RB-69-95

A THEORETICAL STUDY OF TWO-STAGE TESTING

Frederic M. Lord

Office of Naval Research Contract N-00014-69-C-0017
Project Designation NR 150-303
Frederic M. Lord, Principal Investigator

Educational Testing Service
Princeton, New Jersey

December 1969

RB-69-95

A THEORETICAL STUDY OF TWO-STAGE TESTING

Frederic M. Lord

Office of Naval Research Contract N-00014-69-C-0017

Project Designation NR 150-303

Frederic M. Lord, Principal Investigator

Educational Testing Service

Princeton, New Jersey

December 1969

A THEORETICAL STUDY OF TWO-STAGE TESTING

Frederic M. Lord

Educational Testing Service

ABSTRACT

When items cannot be answered correctly by guessing, certain two-stage testing procedures are about as effective over the ability range of interest as the "best" up-and-down procedures studied previously. When answers can be guessed correctly 20 percent of the time, no two-stage procedure is found to match the "best" up-and-down procedures over this ability range. Feet-on-the-desk designs for two-stage procedures may produce poor results.

A THEORETICAL STUDY OF TWO-STAGE TESTING[1]

Frederic M. Lord

Educational Testing Service

A two-stage testing procedure consists of a routing test followed by one of several alternative second-stage tests. All tests are of conventional type. The choice of the second-stage test administered is determined by the examinee's score on the routing test.

The main advantage of such a procedure lies in matching the difficulty level of the second test to the ability level of the examinee. Since conventional tests are usually at a difficulty level suitable for typical examinees in the group tested, two-stage testing procedures are likely to be advantageous chiefly at the extremes of the ability range.

Two-stage testing is discussed by Cronbach and Gleser (1965, chapt. 6), using a decision theory approach. They deal primarily with a situation where examinees are to be selected or rejected. Their approach is chiefly sequential in the special sense that the second-stage test is administered only to borderline examinees. All advantages of this procedure come from varying the amount of testing according to the ability level of the examinee.

In contrast, the present paper is concerned with situations where the immediate purpose of the testing is measurement, not classification. In this paper, the total number of test items administered to a single examinee is fixed. Any advantage of two-stage testing appears as

---

[1]This work was supported in part by contract N-00014-69-C-0017 between the Personnel and Training Research Programs Office, Psychological Sciences Division, Office of Naval Research and Educational Testing Service.

improved measurement. Some empirical studies of such two-stage testing are reported by Linn, Rock, and Cleary (1969), who also cite other references.

The present study attempts to find, under specified restrictions, some good designs for two-stage testing. A "good" procedure is one that provides reasonably accurate measurement for examinees who would obtain near-perfect or near-zero (or near-chance-level) scores on a conventional test.

The particulars at our disposal in designing a two-stage testing procedure include the following:

1. the number of items given to a single examinee (n),
2. the number of alternative second-stage tests available for use,
3. the number of alternative responses per item,
4. the number of items in the routing test (n_1),
5. the difficulty level of the routing test,
6. the method of scoring the routing test,
7. the cutting points for deciding which second-stage test an examinee will take,
8. the difficulty levels of the second-stage tests,
9. the method of scoring the entire two-stage procedure.

It does not seem feasible to locate truly "optimum" designs. The present study has proceeded by investigating several designs, modifying the best of these in various ways, choosing the best of the modifications, and continuing in this fashion as long as any modification can be found that noticeably improves results.

Nearly 200 different two-stage designs have been investigated in this process. Obviously, an empirical investigation of 200 designs would have been out of the question. Instead, theoretical investigations were carried out, with the aid of a high-speed computer, based on item characteristic curve theory.

Specifications

Let us start by restricting our attention to tests composed of dichotomously scored items. The mathematical model to be used assumes that $P_i \equiv P_i(\theta)$, the probability of a correct response to item i, is a generalized normal-ogive function of the examinee's ability (or standing on the trait measured):

$$P_i(\theta) = c_i + (1 - c_i)\Phi[a_i(\theta - b_i)] \quad , \tag{1}$$

where $\Phi(t)$ represents the normal distribution cumulative frequency up to the relative deviate t. This assumes that the items to be used are all homogeneous in the sense that they all measure the same psychological trait.

