RESEARCH MEMORANDUM NUMBER 19

JUNE 1975

BEST TEST DESIGN AND SELF-TAILORED TESTING

BENJAMIN D. WRIGHT AND GRAHAM A. DOUGLAS

Statistical Laboratory

Department of Education

The University of Chicago

BEST TEST DESIGN
AND SELF-TAILORED TESTING

Benjamin D. Wright
University of Chicago

Graham A. Douglas
University of Western Australia

ABSTRACT

Shows how the Rasch model and a few natural decisions about the nature of tests and their targets leads to simple practical procedures for best test design and self-tailored testing. The error *minimization* necessary for best test design is developed and applied. Tables for *converting* scores observed on self-chosen segments of uniform tests into test-free measures and their standard errors are provided. Robustness with respect to design errors in target location and item calibration is evaluated. and a convenient unit of measurement for psychological and educational variables suggested.

INTRODUCTION

When an examiner wishes to measure a person he must obtain an appropriate measuring instrument. He may do this by selecting from available standard forms or he may draw upon a pool of calibrated items and compose a test tailored to the requirements of his measurement target. He may be a psychometrician constructing an integrated series of standard forms from a *national* pool of well-calibrated items or a teacher making a test for a pupil from items in his files. In each case the examiner requires a clear way of thinking about tests and targets which allows him to deduce from the

nature of the target he wishes to measure the characteristics of the best possible test for the job. We will develop and explain such a way of thinking and derive its practical consequences for how to design, construct and administer best tests.

A simple practical procedure is especially urgent if the examiner wishes to bring test and target into the best possible relationship during the process of testing. Tailored testing (Lord, 1971) tries to do this by various stepwise procedures. Unfortunately most schemes are either expensive, requiring computer assistance, or complicated, placing heavy demands on test-taking *ingenuity.*

The.procedure we will develop lends itself to a simple, practical and inexpensive scheme for self-tailored testing. It makes it possible to present a target person with a stan dard booklet of items sequenced in increasing difficulty and to invite him to choose any starting point in the booklet which suits him. Having chosen, he can work up into harder items until he reaches items so difficult he no longer feels he can do his best and/or down into easier items until he reaches items so easy they no longer challenge him.

If the booklet is properly constructed, the examiner can use the target person's performance on such a variable segment of self-chosen items to estimate his ability. If the items in the booklet are more or less equally spaced on a log-odds scale and have been constructed to more or less. fit the Rasch model, then three easy statistics: the sequence number of the easiest item reached, the sequence number of the hardest item reached and (through a simple self-scoring sheet) the number of *intervening* items succeeded on, are sufficient for reading the target person's estimated measure and its standard error directly from a small family of easy-to-use tables. No need for computer assistance, nor machine scoring, nor lengthy calculations. No need for pretests to identify the right individualized test. The process is self-tailoring. As the target person takes the test he finds for himself the items in the test booklet of difficulty best for him.

BASIC **CONCEPTS** OF TEST **DESIGN**

In order develop such a convenient system of measurement, we must decide what it is reasonable to imagine happens when a person to be tested responds to an item used to make measurements. What could a response we might observe tell us about a person we might wish to measure? What part does the item

play? How do item and person interact to show us something about the person's position on an interesting but difficult to observe variable? How shall we think about this "latent" variable along which item and person supposedly interact? The basic concepts we need to establish and clarify are: latent variable, response model, measurement target and test design.

The Latent Variable

A fundamental problem in the systematic development of our knowledge about the world is how to weave a useful connection between experience and idea. Careful observations are the core of our experience. Our motive for bothering to make careful observations is our idea about how we suppose things to be. We know our specific observations to be limited and incomplete yet we try to see in them indications of general ideas *which* we plan to use comprehensively at other places and times not yet experienced. The concept of a "latent variable" is intended to keep this difference between passing experience and persisting idea clearly before us and so to help us weave a useful connection between them.

The distinction between a latent or underlying variable and a manifest or observable one is analogous to the distinction between the parameter of a model and a statistic intended to estimate it. The parameter represents an idealized conception of all that we wish to know. The statistic represents a particular realization of what little we can observe. Our interest is focused on the latent variable and, whilst it can only be known through its observable manifestations, it is the latent variable which is the motive and meaning of our observations.

In mental measurement the "intelligence," "ability" or "achievement" variables along which both items and persons are supposed to be positioned are latent variables. It is through the calibration of items along these latent variables that we transform a person's observable responses into his "unobservable" measurements. Our interest in his actual responses is transient. Once we have used our knowledge of item calibration to extract from his item responses an optimal indication of his measure we have no further interest in the responses themselves and turn instead to our motive for observing these responses in the first place, namely our wish to measure the person on the fundamental but latent variable. In this discussion we will use the term ,"ability" to

refer to person measure and speak of persons as "dumb" and "smart" and of items as "hard" and "easy." But the latent variable could as well be "achievement" or "intelligence" or some "attitude" or "inclination" or any other attribute for which we constructed relevant observations.

But if a variable is latent and we cannot observe it directly, how can we know what it is? How can we arrive at its operational definition? The answer lies in the items we use to make measurements. To be useful these items must be calibrated along the latent variable. Each item must have its own location, the position at which its difficulty matches the ability of a person for whom that item is just right.[1] When the pool of items from which we select the elements for a best possible test has been calibrated on a latent variable, then these items and their locations on the latent variable provide its operational definition. A measurement of a person on the variable will place him among items with difficulties near his estimated ability. The meaning of his position on the variable will be defined by these nearby items.

Origin and Scale

In general the origin and scale of a latent variable are arbitrary. We must make some specific choice in order to proceed with measurement. If we ask, "What is the distance to New York?", we find that a useful answer depends on settling two prior questions: "From where?" and "In what units?" These are requests for an origin and scale. Thus even for the familiar variable "distance" an origin and scale must be defined before we can make useful measurements.

The origin is the place from which distance

[1] We are not ready to derive more exactly what "just right" might mean. For that we need to specify a model for what happens when a person responds to an item. But we can anticipate our final definition: "just right" will mean that the probability of a correct response is exactly one-half, that is "even odds for a correct response," or "a log odds difference between person ability and item difficulty of zero."

along the latent variable will be "measured." In general there is no unique zero place. We have to decide upon a choice convenient and useful for the measurements we want to make. The same is true for scale. We may be motivated to make a choice of origin which frees us from fiddling with negative numbers. We may be motivated to choose a scale which frees us from decimals or uses the decimal point in an informative way. When no more useful choice is in sight we may even nominate a sample of person measurements or item calibrations as our standard and take their mean value as our origin and their standard deviation as our scale. But whatever our pleasure we will be forced to make some choice. In the discussion that follows we will use either the center of our target or the center of our test as our origin and the log odds for a correct response as our scale. At the end we will recommend a transformation of origin and scale with pleasing properties.

The concept of a latent variable is indispensible for acquiring a grasp of what testing is all about and essential for guiding our attempts to measure. A latent variable is operationally defined by the pool of calibrated items which provides the elements for measuring it. How to identify, select and calibrate items along a latent variable is described in Rasch (1960), Wright and Panchapakesan (1969), Andersen (1972a, 1972b, 1973) and Wright and Douglas (1975).

The Response Model

What do we observe when a person of unknown ability tries a calibrated item of some estimated difficulty? Usually just the binary response "right" or "wrong." [2] In order to use this binary observation as a basis for estimating a person's ability and hence for measuring him we must formulate a plausible relationship between his unknown position, or "ability," on the latent variable, the calibrated "difficulty" of the item and the observable "response." We will call this relationship

[2] More complex observations have been proposed and modelled (Andersen, 1972a, Samejima, 1968 and Bock, 1972) but so far they have found few uses. Since our aim is to formulate and develop procedures consistent with common practice, we will not adventure into the possibilities of more complex observations in this article.

the "response model." By what criteria shall we formulate it?

The response must be reasonable. The way persons and items enter into it must be consistent with what we already know. We also want the simplest model which will do the job. Beyond reasonableness and simplicity we want the measurements resulting from the model to be free from irrelevant influences such as how the items happen to have been calibrated and on whom (Rasch, 1960).

In particular we want to be able to adjust our interpretation of the implications of each observed response for the difficulty of the item prov6king it so that we can reach an "item-free" estimate of the person's ability (Wright, 1967). Finally since making measurements will amount to using a collection of well-designed observations to estimate a person's ability we want our response model to have parameters which can be estimated satisfactorily.

What kind of relationship can we expect between the unobservable characteristics of person and item and the observable response their interaction is supposed to produce? No matter how smart a person may be we cannot be sure he will succeed on a particular item. No matter how dumb, we cannot be sure he will fail. Therefore our only reasonable course is to formulate a probabilistic relationship between person ability, item difficulty and the response they are supposed to influence.

How shall ability and difficulty combine to determine this probability for a successful response? We expect a person's chances for success to increase with his ability but to decrease with item difficulty. Either the difference or the ratio of these two parameters will express these expectations. Since differences are simpler to work with, we will use them.

It only remains to formulate these differences in such a way that,no matter what their values, they will define a probability between zero and one. The simplest realization of these reasonable requirements is the response model for binary observations proposed by Georg Rasch (1960).[3]

[3] For further comments on this kind of model see Birnbaum, 1968, 431-34 and 445-46.

This particular response model meets all our requirements. Not only is the relationship between person, item and the probability of a correct response reasonable and simple, but, as the work of Rasch and many others has so amply shown, the calibration of items can be person-free and the measurement of persons can be item-free.

What about the estimation of these person and item parameters? It has always been popular to take the number of right answers on a test as an indication of the ability of a person. That popular practice is just what the Rasch model calls for. In fact, if we take the widespread belief in the fundamental *meaning*-fulness of an unweighted test score as our point of departure and ask what response model this belief implies, *then* we can deduce that the Rasch model is the one and only reasonable model consistent with that belief and its practice. The very same statistic which, according to the Rasch model, leads to an unbiased, consistent and sufficient estimator of person ability is that same unweighted test score which nearly everyone already uses to make their measurements.

The Rasch model for what happens when person v responds to item i with binary response $X_{vi}=1$ (for a correct response) and $X_{vi}=0$ (otherwise) can be expressed as:

$$P\{x_{vi} \mid \beta_v, \delta_i\} = \frac{e^{x_{vi}(\beta_v-\delta_i)}}{1+e^{(\beta_v-\delta_i)}} \qquad (1)$$

where β_v = the ability of person v, and

δ_i = the difficulty of item i.

In order to emphasize that item difficulty in this model is a persisting property of the item and not of the person, we point out the difference between the colloquial use of "difficulty" as in the complaint "That item is easy for you but hard for me.", and the technical use in which the item has its own fixed difficulty and the complaint becomes "You are smart but I am dumb." one value of the response model is that it formulates the relationship between these two kinds of difficulty explicitly. Collo-

quial difficulty is person-bound. It is the probability of a person with a particular ability succeeding on the item. But technical difficulty is person-free. It is the fixed item parameter which joins with person ability in the response model to produce the probability of a correct response.

Practical discussions of how to estimate the parameters of this model are given in Rasch (1960), Wright and Panchapakesan (1969), Wright and Douglas (1975) and Wright and Mead (1975).

