluation.** Robert Perloff, Evelyn Perloff, and Edward Sussna  
569

### Indexes

- Author Index 595
- Subject Index 618
- Cumulative Index of Contributing Authors, Volumes 23 to 27 636
- Cumulative Index of Chapter Titles, Volumes 23 to 27 638