The quantities a_i, b_i, and c_i are parameters describing item i. The ogive $P_i(\theta)$ has its point of inflection at $\theta = b_i$. As θ becomes negatively large, $P_i(\theta)$ approaches its lower asymptote $P_i = c_i$. For fixed c_i, the slope at the point of inflection is proportional to a_i. Thus a_i is thought of as representing <u>item discriminating power</u>, b_i as representing <u>item difficulty</u>, and c_i

as a sort of practical chance-score level. A detailed discussion of these parameters from the present point of view is given by Lord (1969, sections 3, 4).

For the sake of simplicity, let us assume that the available items differ only in difficulty, b_i. They all have equal discriminating power, denoted by a, and equal practical chance-score levels, c. Also, let us consider the case where the routing test and each of the second-stage tests are peaked; that is, each subtest is composed of items all of equal difficulty.

Scoring

For a peaked test, it is known (Birnbaum, 1968, chapter 18) that the number-right score (number of right answers), to be denoted by x, is a sufficient statistic for estimating an examinee's ability θ. Thus at first sight it might seem that there is no problem in scoring a two-stage testing procedure when all subtests are peaked. However, it is clear that different estimates of θ should be used for examinees who obtain the same number-right score, but on different second-stage tests having different difficulty levels.

What is needed is to find a function of the sufficient statistic x that is an unbiased estimator, or at least a consistent estimator, of θ. The maximum likelihood estimator, to be denoted by $\hat{\theta}$, satisfies these requirements, and will be used here.

For an m-item peaked subtest, the likelihood function is

$$L(x|\theta) = \binom{m}{x} P^x Q^{m-x} \quad , \tag{2}$$

where P is the item characteristic function defined by (1) for each of the m items, and where $Q = 1 - P$. Differentiate the logarithm of the likelihood

$$\log L(x|\theta) = \log\binom{m}{x} + x \log P + (m - x)\log Q \tag{3}$$

with respect to θ to obtain

$$\frac{\partial \log L(x|\theta)}{\partial \theta} = \frac{xP'}{P} - \frac{(m - x)P'}{Q}$$

$$= \frac{P'}{PQ}(x - mP) \quad , \tag{4}$$

where P' is the derivative of P with respect to θ. When (4) is set equal to zero, we obtain the likelihood equation

$$P(\hat{\theta}) = \frac{x}{m} \quad . \tag{5}$$

Substituting (5) into (1) and solving for θ, we have

$$\theta[a(\hat{\theta} - b)] = \frac{x/m - c}{1 - c} \quad ,$$

where a, b, and c describe each item of the peaked subtest. The maximum likelihood estimator is found by solving for $\hat{\theta}$:

$$\hat{\theta} = \frac{1}{a}\Phi^{-1}\left(\frac{x/m - c}{1 - c}\right) + b \quad , \tag{6}$$

where Φ^{-1} is the inverse of the function Φ (Φ^{-1} is the relative deviate corresponding to a given normal curve area).

Equation (6) gives a sufficient statistic that is also a consistent estimator of θ having minimum variance in large samples. The separate use of (6) for the routing test and for the second-stage test yields two such estimates, $\hat{\theta}_1$ and $\hat{\theta}_2$, for any given examinee. These are jointly sufficient statistics for θ. They must be combined into a single estimate. However, there is no uniquely good way to do this.

In the present study, $\hat{\theta}_1$ and $\hat{\theta}_2$ are averaged after weighting them inversely according to their (estimated) large-sample variances. This is the weighting that produces a consistent estimator with minimum large-sample sampling variance. Thus an examinee's score $\tilde{\theta}$ on the two-stage test will be proportional to

$$\frac{\hat{\theta}_1}{\text{Var } \hat{\theta}_1} + \frac{\hat{\theta}_2}{\text{Var } \hat{\theta}_2} \quad .$$

Specifically, his overall score is defined as

$$\tilde{\theta} = \frac{\hat{\theta}_1 \hat{V}(\hat{\theta}_2) + \hat{\theta}_2 \hat{V}(\hat{\theta}_1)}{\hat{V}(\hat{\theta}_1) + \hat{V}(\hat{\theta}_2)} \quad , \tag{7}$$

where $\hat{V}$ is an estimate of Var , so that asymptotically

$$E\tilde{\theta} = \frac{\theta \text{ Var } \hat{\theta}_2 + \theta \text{ Var } \hat{\theta}_1}{\text{Var } \hat{\theta}_1 + \text{Var } \hat{\theta}_2} = \theta \quad . \tag{8}$$