The Target

When an examiner plans a measurement there must be a target person or group of persons about whom he wants to know more than he already knows. If he cares about the quality of his proposed measurements then he will want to choose or construct his measuring instrument with the specifics of his target in mind. In order to do this systematically he must begin by setting out as clearly as he can what he expects of his target. Where does he suppose it is located on the latent variable? How uncertain is he of that approximate location? What is the lowest ability he imagines the target could have? What is the highest? How are other possible values distributed in between?

Sometimes an examiner has explicit prior knowledge about his target. He or someone else has measured it before and so he can suggest its probable location and dispersion directly in terms of these prior measures on the latent variable and their standard errors. Sometimes an examiner has sample items calibrated along the latent variable, some of which he or his client believe are probably just right for the target, some of which are nearly too hard and some of which are nearly too easy. Then he can take from the difficulties of these sample items rough indications of the probable center and boundaries of his target.

One way or another the examiner must assemble and clarify his suppositions about his target as well as he can so that he can derive from them the test design which has the best chance of most increasing his knowledge. If he knew enough about his target, he would not measure it. But no matter how little he knows, he always has some idea of where his target is. Being as clear as possible about that prior knowledge is essential for the design of the best possible test.

A target specification is a statement about where on the latent variable we suppose the target to be. We express our best guess by specifying the target's supposed center, its supposed dispersion and its supposed shape or distribution. If we let

M = our best guess as to target location,

S = our best guess as to target dispersion,

D = our best guess as to target distribution,

then we can describe a target G by the expression G(M,S,D) and we can summarize our prior knowledge and hence our measurement requirements for any target we wish to measure by guessing as well as we can values for the three target parameters, M, S, and D.

A picture of a target is given in the upper half of Figure 1. Guessing the supposed location M of a target seems straightforward. But guessing the dispersion S and the distribution D forces us to think through the difference between a target which is a single person and one which is a group. For the single person, S can describe the extent of our uncertainty about where that person is located. The larger our uncertainty, the larger S.

If we can specify boundaries within which we feel fairly sure the person will be found we can set S so that $M \pm kS$ defines these boundaries (where K=2 or 3). Then even if we have no clear idea at all about the distribution D of our uncertainty we can nevertheless expect (thanks to Tchebycheff's inequality) that at least $(1-1/k^2)$ of the possible measures will fall within M + kS.

If we go further and expect that the measures we think possible for the person will pile up near M then we may even be willing to take a normal distribution as a useful way to describe the shape of our uncertainty. In that case we can expect .95 of the possible measures to fall with $M \pm 2S$ and virtually all of them to fall within M + 3S.

We will refer to these two target distributions as the Tchebycheff interval and the normal. We might consider other target distributions. But these two cover an examiner's state of mind with respect to the shape of his target rather well. If he feels unhappy about thinking of his target as approximately normal then it is unlikely that he will have any definite alternative clearly in mind. Thus the most likely

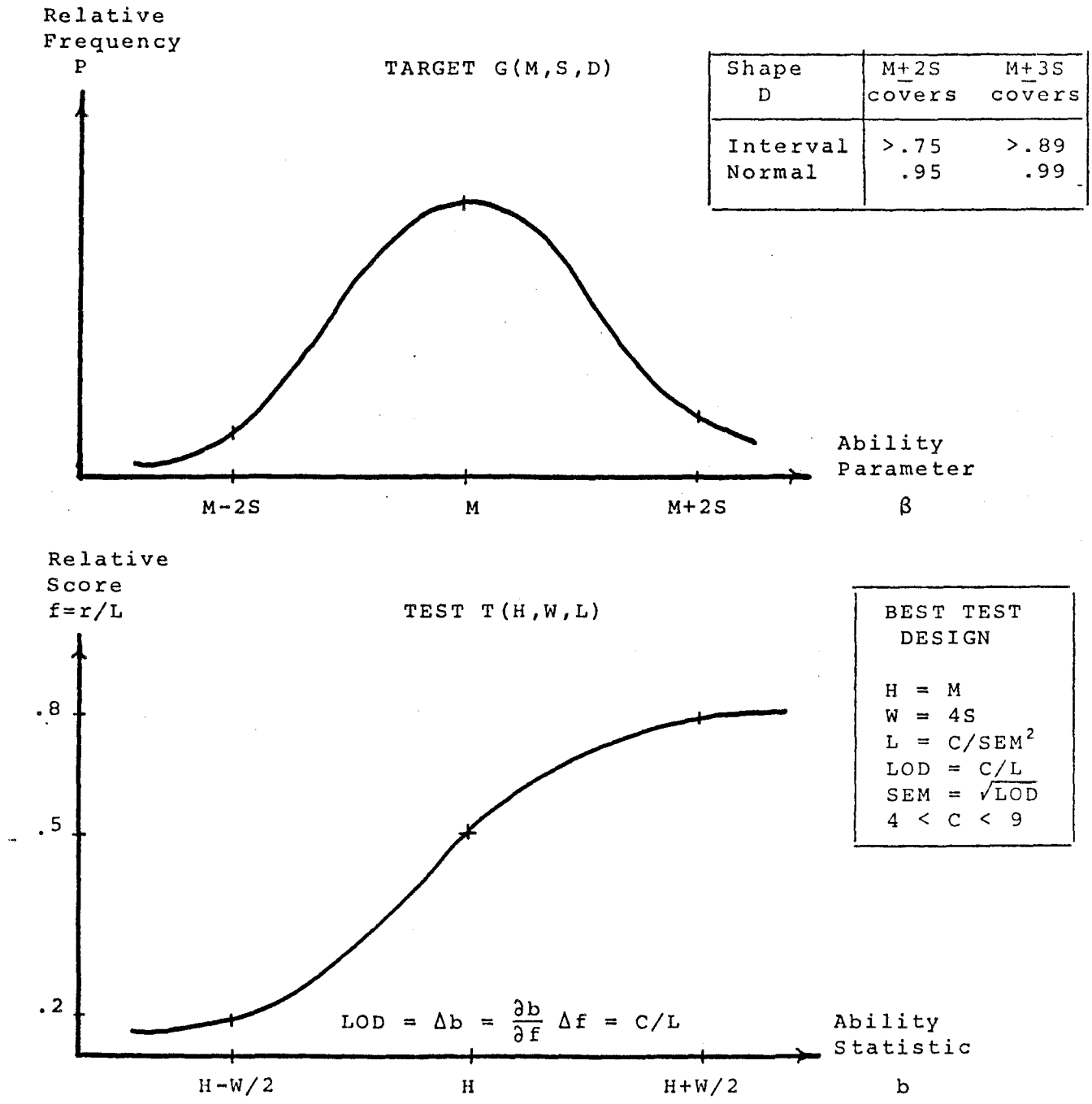

FIG. 1

The Distribution of a Target and the Operation of a Test

alternative to a normal target is one of unknown distribution, captured by a Tchebycheff interval. It is this simplification of possible target shapes to just two reasonable alternativeswhich makes a unique solution to the problem of best test design possible and practical.

If the target is a group rather than an individual then we may take S and D to be our best guess as to the standard deviation and distribution of that group. If we think the group has a more or less normal distribution then we will take that as our best guess for D. Otherwise we can fall back on the Tchebycheff interval.

Finally the examiner must be explicit about how precise he wants his measurement to be. This is his motive for measuring. It is just because his present knowledge is too approximate to suit him that he wants to know more precisely where his target is and, if it is a group rather than an individual, more precisely about its dispersion. But whether the target is an individual or a group our decision about the standard error of measurement SEM will be made in terms of individuals, for that is what we measure.

In the case of a one person target we want SEM to be enough smaller than S to reward our measurement efforts with a useful increase in the precision of our knowledge about where the target person is located. In the case of a group target we want to achieve an improved estimate not only of M, the center of the group, but also of S, its dispersion. The observable variance of measures over the group estimates not only the underlying variance in ability S2 but also the measurement error variance SEM2. Our ability to see the dispersion of our target against the background error of measurement depends on our ability to distinguish between these two components of variance. Since they enter into the observable variance of estimated measures equally, the smaller SEM is with respect to S, the more clearly can we identify and estimate S^2, the component due to target dispersion. Thus for all targets we seek SEM «S.

The Test

A test is a set of suitable items chosen to go together to form a measuring instrument. The complete specification of a test is the set of all parameters which characterize these items. But when we examine a picture of how a test works to transform observed scores into estimated measures, we see that the operating curve is rather simple and lends itself to specification through

just a few test parameters. If we use the simplicity and clarity of the Rasch model as our guide, we can deduce that the only parameters which influence the operating characteristics of a test are the difficulties of its items. When we impose a reasonable fixed distribution on these difficulties, then no matter how many items we use, we can reduce the number of statistics sufficient to characterize the test to three.

In the lower half of Figure 1 we can see from the shape of the test operating curve that its outstanding features are its position along the latent variable, which we will call test **height**, and the range of abilities over which the test can measure more or less accurately, a characteristic caused primarily by the dispersion of the item difficulties which we will call test **width**.

But height and width do not complete the characterization of a test. When we look more closely to see how precisely the test can measure on the latent variable we discover a discontinuity in what we can observe and hence measure. Both its least observable difference LOD and its least believable difference SEM are strongly related to the number of items in the test, that is its **length**. From this we see that any test design can be defined more or less completely just by specifying those three characteristics; height, width and length. If we let

H = the height of the test on the latent variable (i.e., the average difficulty of its designed items),

W = the width of the test in item difficulties (i.e., the range of its designed item difficulties), and

L = the length of the test in number of items,

then we can describe a test design T by the convenient expression T(H,W,L).

In the practical application of best test design we will have to approximate our best design T for a target G from a finite pool of existing items. In order to discriminate in our thinking between the test design T(F1,W,L) and its approximate realization in practice, we will describe an actual test as t(h,w,L), where

h = the average difficulty of its actual items, and

w = an estimate of their difficulty range.

PRINCIPLES OF BEST TEST DESIGN

What is a best test? [4] One which measures best in the region within which measurements are expected to occur. Measuring best means most precisely. So a best test design T(H,W,L) is one with the smallest error of measurement SEM over the target G(M,S,D) for given length L(or, what is equivalent, with smallest L given SEM). But "over the target" implies the minimization of a distribution of possible SEM's. Thus a position with respect to the most likely target distribution must be taken before the minimization of SEM can proceed.

We have brought the profusion of possible target shapes under control by focusing our practice on their two extremes, interval and normal. How shall minimization be specified in each case? For a normal target it seems reasonable to maximize average precision, that is to minimize average SEM, over the whole target. To decide what to do for an interval target we need to know how the SEM varies over possible test scores. When we derive an exact form for the precision of measurement, we see that for ordinary tests with less than three log odds units between adjacent items, precision is a maximum for measurements made at the center of the test and decreases as test and target are increasingly off-center witn respect to one another. For tests centered on their targets this means that, maximizing precision at the boundaries of an interval target is the only way to maximize precision over the target interval. So for interval targets we will maximize precision at the target boundaries.