By a well known theorem,

$$\text{Var } \hat{\theta} = \{\mathcal{E}[\frac{\partial \log L(x|\theta)}{\partial \theta}]^2\}^{-1} \quad .$$

From (4), then,

$$\text{Var } \hat{\theta} = \{\frac{P'^2}{P^2Q^2}\,\mathcal{E}(x - mP)^2\}^{-1} \quad .$$

By (2), x has a binomial distribution with mean mP and variance

$$\mathcal{E}(x - mP)^2 = mPQ \quad ,$$

so that

$$\text{Var } \hat{\theta} = \frac{PQ}{mP'^2} \quad . \tag{9}$$

By (1),

$$P' = (1 - c)a\phi[a(\theta - b_1)] \tag{10}$$

where $\phi(t)$ is the normal curve ordinate at the relative deviate t. In practice, $\hat{V}(\hat{\theta}_1)$ and $\hat{V}(\hat{\theta}_2)$ were obtained by substituting $\hat{\theta}_1$ or $\hat{\theta}_2$, respectively, for θ in the right-hand sides of (9) and (10).

When $x = m$ or $x = cm$, the $\hat{\theta}$ defined by (6) would be infinite. To avoid this, whenever $x = m$, x was in practice replaced by $x = m - 1/2$. Whenever $x \leq cm$ and $x + 1 > cm$, the lower of these two scores was replaced by $(x + 1 + cm)/2$. At the same time, all other scores lower than $(x + 1 + cm)/2$ were also replaced by $(x + 1 + cm)/2$.

The score $\tilde{\theta}$ so constructed from (7) will not have strictly optimum properties for small n ; however, this is typical of estimation problems where (as here) no single sufficient estimator exists. Two-stage testing is on its face a rather inefficient method of tailored testing. Any additional inefficiency from the use of $\tilde{\theta}$ should be of relatively minor importance.

Evaluation of Procedures

If there are n_1 items in the routing test and n_2 items in the second-stage test, there are at most $n_1 n_2$ different possible numerical values for $\tilde{\theta}$. Let $\tilde{\theta}_{xy}$ denote the value of $\tilde{\theta}$ when the number-right scores on the routing test and on the second-stage test are, respectively, x and y . By (2), the frequency distribution of $\tilde{\theta}$ is

$$\text{Prob}(\tilde{\theta} = \tilde{\theta}_{xy}|\theta) = \binom{n_1}{x} P_1^x Q_1^{n_1-x} \binom{n_2}{y} P_2^y Q_2^{n_2-y} \qquad (11)$$

where P_1 is given by (1) with $a_i = a$, $c_i = c$, and b_i equal to the difficulty level (b , say) of the routing test; and where P_2 is similarly given by (1) with $b_i = b(x)$, a numerical function of x assigned in advance by the psychometrician.

Given numerical values for n_1 , n_2 , a , b , c , and for $b(x)$, $x = 0,1,\ldots,n_1$, the exact frequency distribution of the examinee's score $\tilde{\theta}$ for examinees at any given ability level θ can be computed from (11). These frequency distributions contain all possible information relevant for choosing among specified two-stage testing procedures.

In actual practice, it is necessary to summarize somehow the plethora of numbers computed from (11). This has been done here by using the information function $I_x(\theta)$ discussed at some length in Lord (1969, 1971). By definition,

$$I_x(\theta) = \frac{[\frac{\partial}{\partial\theta} \mathcal{E}(x|\theta)]^2}{\mathrm{Var}(x|\theta)} \quad , \tag{12}$$

where x represents whatever test score is used. For two-stage testing with test score $\tilde{\theta}$, the symbol x is replaced by $\tilde{\theta}$.

For a given θ, the denominator of (12) is computed in straightforward fashion from the conditional frequency distribution (11). Denoting the probability in (11) by p_{xy}, we have

$$\mathcal{E}(\tilde{\theta}|\theta) = \sum_{x=0}^{n_1} \sum_{y=0}^{n_2} p_{xy}\tilde{\theta}_{xy} \quad .$$

Since $\tilde{\theta}_{xy}$ is not a function of θ,

$$\frac{\partial}{\partial\theta} \mathcal{E}(\tilde{\theta}|\theta) = \sum_{x=0}^{n_1} \sum_{y=0}^{n_2} \tilde{\theta}_{xy} \frac{\partial p_{xy}}{\partial\theta} \quad .$$

A formula for $\partial p_{xy}/\partial\theta$ is easily written down from (11), from which numerical values of the numerator of (12) are then calculated for given θ. In this way, $I_{\tilde{\theta}}(\theta)$ is evaluated numerically for all ability levels of interest.

The information function $I_x(\theta)$ is (approximately) an index of how effective the testing and scoring procedures are for measuring the

examinee. For a conventional type of test, the value of $I_x(\theta)$ is directly proportional to the number of test items. The numerical value of $I_x(\theta)$ for a single testing procedure ordinarily is not interpreted by itself, but only in comparison to the value of $I_x(\theta)$ for some other procedure. Thus, if $I_{\tilde{\theta}}(\theta)$ for one procedure is r times as large as $I_{\tilde{\theta}}(\theta)$ for a second procedure, this is to be interpreted as representing an improvement in measurement effectiveness equivalent to that obtained by lengthening a conventional test r times.

Explanation of Figures

Figure 1 shows the information functions for five different testing procedures. The two solid curves are benchmarks, with which curves for three two-stage procedures are compared.

The "standard" curve shows the information function for the number-right score on a 60-item peaked test of the conventional type. The items all have the same difficulty level, b , and the same discriminating power, a .

The vertical scale represents amount of information obtained, as a function of ability level, θ , the latter being shown along the horizontal scale. Instead of drawing a different information curve for each pair of values, a and b , that is of interest, it very conveniently turns out to be possible to choose the units of measurement for the horizontal and vertical scales so that a single information curve will be valid for any a and for any b . This has been done