When we derive the SEM from our response model we will discover that it is an inverse function of the information about ability supplied by each item response averaged over the test. Since the most informative items are those nearest the ability being measured and the least informative those farthest away, the precision over the target will depend not only on the distribution

[4] Attempts to answer this question have been made by Birnbaum (1968, p. 465-71) and although many of our ideas are consistent with his efforts we believe that we have taken their implications to a logical and practical conclusion.

of the target but also on the shape of the test. Thus the question of what is a best test also depends on our taking a position with respect to the best distribution of test item difficulties.

What are the reasonable possibilities? If we want to measure a normal target,then a test made up of normally distributed item difficulties might produce the best maximization of precision over the target. This is the conclusion implied in Birnbaum's analysis of information maximization (Birnbaum, 1968, p. 467).

However, normal tests are clumsy to compose. Normal order statistics can be used to define a unique set of item difficulties but this is tedious. More problematic is the odd conception of measuring implied by an instrument composed of normally distributed measuring elements. A normal test would be like a yardstick with rulings bunched in the middle and spread at the ends. Measuring lengths with such an irregularly ruled yardstick would be awkward. In the long run, even for normal targets, our interest becomes spread out evenly over all the lengths which might be measured. Equally spaced rulings are the test shape which serves that interest best. That is the way we construct yardsticks. The test design corresponding to an evenly ruled yardstick is the uniform test in which items are evenly spaced from easiest to hardest (Birnbaum, 1968, p. 466).

Two target distributions, normal and interval, and two test shapes, normal and uniform, produce four possible combinations of target and test. We investigatedall four.

Now we must turn to our response model and derive from it the formulation of SEM so that we can become explicit about which designs maximize precision by minimizing SEM. We want to know what aspects of test design T(H,W,L) and test shape influence SEM and how we can vary these characteristics in response to a target specification G(M,S,D) in order to minimize SEM over that target.

The response model specifies

$$P_{fi} = \frac{e^{b_f - d_i}}{1 + e^{b_f - d_i}} \qquad (2)$$

where P_{fi} = the probability of a correct response at f and i,

b_f = the ability estimates at relative score $f = r/L$,

d_i = the calibrated difficulty of item i.

According to maximum likelihood estimation (Birnbaum, 1968, p. 455-69; Wright and Panchapakesan, 1969, p. 41-44) b_f is estimated from a test of length L and items $\{d_i\}$, i=1,L through the equations

$$f = \sum_{i=1}^{L} P_{fi}/L \quad \text{for} \quad f = 1/L, (L-1)/L \qquad (3)$$

with asymptotic variance

$$S^2_{b_f} = \frac{1}{\sum_{i=1}^{L} P_{fi}(1-P_{fi})} = SEM^2_f , \qquad (4)$$

which is the square of the standard error of measurement at relative score f.

In (3) we see that SEM depends on the sum of $P_{fi}(1-P_{fi})$ over i. This expression is a function of b_f and all the d_i. However fluctuations in P(1-P) are rather mild for P between .2 and .8. To expedite insight into the make-up of SEM, we will reformulate it so that the average value of $P_{fi}(1-P_{fi})$ over i is one component and test length is the other:

$$SEM_f = \sqrt{\frac{L}{\sum_{i=1}^{L} P_{fi}(1-P_{fi})}} \sqrt{\frac{1}{L}} \qquad (5)$$

Thus we factor length out of SEM in order to find a length-free error coefficient.

Resuming our study of the operating curve of a test in Figure 1 we see that the formulation for the least observable difference in ability LOD is $\Delta b = \frac{\partial b}{\partial f}\Delta f$.

Since the least observable increment in relative score Δf is $1/L$, all we need to complete the formulation of LOD is the derivative of b with respect to f, which from (2) and (3) above is

$$\frac{\partial b}{\partial f} = \frac{L}{\sum_{i=1}^{L} P_{fi}(1-P_{fi})} = C_f \qquad (6)$$

But this is the inverse of the average value of $P_{fi}(1-P_{fi})$ over i which we isolated in (5). We will identify this as the error coefficient C_f and note that it is not only a function of f but also of the distribution of item difficulties.

Then $$LOD_f = C_f/L \qquad (7)$$

and $$SEM_f = \sqrt{C_f/L} \qquad (8)$$

With SEM in this form we note that as far as test shape is concerned it is C_f which requires minimization. This will be true whether we use C_{min} to minimize SEM given L or to minimize L given SEM.

The Error Coefficient

What is the nature of this error coefficient C_f? The expression $P_{fi}(1-P_{fi})$ is the information on b_f contained in a response to item i (Birnbaum, 1968, p. 460-68). So its average over i is the average infor-

mation on b_f per item in the test. Thus C_f is the inverse of this average test information. The greater the information obtained the smaller C_f and hence the smaller SEM_f and so the greater the precision.

What values can we expect of C_f? We can approach this question in two ways: in terms of the influence of reasonable values of (b_f-d_i) on P_{fi} or, if we are willing to focus attention on uniform tests, in terms of test width W and the boundary P's, P_{f1} for $i=1$, the easiest item, and P_{fL} for $i=L$, the hardest item.

Beginning with reasonable values of (b_f-d_i), we see that when $b_f=d_i$ and their difference is zero then $P_{fi} = 1/2$ and $P_{fi}(1-P_{fi}) = 1/4$, but when $b_f-d_i=-2$ then $P_{fi} = 1/8$ and $P_{fi}(1-P_{fi}) = 1/9$. Since an average value can never be greater than the maximum nor less than the minimum value, we can use these figures as bounds for C_f.

When $$-2 < b_f - d_i < 2,$$

$$1/8 < P_{fi} < 7/8$$

and $$4 < C_f < 9 \qquad (9)$$

Turning to the bounds we can derive for C_f from test width W and the boundary probabilities P_{f1} and P_{fL} of a uniform test, we can use an expression (Birnbaum, 1968, p. 466)

$$C_{fW} = W/(P_{f1}-P_{fL}) \qquad (10)$$

where W = the item difficulty width of a uniform test,

P_{f1} = the probability of a correct response by b_f on the easiest item on the test, and

P_{fL} = the probability of a correct response by b_f on the hardest item on the test.

When b_f is contained within the difficulty boundaries of

the test, that is $d_1 < b_f < d_L$ and $W>4$ then $1/2 < P_{f1}-P_{fL} < 1$ and C_{fW} must fall between W and 2W, that is

$$W < C_{fW} < 2W \quad \text{for} \quad W>4 \ , \ d_1<b_f<d_L \tag{11}$$

It follows from considerations (9) and (11) that $SEM = \sqrt{C/L}$ is bounded by $2/\sqrt{L} < SEM < 3/\sqrt{L}$ for any test when $-2 < (b_f-d_i) < 2$, and by $\sqrt{W/L} < SEM < \sqrt{2W/L}$ for uniform tests when $W>4$ and $d_1 < b_f < d_L$.

Simplest Rules

Best test design depends on relating the characteristics of test design T(H,W,L) to the characteristics of target G(M,S,D) so that the SEM is minimized in the region of the latent variable where the measurements are expected to take place. The relationship between test and target visible in Figure 1 makes the general principles of best test design obvious. To match test to target we aim the height of the test at the center of the target, spread the width of the test to cover the dispersion of the target and lengthen the test until it provides the precision we require.

In simplest terms the rules for best test design can be specified as:

(i) Center T on G by $H = M$

(ii) Spread T over G by $W = 4S$

(iii) Lengthen T to reach SEM by $L = C/SEM^2$

where C is the error coefficient with values between 4 and 2W, producing a test for which the operating characteristics will be

a least observable difference $LOD < 2W/L$ and

a standard error of measurement $SEM < \sqrt{2W/L}$.

MATHEMATICS FOR ERROR MINIMIZATION

Even though we have been explicit about the nature of the target G and the test design T, the direct minimization of C is impeded because it is an implicit function of its unknowns. We can see this by deriving the asymptotic variance of the ability estimates and hence the C coefficients.

When the item parameters are considered known, the likelihood to be maximized for the estimation of ability parameter β_v is the product over items of subject v's probabilities of success on each item:

$$\mathcal{L}_v = \prod_{i=1}^{L} \frac{e^{x_{vi}(\beta_v-\delta_i)}}{1+e^{\beta_v-\delta_i}} = \frac{e^{\sum_{i=1}^{L} x_{vi}(\beta_v-\delta_i)}}{\prod_{i=1}^{L}[1+e^{\beta_v-\delta_i}]} = \frac{e^{r\beta_v-\sum_{i=1}^{L} x_{vi}\delta_i}}{\prod_{i=1}^{L}[1+e^{\beta_v-\delta_i}]} \qquad (12)$$

The expression on the right demonstrates that the test score r is sufficient for estimating the parameter β_v since the likelihood may be factored into two components, only one of which, $\frac{e^{r\beta_v}}{\prod_{i=1}^{L}[1+e^{\beta_v-\delta_i}]}$, is dependent on β_v, and this component depends on the observations only through r (Birnbaum, 1968, p. 425-29). Taking logarithms we have:

$$L_v = \log \mathcal{L}_v = r\beta_v - \sum_{i=1}^{L} x_{vi}\delta_i - \sum_{i=1}^{L} \log [1+e^{\beta_v-\delta_i}] \qquad (13)$$

Differentiating this log likelihood with respect to β_v gives

$$\frac{\partial L_v}{\partial \beta_v} = r - \sum_{i=1}^{L} \left(\frac{e^{\beta_v-\delta_i}}{1+e^{\beta_v-\delta_i}} \right) = r - \sum_{i=1}^{L} P_{vi} \qquad (14)$$

with second derivative

$$\frac{\partial^2 L_v}{\partial \beta_v^2} = - \sum_{i=1}^{L} \frac{e^{a_v - a}}{s_v (1+e^{v-aa.})^2} < 0 \qquad (15)$$

Since r, or f= r/L, is sufficient for estimating s_v we may re-write (14) and (15) in terms of the L-1 estimates b_f:

$$\frac{aL_f}{ab_f} = Lf - \sum_{i=1}^{L} \frac{e^{b_f - 6_i}}{1+e^{b_f - 6_i}} = Lf - \sum_{i=1}^{L} P_{fi} \qquad (16)$$

$$\frac{a^2 L_f}{ab_f^2} = - \sum_{i=1}^{L} \frac{e^{b_f - a_i}}{b_f - 8i^2} = - \sum_{i=1}^{L} P_{fl}(1-P_{fi}) \qquad (17)$$

where $P_{fi} \quad \frac{e^{b_f - d_i}}{1+e^{b_f - \delta_i}}$

When we set (16) equal to zero we obtain an implicit solution for b_f:

$$f = \sum_{i=1}^{L} P_{fi} / L \qquad (18)$$

In maximum likelihood estimation the negative inverse of the second derivative is the asymptotic error variance of the estimate:

$$SEM_f^2 = \frac{}{\sum_{i=1}^{L} P_{fi}(1-P_{fi})} C_f/L \qquad (19)$$

Hence

$$C_f \quad L(SEMf) = \frac{L}{\sum_{i=1}^{L} P_{fi}(1-P_{fi})} \qquad (20)$$

Iterative techniques (such as Newton-Raphson) may be employed to obtain solutions for b_f and C_f (Birnbaum, 1968, p.455-59). However, if we are willing to specify the distribution of $\{\delta_i\}$ and to approximate the discrete set of δ_i with a continuum, we can derive explicit solutions for both b_f and C_f which are excellent and useful approximations to the maximum likelihood estimates.