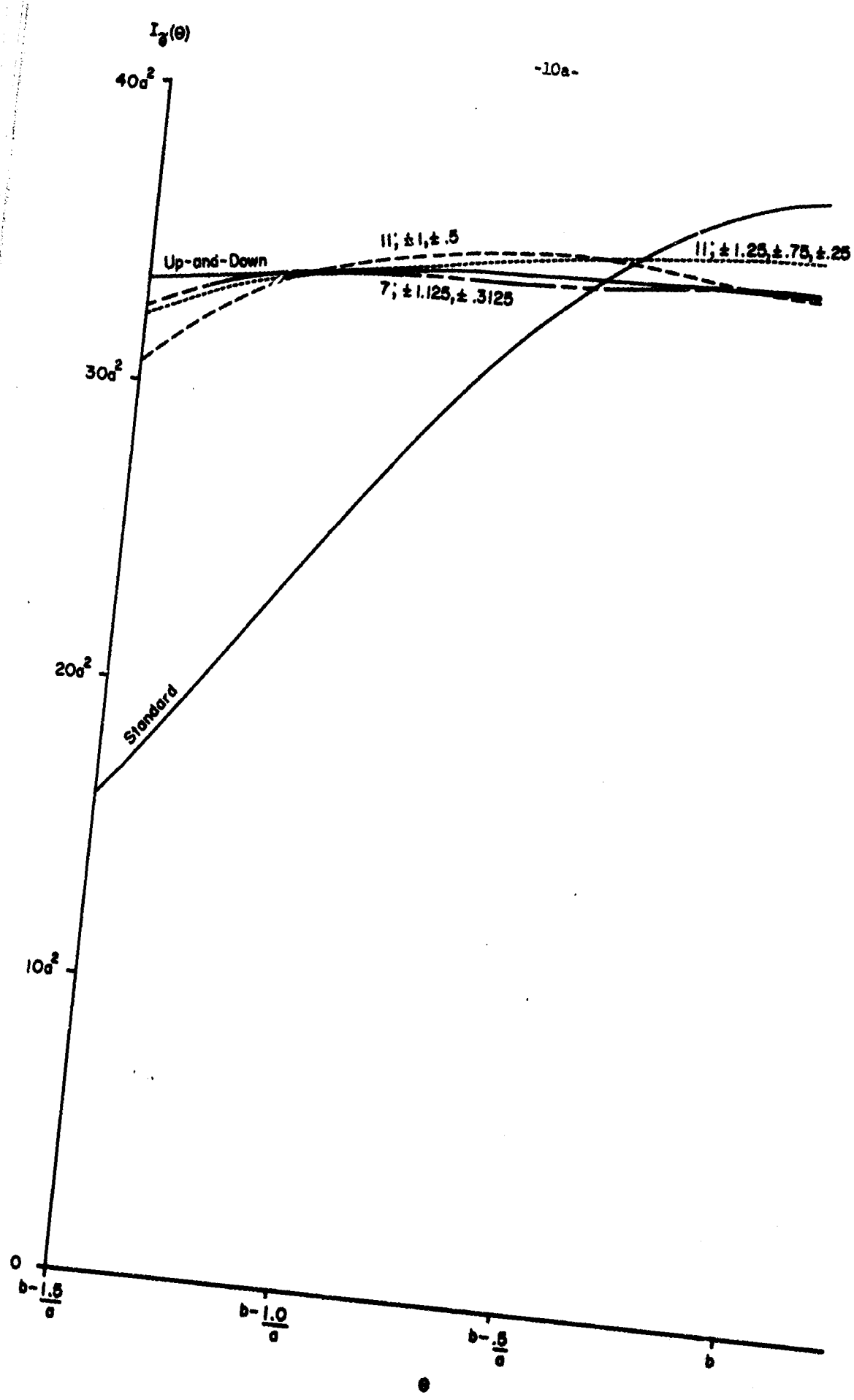

Fig. 1 Two-Stage Testing, $n=60$, $a=0$.

for the figures shown here, which explains why the scale values shown along the horizontal and vertical scales are functions of a and b. Only information curves symmetrical about $\theta = b$ were investigated when $c = 0$. For this reason, only the left portion of each curve is shown in Figure 1.

Although we are directly concerned here with testing single individuals (there may be just one examinee, not a group), the reader needs to know what range of θ is of concern to him. If a 60-item peaked test with $c = 0$, $b = 0$, and $a = 1.00$ is administered to a group in which θ is normally distributed with $\mu_\theta = 0$ and $\sigma_\theta = 0.5$, the test reliability will be 0.90 (see Lord, 1969, section 4). If a reliability of .90 is roughly what the reader would expect for 60-item tests and examinee groups that he is concerned with, and if his groups have roughly a normal distribution of ability, then roughly two-thirds of his examinees should fall between $\theta = -0.5$ and $+0.5$, that is (since $b = 0$ and $a = 1$), between $\theta = b - 0.5/a$ and $\theta = b + 0.5/a$. If the reader is interested only in this subrange of ability, he will not find it profitable to use two-stage testing of the kinds considered here. It is assumed, therefore, that he may be interested in the range from $\theta = b - 1.5/a$ to $\theta = b + 1.5/a$, or perhaps, from $\theta = b - 1/a$ to $\theta = b + 1/a$.

Suppose, next, that the reader is concerned about a testing situation where $c = 0$, $b = 0$, $a = .50$ and $\mu_\theta = 0$, $\sigma_\theta = 1.0$. The test items

here are only half as discriminating as those considered before, but the group tested is twice as heterogeneous. These changes offset each other, so that a 60-item peaked test will again have a reliability of 0.90. The reader will still look at the same segments of the information curves as before, since the range from $\theta = b - 1.5/a$ to $\theta = b + 1.5/a$, for example, will still (since $a = .50$) cover 6 standard deviation units of ability for the group tested.

Suppose, next, that the reader is concerned with an unusual testing situation where a 60-item peaked test typically has a reliability of .80. This may occur either because his items have low values of a or because his examinees are rather homogeneous. A reliability of .80 will be found if $a = .33$ and $\sigma_\theta = 1$; or if, alternatively, $a = 1$ and $\sigma_\theta = .33$. In either case, the range from $\theta = b - 1/a$ to $\theta = b + 1/a$ covers 6 standard deviation units of ability in the group tested. In this case the reader will probably wish to ignore the left third of Figure 1 as representing extreme ability levels so rare that they can be neglected.

Suppose, finally, that the reader is concerned with an unusual situation where a 60-item peaked test typically has a reliability of .97. This would occur if $a = 1.00$ and $\sigma_\theta = 1$, or if $a = .50$ and $\sigma_\theta = 2$. In this case, the range from $\theta = b - 1.5/a$ to $\theta = b + 1.5/a$ covers only 3 standard deviations, representing the middle 87 percent of the group tested.

There is no assumption of a normal or other frequency distribution underlying the figures. The point is simply that the reader needs to know what range of θ is of interest to him. If his examinees are

asymmetrically distributed, or if he is chiefly interested in only part of the ability range of the group tested, then he will pick the portion of Figure 1 that interests him accordingly.

In choosing among two-stage testing procedures, a procedure can be eliminated if computations show that its information curve is always lower than the curve of some other procedure, regardless of θ level. Commonly, however, information curves cross, showing that one procedure provides better measurement at certain ability levels, whereas another procedure is better at other levels.

As already pointed out, an examiner who wants accurate measurement for typical examinees in the group tested and is less concerned about accurate measurement at the extremes should use a peaked conventional test. If a two-stage procedure is to be really valuable, it will usually be because it provides good measurement for extreme as well as for typical examinees. For this reason, the main effort in the present study has been to find two-stage procedures with information curves similar to (or better than) "up-and-down" curves shown in the figures. These last are benchmark curves, chosen as the "best" of those obtained by the up-and-down method of tailored testing (see Lord, 1969). The up-and-down curve shown here in Figure 1 is the curve labeled $ad = .20$, $c = 0$ shown there in Figure 7.6.

Results for 60-Item Tests with No Guessing

Surprisingly, Figure 1 shows that when there is no guessing it is possible to approximate the measurement efficiency of a 60-item up-and-down tailored testing procedure by a 60-item two-stage procedure

throughout the ability range from $\theta = b - 1.5/a$ to $\theta = b + 1.5/a$. The effectiveness of the two-stage procedures shown falls off rather sharply outside this ability range, but this range is adequate or more than adequate for most testing purposes (as explained in the preceding section).

The label "11; ±1, ±.5" indicates that the routing test contains $n_1 = 11$ items (at difficulty b), and that there are four alternative 49-item second-stage tests with difficulty levels $b - 1/a$, $b - .5/a$, $b + .5/a$, and $b + 1/a$. The cutting points on this routing test are equally spaced in terms of number-right scores, x_1: if $x_1 = 0\text{-}2$, the examinee is routed to the easiest second-stage test; if $x_1 = 3\text{-}5$, to the next easiest; and so on.

The label "7; ±1.125, .3125" is similarly interpreted, the examinees being routed according to the score groupings $x_1 = 0\text{-}1$, $x_1 = 2\text{-}3$, $x_1 = 4\text{-}5$, $x_1 = 6\text{-}7$. The label "11; ±1.25, ±.75, ±0.25" similarly indicates a procedure with six alternative second-stage procedures, assigned according to the groupings $x_1 = 0\text{-}1$, $x_1 = 2\text{-}3$, ... $x_1 = 10\text{-}11$.

A 60-item up-and-down procedure in principle requires 1,830 items before testing can start. In practice, 600 items might be adequate without seriously impairing measurement. Two of the two-stage procedures shown in Figure 1 require slightly more than 200 items.

The two-stage procedures shown in Figure 1 are the "best," also the last ones tried, out of sixty-odd 60-item procedures studied with $c = 0$. None of the two-stage procedures that at first seemed promising according to armchair estimates turned out particularly well. From

Table 1

Information for Various 60-Item Testing Procedures with $c = 0$

Procedure*	Information** at θ = b-1.5/a	b-1/a	b-0.5/a	b
Up-and-down (benchmark)	33.5	34.3	34.9	35.1
7; ± 1.125, ± .3125	32.5	34.4	34.5	35.1
7; ± 1 , ± .25	31.1	34.2	35.1	35.8
7; ± 1 , ± .25*	27.0	31.4	35.8	37.0
7; ± 1.25 , ± .25	33.2	33.7	33.7	35.1
7; ± .75 , ± .25	28.0	33.7	35.9	36.5
11; ± 1 , ± .25	30.4	34.1	35.5	36.8
11; ± 1 , ± .5	30.6	34.8	35.6	34.9
11; ± 1.25 , ± .375	32.6	34.0	34.6	35.5
3; ± .75 , ± .25	27.6	32.9	34.9	35.2
3; ± .75 , ± .5	28.0	33.8	34.0	33.4
7; ± .75	28.6	34.4	34.5	31.4
7; ± .5	24.4	32.9	36.0	34.9
3; ± .5	24.5	32.5	34.5	34.4

*All cutting points are equally spaced, except for the starred procedure, which has score groups $x_1 = 0$, $x_1 = 1\text{-}3$, $x_1 = 4\text{-}6$, $x_1 = 7$.