Uniform Items

A continuous uniform density function for δ with mean $\delta.$ and range W is given by

$$g(\delta) = \begin{cases} 1/W & \delta.-W/2 < \delta < \delta.+W/2 \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

Our response model, written as a function of the variable δ, is

$$P_f(\delta) = \frac{e^{b_f-\delta}}{1+e^{b_f-\delta}} \tag{22}$$

By equating the relative score f with its expected value $P_f(\delta)$ we obtain

$$f = \mathcal{E}_\delta [P_f(\delta)] = \int_{\delta.-W/2}^{\delta.+W/2} P_f(\delta) g(\delta) d\delta = \int_{\delta.-W/2}^{\delta.+W/2} \frac{e^{b_f-\delta}}{1+e^{b_f-\delta}} \frac{1}{W} d\delta \tag{23}$$

which is of the standard form

$$\int [h'(x)/h(x)]\, dx$$

with solution

$$f = \left(\frac{1}{W}\right) \log \left(\frac{1+e^{b_f-\delta.+W/2}}{1+e^{b_f-\delta.-W/2}} \right) \tag{24}$$

Solving for e^{fW} leads to

$$e^{b_f - \delta_.} = \frac{e^{W(f-1/2)}[1-e^{-fW}]}{[1-e^{-(1-f)W}]} \tag{25}$$

from which, taking logarithms, an explicit approximation for b_f results:

$$b_f = \delta_. + W(f-1/2) + \log\left[\frac{1-e^{-fW}}{1-e^{-(1-f)W}}\right] \tag{26}$$

where $b_f = \delta_. + \log \frac{f}{1-f}$ when $W = 0$

and $b_{.5} = \delta_.$ when $f = .5$

Turning to a similar approximation for the error coefficient C_f we approximate the average over L discrete δ_i by the integral over $g(\delta)$ to arrive at

$$\frac{1}{C_f} = \frac{\sum_{i=1}^{L} P_{fi}(1-P_{fi})}{L} = \int_{\delta_.-W/2}^{\delta_.+W/2} P_f(\delta)[1-P_f(\delta)]g(\delta)\,d\delta \tag{27}$$

This integral can be evaluated by noting that $P_f(\delta)[1-P_f(\delta)]$ is the logistic density function, $\psi(\delta)$, whose integral is the logistic cumulative function $\Psi(\delta)$. Whence

$$C_{fW} = \frac{W}{\{\Psi(b_f-\delta_.+W/2) - \Psi(b_f-\delta_.-W/2)\}} \tag{28}$$

which is approximately the ratio of test width W to the difference between the probabilities of success on the easiest item at δ_1 and hardest item at δ_L:

$$C_{fW} = \frac{W}{P_{f1} - P_{fL}} \tag{29}$$

This error coefficient may be expressed in terms of b_f as

$$C_{b_f W} = \frac{W[1+e^{b_f-\delta.+W/2}][1+e^{b_f-\delta.-W/2}]}{[e^{b_f-\delta.+W/2} - e^{b_f-\delta.-W/2}]} \tag{30}$$

and by replacing b_f by f and W as given by (26) as

$$C_{fW} = \frac{W\,[1-e^{-W}]}{[1-e^{-fW}][1-e^{-(1-f)W}]} \tag{31}$$

where $C_{f,0} = \frac{1}{f(1-f)}$ when $W = 0$

and $C_{.5,0} = 4$ when $W = 0$, $f = .5$.

Our final approximation for uniform items is for the average value of C_{bW} over the latent variable when the target distribution of b is normal:

$$D(b) \sim N[M,S^2].$$

We proceed by taking the expected value of C_{bW} over the range of the normal distribution in the metric of b.

$$\bar{C}_W = \mathcal{E}_b[C_{bW}] = \int_{-\infty}^{\infty} C_{bW}\, dD(b)$$

$$= \frac{W}{\sqrt{2\pi}S} \int_{-\infty}^{\infty} \frac{[1+e^{b-\delta.+W/2}][1+e^{b-\delta.-W/2}]e^{-[b-M]^2/2S^2}}{[e^{b-\delta.+W/2} - e^{b-\delta.-W/2}]}\, db \tag{32}$$

Although the algebra is tedious there is no obstruction to evaluating this integral:

$$\bar{C}_W = \frac{W}{[e^{W/2}-e^{-W/2}]}\,[e^{W/2}+e^{-W/2}]+e^{M+S^2/2-\delta.}+e^{-M+S^2/2+\delta.} \qquad (33)$$

Where $C_0 = 4$ when $W = 0$, $M = \delta.$ and $S = 0$.

Normal Items

Our approximation for a test with normally distributed items is based on the similarity in shape of the normal and logistic distributions. If a scaling factor of 1.7 is used for the logistic, the difference between their cdf's is

$$|\Phi(x)-\Psi(1.7x)| < .01 \quad \text{for all} \quad x. \qquad (34)$$

In terms of the corresponding pdf's, we obtain by differentiation,

$$|\phi(x/1.7)/1.7-\psi(x)| < .015 \quad \text{for all } x^{(5)}. \qquad (35)$$

Our procedures for obtaining the approximations for normal items are identical to those for uniform items except that the above inequalities connecting the normal and logistic distributions are required to solve the integrals.

For our first approximation, with

$$g(\delta) \sim N[\delta.,Z^2]$$

we set

$$f = \frac{1}{\sqrt{2\pi}z}\int_{-\infty}^{\infty} \Psi[b-\delta]\; g(\delta)\; d\delta \qquad (36)$$

---

[5] Results (34) and (35) were verified numerically by comparing the respective functions for values of x ranging from -4.5 to 4.5. In the case of (34) the difference at $x = 0$ is zero and the maximum difference of .01 occurs at ± 2.0. In the case of (35) the difference at $x = 0$ is .015, the maximum, whereas the minimum difference of .003 occurs at ± 3.5.

and after replacing the cdf of the logistic by its normal approximation we find that

$$f = \Phi\left[\frac{b - \delta.}{\sqrt{2.9 + Z^2}}\right] \tag{37}$$

We use inequality (34) again to obtain

$$b_f = \delta. + \sqrt{1 + \left(\frac{Z}{1.7}\right)^2}\ \text{log}\ \left(\frac{f}{1-f}\right) \tag{38}$$

The expression under the radical acts to expand the initial estimates ($\log\left[\frac{f}{1-f}\right]$). The larger Z, the greater this expansion. Setting $Z = 0$ and then $f = .5$ produce

$b_f = \delta. + \text{log}\ \left(\frac{f}{1-f}\right)$ and $b_{.5} = \delta.$ as in (25).

Similarly for C_{fZ}, inequality (34) produces:

$$C_{fZ} = \sqrt{2\pi}\ \sqrt{2.9+Z^2}\ e^{[\text{log}(\frac{f}{1-f})]^2/5.8} \tag{39}$$

or, in the metric of b_f,

$$C_{b_f Z} = \sqrt{2\pi}\ \sqrt{2.9+Z^2}\ e^{[b-\delta.]^2/2[2.9+Z^2]} \tag{40}$$

where $C_{b,0} = 1.7\ \sqrt{2\pi} = 4.26 \neq 4$, when $b = \delta.$ and $Z = 0$.

The bias in this approximation is a result of the application of the inequalities (34) and (35). In terms of standard errors of measurement, $SEM = \sqrt{C/L}$, it is extremely small,

since $\sqrt{4.26} = 2.06 \simeq 2.0$.[6]

Finally for an average C_{bZ} over a normal target we find

$$\overline{C}_Z = \frac{\sqrt{2\pi}\ (2.9+Z^2)}{\sqrt{2.9 + Z^2 - S^2}}\ e^{\frac{[M-\delta.]^2}{2[2.9+Z^2-S^2]}} \qquad (41)$$

subject to the restriction that $2.9 + Z^2 - S^2 > 0$. Setting $M = \delta.$, $S = 0$ and $Z = 0$ shows the same bias in $\overline{C}_Z$ as (39).

we have investigated the accuracy of each of these six approximations (26, 30, 33, 38, 40, 41) for a wide variety of test designs. The uniform approximations are exceptionally accurate, resting as they do on the replacements of a discrete uniform by a continuous uniform distribution. The normal approximations are less accurate but produce ability estimates which are within .1 log odds units of the maximum likelihood estimates for a range of cases wider than those likely to be encountered in practice (Douglas, 1975)

MINIMIZATION OF C_f AND $\overline{C}$ FOR NORMAL AND UNIFORM ITEMS

We will deal with normal items first because those approximations for C add $\overline{C}$ lend themselves to differentiation and thus enable us to minimize Z algebraically. To minimize C_{fZ} at the interval target boundary of $M\pm 2S$ we place

$$(b-\delta.)^2 = 4S^2 \quad \text{in} \quad (40) \quad \text{and}$$

$$C_Z = \sqrt{2\pi}\ \sqrt{2.9+Z^2}\ \ e^{2S^2/(2.9+Z^2)} \qquad (42)$$

---

[6] The reason that the approximation for b_f in the special case is exact and the approximation for C_{fZ} is biased is because only the first inequality (34) is employed in obtaining b_f whereas both are required for C_{fZ}. The difference in (34) at $X=0$ is zero but the difference in (35) at zero is the maximum .015.

By setting the first derivative of C_Z with respect to Z equal to zero, and noting that the second derivative is negative when $4S^2 - 2.9 < 0$ we see that $Z = 0$

gives $C_{Z,min} = \sqrt{2\pi}\ (1.7)\ e^{2S^2/2.9} = 4.26\ \ e^{2S^2/2.9}$

when $S < \frac{1.7}{2} = .85$ and $Z = \sqrt{4S^2 - 2.9}$

gives $C_{Z,min} = 2\sqrt{2\pi}\ e^{1/2}\ S = 8.27S$ when $S > \frac{1.7}{2} = .85.$

For the average $\bar{C}_Z$ we find by a similar process that

$Z = 0$ gives $\bar{C}_{Z,min} = \frac{1.7\sqrt{2\pi}}{\sqrt{1-(\frac{S}{1.7})^2}} = \frac{4.26}{\sqrt{1-(\frac{S}{1.7})^2}}$ when $S < \frac{1.7}{\sqrt{2}} = 1.2$

and $Z=\sqrt{2S^2-2.9}$ gives $\bar{C}_{Z,min} = 2\sqrt{2\pi}\ S=5S$ when $S > \frac{1.7}{\sqrt{2}} = 1.2.$

These explicit minima allow for easy evaluation of C and $\bar{C}$ over various values of target standard deviation S. (Douglas, 1975).

In the case of uniform items we were forced to determine the values of W which minimize C_{fW} and $\bar{C}_W$ numerically because we could not solve expressions (30) and (33) after differentiation. Therefore we computed C_{fW} and $\bar{C}_W$ for a wide variety of combinations of f and W and obtained the minimizing W's and their related C_{min} and $\bar{C}_{min}$ by inspection. (Douglas, 1975).