**All information values are to be multiplied by a^2 .

this experience, it seems that casually designed two-stage tests are likely to provide fully effective measurement only over a relatively narrow range of ability, or possibly not at all.

Discussion of Results for 60-Item Tests with No Guessing

Table 1 shows the information at four different ability levels obtainable from some of the better procedures. The following generalizations are tentative and may not hold in situations quite different from those studied here.

Length of routing test. If the routing test is too long, not enough items are left for the second-stage test, so that measurement may be effective near $\theta = b$, but not at other ability levels. If the routing test is too short, then examinees are poorly allocated to the second-stage tests. In this case, if the second-stage tests all have difficulty levels near b, then effective measurement may be achieved near $\theta = b$ but not at other ability levels; if the second-stage tests differ considerably in difficulty level, then the misallocation of examinees may lead to relatively poor measurement at all ability levels. The results shown in Table 1 and Figure 1 suggest that $n_1 = 3$ is too small and $n_1 = 11$ is too large for the range $b \pm 1.5/a$ in the situation considered, assuming that no more than four second-stage tests are used.

Number of second-stage tests. There cannot usefully be more than n_1 second-stage tests. The number of such tests will also often be limited by considerations of economy. If there are only two second-

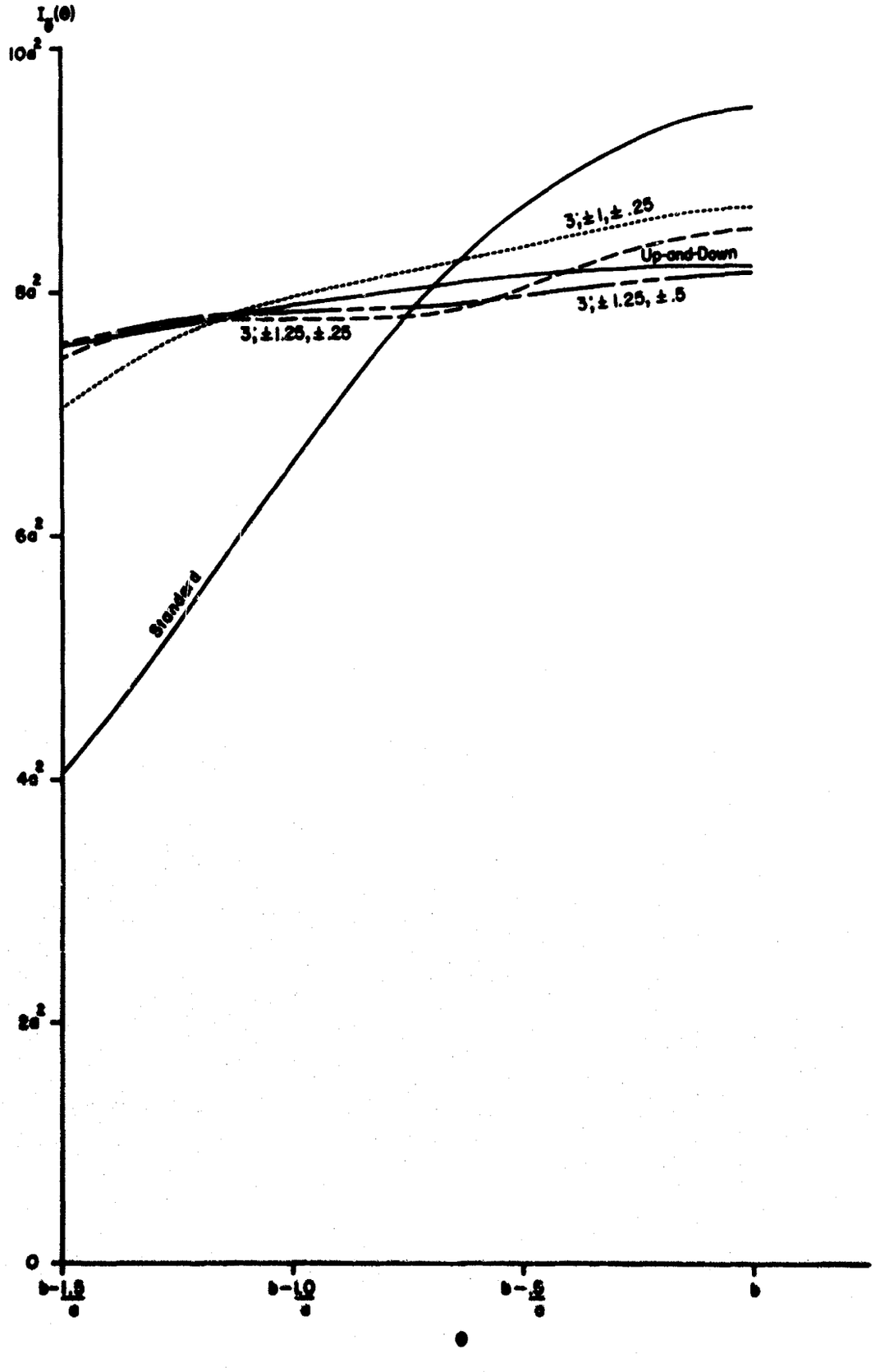

Fig. 2 Two-Stage Testing, n=15, c=0.

stage tests, good measurement may be obtained in the subranges of ability best covered by these tests, but not elsewhere (see "7; ±.75" in Table 1). On the other hand, a short routing test cannot make sufficiently accurate allocations to justify a large number of second-stage tests. In the present study, the number of second-stage tests was kept as low as possible; however, at least four second-stage tests were required to achieve effective measurement over the ability range considered.

Difficulty of second-stage tests. If the difficulty levels of the second-stage tests are all too close to b, there will be poor measurement at extreme ability levels (see "7; ±.75, ±.25" in Table 1). If the difficulty levels are too extreme, there will be poor measurement near $\theta = b$.