Table 1 lists the values of the minimum error coefficients C_{min} and $\bar{C}_{min}$ for both uniform and normal items and the corresponding W in the case of uniform items. The uniform approximation is exact for all practical purposes. But the normal approximation contains a bias due to the interchange of logistic and normal pdf's during the development of the approximation. As a result the C_{min} and $\bar{C}_{min}$ values for the normal approximation are too large above the dotted lines

TABLE 1

Minimum Error Coefficients C_{min} and $\bar{C}_{min}$ for Interval and Normal Targets under Uniform and Normal Tests

Target Std. Dev.	Interval Target			Normal Target		
	Normal Test	Uniform Test		Normal Test	Uniform Test	
S	C_{min}	C_{min}	W	$\bar{C}_{min}$	$\bar{C}_{min}$	W
0.00	4.26	4.00	0	4.26	4.00	0
0.25	4.45	4.25	0	4.31	4.06	0
0.50	5.07	5.09	0	4.46	4.27	0
0.75	6.29*	6.62	2.60	4.75	4.65	0
1.00	8.27	8.18	5.00	5.27	5.30	0
1.25	10.33	9.60	6.75	6.27	6.34	1.88
1.50	12.40	10.94	8.30	7.52	7.54	4.12
1.75	14.47	12.22	9.75	8.77	8.80	5.88
2.00	16.53	13.46	11.00	10.03	10.11	7.38

*The approximation value is 6.29 but the true value for a normal test at S = 0.75 is 6.70 which is greater than 6.62.

MEASUREMENC PROCEDURES

When we compare the normal and uniform tests on interval targets there is no doubt that the uniform test is more precise. This is especially true when S exceeds 1.

When we compare the normal and uniform tests on normal targets they differ so little as to seem equivalent in their measuring precision. Since uniform tests are better for interval targets and equally good for normal targets there is no motivation for considering normal tests any further. Therefore we will focus the rest of our best test design strategy on uniform tests. Table 2 gives optimal uniform test widths for a variety of normal and interval targets.

For best test design on either interval or normal targets we select a set of equivalent items (where W = 0) or a set of uniform items with W as indicated by Table 2. For example, if the target is thought to be approximately normal with presumed.standard deviation S = 1.5, the optimum test width W is 4. Note that for any value of S a smaller W is indicated when a normal target shape is expected.

Table 2 also shows the efficiency of a simple rule for relating test width W to target dispersion S. The rule W = 4S comes close to the optimal W for narrow interval targets and for wide normal targets. When we are vague about where our target is we are also vague about its boundaries. That is just the situation where we would be willing to use a normal distribution as the shape of our target uncertainty. When our target is narrow however, that is the time when we are rather sure of our target boundaries but, perhaps, not so willing to specify our expectations as to its precise distribution within these narrow boundaries. To the extent that interval shapes are natural for narrow targets while normal shapes are inevitable for wide targets, W = 4S is a useful simple rule.

The efficiency of this simple rule for normal and interval targets is given in the final columns of Table 2. There we see that it is hardly ever less than 90 per cent efficient; if we cross over from an interval target to a normal target as our expected target dispersions exceeds 1.4, then the efficiency is never less than 95 per cent.

TABLE 2

Values of W for Best Uniform
Tests on Normal and Interval Targets

Target Std. Dev.	Normal Target	Interval Target	Simple Rule*	Efficiency**	
S	W over $N(M,S^2)$	W at $(M \pm 2S)$	W=4S	Normal	Interval
.5	0	0	2	94	97
.6	0	0	2		
.7	0	2	3	90	100
.8	0	3	3		
.9	0	4	4		
1.0	0	5	4	89	98
1.1	0	6	4		
1.2	1	6	5	92	96
1.3	2	7	5		
1.4	3	7	6		
1.5	4	8	6	96	91
1.6	5	9	6		
1.8	6	10	7	98	87
2.0	8	11	8	99	84

*This Simple Rule is conservative for narrow targets and more practical since available items are bound to spread some. It is also close to the normal target optimum for wide targets, which is reasonable in the face of substantial target uncertainty.

**

Efficiency $\equiv C_W / C_{4S} = L_W / L_{4S}$

where C_W = minimum error coefficient for optimal W.
C_{4S} = error coefficient for W = 4S.
L_W = length of optimal test of width W.
L_{4S} = length of equally precise test of width 4S.

Our investigations have shown that given a target M, S and D there exists an optimal test design H and W from which may be generated a unique set of L uniformly distributed item parameters $\{\delta_i\}$. However, this design is an idealization and cannot be perfected in practice. Real item pools are finite and each item difficulty is only an estimate of its corresponding parameter and hence inevitably subject to calibration error. No examiner will ever be able to select the exact items stipulated by his best test design $\{\delta_i\}$. Instead he must attempt to select from among the items he has available, a real set of $\{d_i\}$ which come as close as possible to his ideal design $\{\delta_i\}$.

Thus parallel to the design specification T(H,W,L) we must write the test description t(h,w,L) characterizing the actual test $\{d_i\}$ which can be constructed in practice. This raises the problem of estimating h and w from the set of items difficulties $\{d_i\}$. We will take their observed mean d. as h. As for their width, W, we investigated a number of alternatives: the range $w_s = \sqrt{12S_d^2}$, of a uniform distribution in which S_d^2 was the variance of d_i, the observed "first" range $w_1 = (d_L - d_1)(\frac{L}{L-1})$ and the observed "second" range $w_2 = ((d_L + d_{L-1} - d_2 - d_1)/2)(\frac{L}{L-2})$. There were few differences among the results we obtained from these alternatives for w. Our preference for w_2 is based on the observation that it is a slightly more stable estimate of W than w_1 and even w_s for short tests.

The approximations derived in order to minimize C may be used to set up two tables which allow an examiner to read off the ability estimate and its standard error for any relative score f which is observed on any test of width w and length L.

Table 3 gives us the position of the estimated measure b_f, relative to test height h, for any observed relative score f on a test of width w. Since this test t(h,w,L) will not in general be centered at zero, we must then add the test height h estimated by d. to the tabled values in order to arrive at the final estimated measure b_f. If we identify the entries in Table 3 as x_{fw} then the final ability estimate b_f is:

$$b_f = d. + x_{fw} \qquad \text{when} \quad f > .5$$

$$b_f = d. - x_{(1-f)w} \qquad \text{when} \quad f < .5.$$

For example, if the test characteristics are t(1,5,30) then a person who obtains a score of 12, or 40 per cent of the items correct, would have an ability estimate of

$$b_{.4} = 1.0 - x_{.6,5}$$

$$= 1.0 - 0.6$$

$$= 0.4.$$

Table 4 gives us the square root of the error coefficient, C_{fw} in preparation for obtaining an estimate of the standard error of measurement,

$$SEM = \sqrt{C_{fw}} \; / \; \sqrt{L}\,.$$

We are now in a position to give explicit, objective and systematic rules for the design and use of a best possible test.

COMPLETE RULES FOR BEST TEST DESIGN AND MEASUREMENT

For <u>design</u> T(H,W,L) on <u>target</u> G(M,S,D)

1. From our hypothesis about M we derive H = M.
2. From our hypothesis about S we derive an optimum W either by consulting Table 2, or by using the simple rule W = 4S.
3. From our requirements for measurement precision, namely the value of SEM we seek, we derive $L = C/SEM^2$ where C either comes from the

TABLE 3

Relative Ability Estimates x_{fw} in Log Odds for Uniform Tests

f \ w	0	1	2	3	4	5	6	7	8	9	10	11	1-f
.50	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	.50
.52	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	.48
.54	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.5	0.5	.46
.56	0.2	0.2	0.3	0.3	0.3	0.4	0.4	0.4	0.5	0.6	0.6	0.7	.44
.58	0.3	0.3	0.3	0.4	0.4	0.5	0.5	0.6	0.7	0.7	0.8	0.9	.42
.60	0.4	0.4	0.4	0.5	0.5	0.6	0.7	0.7	0.8	0.9	1.0	1.1	.40
.62	0.5	0.5	0.5	0.6	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	.38
.64	0.6	0.6	0.6	0.7	0.7	0.8	0.9	1.1	1.2	1.3	1.4	1.6	.36
.66	0.7	0.7	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.5	1.6	1.8	.34
.68	0.8	0.8	0.8	0.9	1.0	1.1	1.2	1.4	1.5	1.7	1.8	2.0	.32
.70	0.8	0.9	0.9	1.0	1.1	1.2	1.4	1.5	1.7	1.9	2.1	2.2	.30
.72	0.9	1.0	1.0	1.1	1.2	1.4	1.5	1.7	1.9	2.1	2.3	2.5	.28
.74	1.0	1.1	1.1	1.2	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.7	.26
.76	1.2	1.2	1.2	1.3	1.5	1.6	1.8	2.0	2.2	2.5	2.7	2.9	.24
.78	1.3	1.3	1.4	1.5	1.6	1.8	2.0	2.2	2.4	2.7	2.9	3.2	.22
.80	1.4	1.4	1.5	1.6	1.8	1.9	2.2	2.4	2.6	2.9	3.1	3.4	.20
.82	1.5	1.5	1.6	1.7	1.9	2.1	2.3	2.6	2.8	3.1	3.4	3.7	.18
.84	1.7	1.7	1.8	1.9	2.1	2.3	2.5	2.8	3.0	3.3	3.6	3.9	.16
.86	1.8	1.8	1.9	2.1	2.3	2.5	2.7	3.0	3.3	3.6	3.9	4.2	.14
.88	2.0	2.0	2.1	2.3	2.5	2.7	2.9	3.2	3.5	3.8	4.2	4.5	.12
.90	2.2	2.2	2.3	2.5	2.7	2.9	3.2	3.5	3.8	4.1	4.5	4.8	.10
.92	2.4	2.5	2.6	2.7	2.9	3.2	3.5	3.8	4.1	4.4	4.8	5.2	.08
.94	2.8	2.8	2.9	3.1	3.3	3.5	3.8	4.1	4.5	4.8	5.2	5.6	.06
.96	3.2	3.2	3.3	3.5	3.7	4.0	4.3	4.6	5.0	5.3	5.7	6.1	.04

If $f > .5$, $b_f = d. + x_{fw}$

$f < .5$, $b_f = d. - x_{(1-f)w}$

$$x_{fw} = w(f-.5) + \ln([1-\exp(-wf)]/[1-\exp(-w(1-f))])$$

TABLE 4

Error Coefficient $\sqrt{C_{fw}}$ for SEM = $\sqrt{C_{fw}}/\sqrt{L}$ in Log Odds

f \ w	0	1	2	3	4	5	6	7	8	9	10	11	1-f
.50	2.0	2.0	2.1	2.2	2.3	2.4	2.6	2.7	2.9	3.0	3.2	3.3	.50
.52	2.0	2.0	2.1	2.2	2.3	2.4	2.6	2.7	2.9	3.0	3.2	3.3	.48
.54	2.0	2.0	2.1	2.2	2.3	2.4	2.6	2.7	2.9	3.0	3.2	3.3	.46
.56	2.0	2.0	2.1	2.2	2.3	2.4	2.6	2.7	2.9	3.0	3.2	3.3	.44
.58	2.0	2.0	2.1	2.2	2.3	2.4	2.6	2.7	2.9	3.0	3.2	3.3	.42
.60	2.0	2.1	2.1	2.2	2.3	2.5	2.6	2.7	2.9	3.0	3.2	3.3	.40
.62	2.1	2.1	2.1	2.2	2.3	2.5	2.6	2.8	2.9	3.1	3.2	3.3	.38
.64	2.1	2.1	2.2	2.2	2.4	2.5	2.6	2.8	2.9	3.1	3.2	3.4	.36
.66	2.1	2.1	2.2	2.3	2.4	2.5	2.6	2.8	2.9	3.1	3.2	3.4	.34
.68	2.1	2.2	2.2	2.3	2.4	2.5	2.7	2.8	3.0	3.1	3.2	3.4	.32
.70	2.2	2.2	2.3	2.3	2.4	2.6	2.7	2.8	3.0	3.1	3.2	3.4	.30
.72	2.2	2.2	2.3	2.4	2.5	2.6	2.7	2.9	3.0	3.1	3.3	3.4	.28
.74	2.3	2.3	2.3	2.4	2.5	2.6	2.8	2.9	3.0	3.2	3.3	3.4	.26
.76	2.3	2.4	2.4	2.5	2.6	2.7	2.8	2.9	3.1	3.2	3.3	3.4	.24
.78	2.4	2.4	2.5	2.6	2.6	2.8	2.9	3.0	3.1	3.2	3.4	3.5	.22
.80	2.5	2.5	2.6	2.6	2.7	2.8	2.9	3.1	3.2	3.3	3.4	3.5	.20
.82	2.6	2.6	2.7	2.7	2.8	2.9	3.0	3.1	3.2	3.4	3.5	3.6	.18
.84	2.7	2.7	2.8	2.8	2.9	3.0	3.1	3.2	3.3	3.4	3.5	3.6	.16
.86	2.8	2.9	2.9	3.0	3.1	3.2	3.3	3.4	3.4	3.5	3.6	3.7	.14
.88	3.1	3.1	3.1	3.2	3.3	3.3	3.4	3.5	3.6	3.7	3.8	3.9	.12
.90	3.3	3.3	3.4	3.5	3.6	3.7	3.7	3.8	3.9	3.9	4.0	4.1	.10
.92	3.7	3.7	3.7	3.8	3.8	3.9	4.0	4.0	4.1	4.2	4.3	4.3	.08
.94	4.2	4.2	4.2	4.3	4.4	4.5	4.5	4.5	4.6	4.6	4.7	4.8	.06
.96	5.1	5.1	5.1	5.2	5.2	5.2	5.3	5.4	5.4	5.5	5.5	5.6	.04

Simple Rule: when $w < 4$, $2 < \sqrt{C} < 3$

$w > 4$, $\sqrt{w} < \sqrt{C} < \sqrt{2w}$

$$C_{fw} = [w(1-\exp(-w))]/[(1-\exp(-wf))(1-\exp(-w(1-f)))]$$

value minimized by W given in Table 4, or is approximated by the simple rule C = 8S.

4. From these H, W and L we generate the design set $\{\delta_i\}$ according to the formula:

$$\delta_i = H - \frac{W}{2L}[L-2i+1] \qquad i = 1,L$$

For test t(h,w,L) from design T(H,W,L)

5. Commencing with δ_i we select items d_i from the item pool such that they best approximate the set $\{\delta_i\}$, i.e., by minimizing the discrepancy $e_i = d_i - \delta_i$.

6. We calculate $h = d. = \sum_{i=1}^{L} d_i/L$ and

and $w = ((d_L + d_{L-1} - d_2 - d_1)/2)(\frac{L}{L-2})$.

7. We administer the set of $\{d_i\}$ as the test t(h,w,L), obtain b_f for f and w as given in Table 3 and SEM = $\sqrt{C_{fw}} / \sqrt{L}$ with C_{fw} for f and w as given in Table 4.

BEST TEST PERFORMANCE

Confidence in the use of these techniques depends on a knowledge of their functioning over a variety of typical testing situations. We investigated their performance with a simulation study designed to check on two major threats to the successful functioning of this testing strategy.

(i) By simulating a person at a fixed point on the target (say ~=0) and subjecting him to a series of tests with centers increasingly removed from his ability level we evaluated the effectiveness of this technique as tests are increasingly "off-center" with respect to their targets.

(ii) By introducing increasing magnitudes of random disturbance into the designed item calibrations $\{d_i\}$ we simulated and then evaluated the practica: situation where the actual test realized t(h,w,L) departs increasingly from its generating design T(H,W,L).

The simulation was set up according to the following steps:

1. Set person ability at $\beta = 0$.
2. Consider values of test height $H = 0,1,2$, test width $W = 2,5,8$ and test length $L = 10,30,50$.
3. Determine the item spacing parameter $V = W/L$ and set the disturbance standard deviation at $\sigma = V/2, 2V$ so that the magnitude of discrepancy is keyed to multiples of the spacing parameter.
4. Generate $\{\delta_i\}$ from H, W, and L and $\{d_i = \delta_i + e_i\}$ from $e_i \sim N[0,\sigma^2]$.
5. Calculate h and w from $\{d_i\}$.
6. Use a random number generator and the response model with $\beta = 0$ and $\{d_i\}$ to obtain a raw score r for each simulated person.
7. Obtain values of b_f and SEM from Tables 3 and 4.
8. Repeat steps (6) and (7) 100 times to simulate 100 independent replications of the application of this test to this person.
9. Summarize these 100 replications in terms of:

Mean ability estimate or bias, since β=0:	BIAS
Standard deviation of ability estimates:	SA
Standard error of bias:	SE=SA/10
Significance of the bias:	BIAS/SE
Mean standard error of measurement:	ME
Accuracy of SEM as an estimator of SA:	SA/ME

From the results of this simulation study we have derived a set of general rules regarding the degree to which an examiner may depart in practice from a uniform spacing of item difficulties before these measurement procedures produce unacceptably disturbed ability estimates. These rules outline the area of possible test designs (combinations of H, W, and L) within which the bias in measurement is limited to less than .1 log odds ability units. They allow the test designer to incur item discrepancies $e_i = d_i - \delta_i$, that is item calibration errors, as large as 1.0. This may appear to be unnecessarily generous, since it permits use of an item difficulty of 2.0, say, when the design calls for 1.0, but it is offered as an upper limit because we found a large area of the test design domain to be exceptionally robust with respect to independent item discrepancies.

ROBUST REGION OF TEST DESIGN DOMAIN

When no item deviates more than 1 log odds unit from its designed value			
and	($\lvert H-\beta \rvert$<2, W<8, L>10)	then BIAS <.2,	BIAS/SEM <.4
or	($\lvert H-\beta \rvert$<1, W<8, L>30)	then BIAS <.1,	BIAS/SEM <.3

When no item deviates more than 0.5 log odds unit from its designed value			
and	($\lvert H-\beta \rvert$<2, W<8, L>30)	then BIAS <.1,	BIAS/SEM <.2
or	($\lvert H-\beta \rvert$<1, W<8, L>10)	then BIAS <.1,	BIAS/SEM <.2

As test length increases above 30 items, virtually no reasonable testing situation risks a measurement bias large enough to notice. Tests in the neighborhood of 30 items or more, of almost any reasonable width (W < 8) and which come within 1 log odds unit of their targets (IH-Sj<1) are for all practical purposes free from bias caused by random deviations in actual item calibrations of magnitude less than 1 log odds unit. Only when tests are as short as 10 items, wider than 2 log odds units and more than 2 log odds units off-target does the measurement bias caused by item calibration error exceed .2.

Table 5 lists summary statistics for some tests with L = 10.

In Table 5 we see that on the short, narrow test, 2 log odds units above its target, test bias is .25 units with BIAS/SE = 4.2, *indicating* that abilities are overestimated. With only 10 items and a person with β = 0, 2 log odds units below test height H = 2, the expected raw score is less than 2 items. In order to obtain finite ability estimates we are forced to discard zero raw scores thus truncating our distribution of estimates. Hence our simulations in this *instance* produce estimates which are too high and standard errors which are too small leading to an SA/ME ratio well below the expected value of 1.0.

On the short, wide, off-target test the bias of -.29 log odds units is due to the large discrepancy (maximum .8) between design and test. This discrepancy also threatens the short, wide test even when it is on-target. There the

TABLE 5

The Bias in a Short Ten Item
Test Caused by Design Discrepancies

	Narrow Test w = 2 Max. Discrepancy e = .2		Wide Test w = 8 Max. Discrepancy e = .8	
On Target H = B	BIAS	- .03	BIAS	.18
	SA	.77	SA	1.09
	SE	.08	SE	.11
	BIAS/SE	-0.38	BIAS/SE	1.64
	ME	.70	ME	.95
	SA/ME	1.09	SA/ME	1.15
Off Target H=B+2	BIAS	.25	BIAS	- .29
	SA	.58*	SA	.91
	SE	.06	SE	.09
	BIAS/SE	4.16	BIAS/SE	-3.22
	ME	.90	ME	1.01
	SA/ME	.64	SA/ME	.90

Note: Max. Discrepancy $e=d-\delta$ is the largest discrepancy between the designed item difficulty δ and the realized item difficulty d.

BIAS = mean of 100 ability estimates.

SA = standard deviation of 100 ability estimates.

SE = SA/10 standard error of BIAS.

ME = mean SEM of 100 ability estimates.

*Caused by severe truncation of ability estimate distribution due to many zero scores.

bias is .18 log odds units. But BIAS/SE = 1.6 only and .18 is less than .19 = .18/.95 SEM units, so this short, wide test may function well enough when it is more or less on target, since its maximum bias is lost in its SEM.

The short, narrow, on-target test serves as a comparison for the above. Here the bias of -.03 is small, the ratio of ability estimate standard deviation SA to the average standard error of those estimates ME is 1.08, virtually one and so the measurement technique is working well.

EXAMPLES

Example 1

This client has very little knowledge about his target population other than the fact that the persons are, in his best guess, more or less centered on the item pool. Thus his target center is at M = 0 and and we need to come to his aid in selecting his target boundaries. From the appropriate pool of items we select a small number of items whose difficulties range in value from about 2.0 to 4.0. We then ask the client to select that item (or items) which he believes will be too difficult for all but 5 or 10 per cent of his target. Then the calibrated difficulty of that item on the latent variable gives us an approximate upper bound for his target. The lower bound can be obtained by symmetry or by repeating the same steps with easy items.

Let us assume that the suggested upper bound is 4.0 and that the client would like his SEM to be less than 0.5 log odds units. The following sequence of steps would be carried out.

1. M = 0 implies H = 0.
2. Since the upper bound is 2S = 4, the recommended W from Table 2 is 11, if the criterion is the minimization of standard errors at the boundary of the interval target (or 8, if the criterion is the minimization of SEM over a normally distributed target).
3. Using either Table 4 or the simple rule C = 8S provides an upper bound for C of 16. This implies for an SEM = 0.5 that L<C/SEM^2<16/.25<64, or a length of about 60 items.
4. These 60 items are selected from the pool such that they best approximate a uniform distribution with its center at 0.0 and a range of 11. Let us assume that the three easiest items available are -4.0, -4.2

and -4.2 and that the three most difficult are 4.5, 4.3 and 4.2. These six values imply an actual test width of w = 8.4. Finally, let us assume that the actual mean of the 60 item difficulties is d. = 0.4 =h. Thus, while our design was T(0,11,60), our best available test turned out to be t(o.4,8.4,60).

5. This test is administered and we obtain the set of raw scores r. These are converted to relative scores f. If we approximate w = 8.4 by w = 8 we may use that column of Table 3 and h = 0.4 to estimate ability for any f. A person with 42 items correct has f = .7. The corresponding entry in column w = 8 of Table 3 is 1.7 and his corresponding ability estimate is b.7 = 0.4 + 1.7 = 2.1.

6. In order to determine the SEM associated with this ability estimate of 2.1 we go to column w = 8 of Table 4 and°find that

VC - = 3.0 and so SEM = /C / YL = 3/ 60 = 0.4.

Example 2 Compound and Multiple Targets

We have talked about single targets, whether they represent an individual or a group, in order to come to grips with the conceptual and practical problem of the best test design. In practice, however, we may be confronted with several targets spaced out on the latent variable along which we are measuring. Sometimes we want to design a single test which is best for such a series of targets. In that case, the solution is straightforward and simple. Then we have to assume that the series of targets are in fact one large target and from the combined features of this series extract the specifications of that single compound target. The best test design then is the one which minimizes precision or test length for that single compound target.

However, when two or more targets are far enough apart on the latent variable, we will discover that the precision of measurement accessible through a single test designed for all targets combined is unsatisfactory. This will motivate us to see what we can do to increase precision. The only course of action which can accomplish this is to design several tests tailored more exactly to the individual components in the compound target, that is, to design a best test for each of several sub-targets. Of course, this can lead to subtests with no items in common. It might seem that this would place us in a situation where we could no longer compare the results of one

subtest on one sub-target with those from another. But if we remember that our items are calibrated on a common latent variable and that the measures we extract from our various tests will also be on this latent variable, then we see that we can compare the positions of different sub-targets by estimating their location and dispersion on the common variable from measurements made from the individualized tests tailored to each of those targets.

If it should become important to estimate how well persons in one target might have done on a test suitable for another target had they also taken that other test, we can utilize the ability estimates of that target and the difficulties of the items on the other test which they have not taken to estimate the proportion of those untaken items which they would have probably gotten correct had they taken them.

Consider the typical research in which investigators have studied the differences between a control group A and an experimental group B after the latter has received an experi mental treatment. Let us assume that there is evidence to suggest that the treatment has a dramatic effect with the consequence that, group B is expected to end up 3.5 units above group A. To be specific we will allocate the following working specifications for the two targets:

G_A (-1.5, 1.5, D) and G_B (2, 1, D).

Not only is G_A expected to be at a different location on the latent variable but its dispersion is expected to be greater.

Ideally we should construct two separate tests for these two distinct targets. That this would be the preferable strategy will be demonstrated by considering what happens when we compare a compound target strategy with a multiple target strategy. When the two targets are compounded we arrive at G_{AB} (-.25, 2.125, D) by calculating the lower boundary $-1.5 - 2(1.5) = -4.5$ for G_A, the upper boundary $2 + 2(1) = 4$ for G_B, using their mean -.25 as M_{AB} and one-fourth their difference 8.5/4 = 2.125 as S_{AB}. An S of about 2 implies optimum W of 11 for minimization of SEM at the boundaries. To avoid unnecessary complications in this example we will assume that the optimum test is always available (i.e., that W = w and H = h) and that L = 40. Under the multiple target strategy the two optimum tests would then be T_A (-1.5, 8, 40) and T_B (2, 5, 40) while the one compound test would be T_{AB} (-.25, 11, 40).

In Table 6 we have listed the expected results for ten persons ranging in ability from -4.5 to 4.0. The first five may be considered to belong to G_A and the second five to G_B. For every one of these 10 persons, the SEM under a multiple strategy is smaller than under the compound strategy. We can convert these differences into the number of items which could be saved by using the multiple strategy. The middle section of Table 5 shows that if we used only 30 items for T_A and 25 items for T_B we would obtain approximately the same SEM's as achieved on the 40 items compound test. The right section of Table 5 shows the raw scores and SEM's which we would predict if the G_A persons had taken the T_B test and vice versa.

If we were interested in knowing the difference between the two groups in terms of the proportion of items from the pool which each group had correct, then the right section of Table 6 shows that if persons at the center of G_B (e.g., person number #8) had taken T_A their expected score would have been $r = 35$ whereas those in G_A at its center earned score $r = 20$. Thus we can estimate an expected 15 item advantage in score for persons at the center of group B when compared with persons at the center of group A.

SELF-TAILORED TESTING?

When an examiner wishes to measure a person efficiently and objectively he needs not only insight into how the variable in which he is interested is manifested in responses to test items but also an unambiguous and practical statistical technique for identifying which items are qualified to serve as a basis for measurement and for calibrating these qualified items on the variable so that responses to them can be used to estimate measurements. The plausible and statistically sound Rasch response model for selecting and calibrating useful test items and subsequently for measuring persons leads not only to an operational definition of the variable but also to a coherent and practical system for optimal measurement and hence best test design. Within such a system self-tailored testing becomes an easily available by-product.

The person to be measured can be handed a booklet of test items more or less equally spaced in increasing difficulty from easiest to hardest and invited to choose any starting place in the booklet with which he feels comfortable. From that self-chosen starting point the examinee can work at his own will and speed in either direction, forward into harder items or backward into easier ones, until he reaches his own performance limits or runs out of time. Whatever the

TABLE 6

Measurement of two Groups by Multiple and Compound Tests and the Prediction of Each Group on the Alternative Test

Group	Subject	Ability	Compound Test Score	Compound Test SEM L = 40	Multiple Tests Test A Score	Multiple Tests Test A SEM L = 40	Multiple Tests Test A SEM L = 30*	Cross-Over Performance Test B Score	Cross-Over Performance Test B SEM L = 40
A	1	-4.5	6	.58	6	.52	.60	1	1.00
A	2	-3.0	10	.54	13	.47	.55	1	.95
A	3	-1.5	16	.52	20	.45	.53	2	.70
A	4	0.0	21	.52	27	.47	.55	8	.44
A	5	1.5	26	.54	34	.52	.60	17	.38
					Test B Score	Test B SEM L = 40	Test B SEM L = 25*	Test A Score	Test A SEM L = 40
B	6	0.0	21	.52	8	.45	.58	27	.47
B	7	1.0	24	.52	14	.39	.50	32	.48
B	8	2.0	28	.54	20	.38	.48	35	.57
B	9	3.0	31	.55	26	.39	.50	37	.73
B	10	4.0	34	.58	32	.44	.56	39	1.00

*Maximum length necessary to match SEM of compound test.

level and length of this self-chosen seqment, all that are needed to obtain an objective, item-free person measure and its standard error are the serial numbers of the easiest and hardest items tried and the number of successes in between.

These three observations are sufficient to look up in a simple series of tables the person's estimated measure and the standard error of that estimate. That each examinee may work a different segment of items and so perform on his own uniquely self-tailored test, does not interferein any way with the objective comparison of examinees. All measures based on responses to calibrated items are on the single common scale defined by that calibration. Comparisons between persons are made on this common scale and are quite independent of which items were used to make the measurement or why. Of course, the suitability of test items for this kind of use must be planned for in their writiAl~. and carefully developed in their editing. The items must be examined statistically and their qualifications as instruments of measurement demonstrated. But the ubiquitous use of unweighted test scores as sensible and useful measures depends, at bottom, on satisfying just these conditions. If items do not qualify for the self-tailored testing described, then they do not qualify for inclusion in any unweighted test score.

BIBLIOGRAPHY

Andersen, E. B. The numerical solution of a set of conditional estimation equations. Journal of the Royal Statistical Society, 1972, 34, 42-54.

Anderson, E. B. A computer program for solving a set of conditional maximum likelihood equations arising in the Rasch model for questionnaires. Research Memorandum 72-6, *Princeton, N.J.:* Educational Testing Service, 1972.

Andersen, E. B. *Conditional* Inference *and* Models for Measuring, Copenhagen: Mentalhygiejnisk Forlag, 1973.

Birnbaum, A. Some latent trait models and their use in infering an examinee's ability. In F. Lord and M. Novick (Eds.), Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-51.

Douglas, G. A. Test design strategies for the Rasch psychometric model. Unpublished doctoral dissertation, Department of Education, University of Chicago, 1975.

Lord, F. M. Some test theory for tailored testing. In W. K. Holtzman (Ed.), Computer Assisted Instruction. New York: Harper and Row, 1971.

Rasch, G. A mathematical theory of objectivity and its consequences for model *construction.* Report from European Meeting on Statistics, Econometrics and Management Sciences, Amsterdam, 1968.

Samejima, F. Estimating latent ability using a pattern of graded scores. Psychometrika Monograph Supplement, No. 17, Richmond, Va.: The William Byrd Press, 1969.

Wright, B. D. Sample-free test calibration and person measurement. In Proceedings of the 1967 *Invitational* Conference on Testing Problems, Princeton, N.J.: Educational Testing Service, 1967.

Wright, B. D. and Panchepakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-37.

Wright, B. D. and Mead, R. J. Calfit: Sample-free item calibration with a Rasch measurement model. Research Memorandum, No. 18. Statistical Laboratory, Department of Education, University of Chicago, 1975.

Wright, B. D. and Douglas, G. A. Better procedures for sample-free item analysis. Research Memorandum, No. 20. Statistical Laboratory, Department of Education, University of Chicago, 1975.

APPENDIX A

BRYTES

A CONVENIENT METRIC FOR MEASUREMENT

The units of measurement we have been using so far flow directly from the simple logistic response model. Since the estimated probability of a correct response is

$$P_{fi} = \frac{e^{b_f - d_i}}{1 + e^{b_f - d_i}}$$

the odds for a correct response are

$$O_{fi} = P_{fi} / (1 - P_{fi}) = e^{b_f - d_i}$$

and the natural log odds for a correct response are

$$\log O_{fi} = b_f - d_i$$

Thus we have been working in differences between person abilities and item difficulties expressed in natural log odds units. But our origin and scale are quite arbitrary. We may consider whether there is a translation and scaling of these units which would make our measurement metric especially convenient.

Origin

The units must be centered somewhere. Initially we center them locally either on our tests or targets. But in the development of a substantial pool of items marking an important latent variable we would want to establish a standard reference point with some useful meaning in the sphere of measurements. We might pick the easiest possible item (or the least conceivable ability) as the origin and calibrate all

harder items (and greater difficulties) with respect to that zero point. Or we might pick a standard group of items (or persons) and use their center as the origin. If we do not like negative numbers we can shift our whole scale upward so that all calibrations and measurements which can occur fall above zero.

Scale

Unless we can relate our measurements to some units with meaning outside the statistical characteristics of our measurement system, we may as well use the least observable difference in ability, LOD, and/or the least believable difference, SEM, as our smallest unit. Our aim would be to construct our scale so that it conveys increments in SEM and LOD clearly and conveniently. To do this we might try to arrange things so that the minimum SEM we expect to encounter is about 1 or a bit larger while the corresponding minimum LOD is a.bit less than 1 but larger than .1. If we did this, we could round our work to the nearest tenth and use the decimal point to remind us of the difference between LOD and SEM.

The derivation from our response model of LOD and SEM gives us, in natural log odds units,

$$LOD = C/L, \quad \text{where} \quad 4 < C < 9 \quad \text{when} \quad -2 < (\beta - \delta) < 2$$

and $$SEM = \sqrt{LOD} > LOD, \quad \text{as long as} \quad L > C.$$

In addition, we will often want to compare two, more or less, independent measurements with one another to see if they are significantly different. For that it will be useful to define a least significant difference, LSD, as two standard errors of the difference between similarly precise but independent measures,

$$LSD = 2\sqrt{2}\ SEM.$$

In order to discover a scaling factor which will separate SEM from LOD at the decimal point we will consider two boundary situations. First, we seldom measure with tests shorter than 20 items. If we consider the worst precision of such a "short" test, which is at its edge, say 2 log odds units off-center, then we will have a maximum error coefficient C of about 9 and a maximum LOD = 9/20 = .45. Second, we can seldom afford to administer more than 100 items and even at the center of such a "long" test the error coefficient C cannot be less than 4. This produces a minimum LOD = 4/100 = .04.

We have combined these boundary values in the top half of Table 1A and calculated from them the related boundary values of SEM and LSD. We can see that if we want to use the.decimal point to mark the difference between LOD's < 1 and SEM's > 1 then the scaling factor we need is in the neighborhood of 5.

We are working with relationships between log odds and probabilities of a correct response. It might be convenient to have equal increments in log odds attached to easily remembered probabilities. Even odds imply P = .5. Three to one implies P = .75. Nine to one implies P = .90. We see that the exponents of 3 form a nice system.. The probability of a correct response moves from .10 to .25 to .50 to .75 to .90 in equal increments of tog 3. Since the scale of tog 3 in log odds units is 1/tog 3 = .91, we can obtain a scaling factor near 5 by multiplying .91 by 5 to produce the factor 4.55. This scaling factor sets the standard increments at five units each as Table 2A shows.

Our notation for ability measures in natural log odds units has been b. Now we will use B = 4.55b as our notation for our scaled ability measures. Instead of referring to these two metrics as large and small "b" we will call them "log odds units" and "brytes."

The boundary values of LOD, SEM and LSD in brytes are given in the bottom of Table 1A. There we can see that we would not expect an LOD below .2 brytes so that one decimal place in our calibrations and measurements will be quite accurate enough. We also see that SEM ranges between 1 and 3 brytes, so that our least believable differences will run just greater than one bryte. Thus the decimal point, in the bryte scale, nicely separates SEM from LOD. Finally, we see that differences between independent measures must get up to around 4 or 5 brytes before we can begin to take them seriously.

The essential best test design tables for best test design and use in brytes are given in Tables 3A through 5A.

TABLE 1A

Typical Boundary Values for LOD, SEM, LSD
In Natural Log Odds and Brytes = 4.55 Log Odds

Log Odds	LOD	SEM	LSD
Center of Long Test	.04	.20	.57
Edge of Short Test	.45	.67	1.88
Brytes			
Center of Long Test	.2	1	3
Edge of Short Test	2	3	9

where : Center-Long means $b - d. = 0$ $L = 100$ $C = 4$
Edge-Short means $|b - d.| = 2$ $L = 20$ $C = 9$

and Least Observable Difference $LOD = C/L$

Least Believable Difference $SEM = \sqrt{LOD}$

Least Significant Difference $LSD = 2\sqrt{2}\ SEM$

TABLE 2A

Relation Between Relative
Brytes and the Probability of Success

Person Ability Relative to Item Difficulty B - D	Probability of a Correct Response P_{BD}
-20	.01
-15	.04
-10	.10
- 5	.25
0	.50
5	.75
10	.90
15	.96
20	.99

where $$P_{BD} = \frac{e^{(B-D)/4.55}}{1 + e^{(B-D)/4.55}}$$

one Bryte = 4.55 Log Odds

4.55 = 5/log 3.

TABLE 3A

Values of W for Best Uniform Tests
With Normal and Interval Targets in Brytes

Target Std. Dev. S	Normal Target W over $N(M,S^2)$	Interval Target W at $(M\pm 2S)$	Simple Rule W=4S
2.0	0	0	8
2.5	0	0	10
3.0	0	5	12
3.5	0	10	14
4.0	0	15	16
4.5	0	20	18
5.0	0	25	20
5.5	5	30	22
6.0	10	30	24
6.5	15	35	26
7.0	20	40	28
8.0	25	45	32
9.0	35	50	36

TABLE 4A

Relative Ability Estimates X_{fw} in Brytes for Uniform Tests

f \ w	0	5	10	15	20	25	30	35	40	45	50	1-f
.50	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	.50
.52	0.4	0.4	0.4	0.4	0.5	0.6	0.6	0.7	0.8	0.9	1.0	.48
.54	0.7	0.7	0.8	0.9	1.0	1.1	1.3	1.5	1.6	1.8	2.0	.46
.56	1.1	1.1	1.2	1.3	1.5	1.7	1.9	2.2	2.5	2.7	3.0	.44
.58	1.5	1.5	1.6	1.8	2.0	2.3	2.6	2.9	3.3	3.7	4.0	.42
.60	1.8	1.9	2.0	2.2	2.5	2.9	3.2	3.7	4.1	4.6	5.1	.40
.62	2.2	2.3	2.4	2.7	3.0	3.4	3.9	4.4	4.7	5.5	6.1	.38
.64	2.6	2.7	2.9	3.2	3.6	4.0	4.6	5.2	5.8	6.4	7.1	.36
.66	3.0	3.1	3.3	3.6	4.1	4.6	5.3	5.9	6.6	7.4	8.1	.34
.68	3.4	3.5	3.8	4.1	4.6	5.3	5.9	6.7	7.5	8.3	9.1	.32
.70	3.9	3.9	4.2	4.6	5.2	5.9	6.6	7.5	8.3	9.2	10.2	.30
.72	4.2	4.4	4.7	5.2	5.8	6.5	7.3	8.2	9.2	10.2	11.2	.28
.74	4.8	4.9	5.2	5.7	6.4	7.2	8.1	9.0	10.1	11.2	12.3	.26
.76	5.2	5.4	5.7	6.3	7.0	7.8	8.8	9.9	11.0	12.1	13.3	.24
.78	5.8	5.9	6.3	6.9	7.6	8.6	9.6	10.7	11.9	13.1	14.4	.22
.80	6.3	6.4	6.8	7.5	8.3	9.3	10.4	11.6	12.9	14.2	15.5	.20
.82	6.9	7.0	7.5	8.1	9.0	10.1	11.2	12.5	13.8	15.2	16.7	.18
.84	7.5	7.7	8.2	8.9	9.8	10.9	12.1	13.5	14.9	16.3	17.9	.16
.86	8.3	8.4	8.9	9.7	10.6	11.8	13.1	14.5	16.0	17.5	19.1	.14
.88	9.1	9.2	9.7	10.5	11.6	12.8	14.1	15.6	17.1	18.8	20.4	.12
.90	10.0	10.2	10.7	11.5	12.6	13.9	15.3	16.8	18.4	20.1	21.8	.10
.92	11.1	11.3	11.9	12.7	13.9	15.2	16.6	18.2	19.9	21.6	23.4	.08
.94	12.5	12.7	13.3	14.2	15.4	16.8	18.3	19.9	21.7	23.5	25.3	.06
.96	14.5	14.7	15.3	16.2	17.4	18.9	20.4	22.1	23.9	25.8	27.7	.04

If $f > .5$, $B_f = D. + X_{fw}$

$f < .5$, $B_f = D. - X_{(1-f)w}$

TABLE 5A

Error Coefficient $\sqrt{C_{fw}}$ for SEM = $\sqrt{C_{fw}}$ / $\sqrt{L}$ in Brytes

f \ w	0	5	10	15	20	25	30	35	40	45	50	1-f
.50	9.1	9.2	9.5	10.0	10.7	11.4	12.1	12.9	13.7	14.4	15.1	.50
.52	9.1	9.2	9.5	10.0	10.7	11.4	12.1	12.9	13.7	14.4	15.1	.48
.54	9.1	9.2	9.6	10.1	10.7	11.4	12.1	12.9	13.7	14.4	15.2	.46
.56	9.2	9.3	9.6	10.1	10.7	11.4	12.2	12.9	13.7	14.4	15.2	.44
.58	9.2	9.3	9.7	10.1	10.8	11.5	12.2	12.9	13.7	14.4	15.2	.42
.60	9.3	9.4	9.7	10.2	10.8	11.5	12.2	13.0	13.7	14.5	15.2	.40
.62	9.4	9.5	9.8	10.3	10.9	11.6	12.3	13.0	13.8	14.5	15.2	.38
.64	9.5	9.6	9.9	10.4	11.0	11.6	12.4	13.1	13.8	14.5	15.2	.36
.66	9.6	9.7	10.0	10.5	11.1	11.7	12.4	13.1	13.9	14.6	15.3	.34
.68	9.8	9.9	10.2	10.6	11.2	11.8	12.5	13.2	13.9	14.6	15.3	.32
.70	9.9	10.0	10.3	10.8	11.3	12.0	12.6	13.3	14.0	14.7	15.4	.30
.72	10.1	10.2	10.5	11.0	11.5	12.1	12.8	13.4	14.1	14.8	15.4	.28
.74	10.4	10.5	10.8	11.2	11.7	12.3	12.9	13.6	14.2	14.9	15.5	.26
.76	10.7	10.8	11.0	11.4	12.0	12.5	13.1	13.8	14.4	15.0	15.7	.24
.78	11.0	11.1	11.3	11.7	12.2	12.8	13.4	14.0	14.6	15.2	15.8	.22
.80	11.4	11.5	11.7	12.1	12.6	13.1	13.7	14.3	14.8	15.4	16.0	.20
.82	11.8	11.9	12.2	12.5	13.0	13.5	14.0	14.6	15.1	15.7	16.2	.18
.84	12.4	12.5	12.7	13.1	13.5	14.0	14.5	15.0	15.5	16.1	16.6	.16
.86	13.1	13.2	13.4	13.7	14.1	14.6	15.1	15.5	16.0	16.5	17.0	.14
.88	14.0	14.1	14.3	14.6	15.0	15.4	15.8	16.3	16.7	17.2	17.6	.12
.90	15.2	15.2	15.4	15.7	16.1	16.4	16.8	17.2	17.6	18.1	18.5	.10
.92	16.8	16.8	17.0	17.3	17.6	17.9	18.3	18.6	19.0	19.4	19.7	.08
.94	19.2	19.2	19.4	19.6	20.1	20.3	20.4	20.8	21.1	21.4	21.7	.06
.96	23.2	23.2	23.4	23.6	23.8	24.0	24.3	24.5	24.8	25.0	25.3	.04

Simple Rule: when $w < 20$, $9 < \sqrt{C} < 15$

$w > 20$, $\sqrt{4w} < \sqrt{C} < \sqrt{9w}$