Cutting points on routing test. It is clearly important that the difficulty levels of the second-stage tests should match the ability levels of the examinees allocated to them, as determined by the cutting points used on the routing test. It is difficult to find an optimal match by the trial-and-error methods used here. Although many computer runs were made using unequally spaced cutting points, like those indicated in the footnote to Table 1, equally spaced cutting points turned out better. This matter deserves more careful study.

Results for 15-Item Tests with No Guessing

Some 40-odd different procedures were tried out for the case where a total of $n = 15$ items with $c = 0$ are to be administered to each

Table 2

Information for Various 15-Item Testing Procedures with $c = 0$

Procedure*	Information** at $\theta =$ b-1.5/a	b-1/a	b-0.5/a	b
Up-and-down (benchmark)	7.6	7.9	8.1	8.2
3; ± 1.25, ± .5	7.6	7.8	3.0	8.2
3; ± 1.25, ± .25	7.4	7.8	8.0	8.5
3; ± 1 , ± .25	7.0	8.0	8.4	8.7
7; ± 1.25, ± .5	6.5	7.6	8.4	8.5
5; ± 1.5 , ± 1 , ± .5	7.2	7.7	8.0	8.1
4; ± 1 , 0*	7.1	8.0	8.0	8.0
2; ± 1 , 0	7.2	8.0	8.0	7.9
3; ± .25	4.8	7.1	8.7	9.1
7; ± 1	6.2	7.8	8.0	7.5

*All cutting points are equally spaced, except for the starred procedure, which has score groups $x_1 = 0\text{-}1$, $x_1 = 2$, $x_1 = 3\text{-}4$.

**All information values are to be multiplied by a^2 .

examinee. The "best" of these--those with information curves near the up-and-down benchmark--are shown in Figure 2. The benchmark here is again one of the "best" up-and-down procedures (see Stocking, 1969, Fig. 2, curve labeled "A" and "ad = .50").

Table 2 shows results for various other two-stage procedures not quite as "good" as those in Figure 2. In general, these others either did not measure well enough at extreme ability levels, or else did not measure well enough at $\theta = b$. The results for $n = 15$ seem to require no further comment, since the general principles are the same as for $n = 60$.

Results for 60-Item Tests with Guessing

About 75 different 60-item two-stage procedures with $c = .20$ were tried out. The "best" of these are shown in Figure 3 along with an appropriate benchmark procedure (see Lord, 1969, Fig. 7.8, curve labeled "ad = .25, H = 1, L = 2").

Apparently, when items can be answered correctly by guessing, two-stage testing procedures are not as effective for measuring at extreme ability levels as are the better up-and-down procedures. Unless some really "good" two-stage procedures were missed in the present investigation, it appears that a two-stage test might require ten or more alternative second stages in order to measure well throughout the range shown in Figure 3. Such tests were not studied here because the cost of producing so many second stages may be excessive. Very possibly, a three-stage procedure would be preferable.

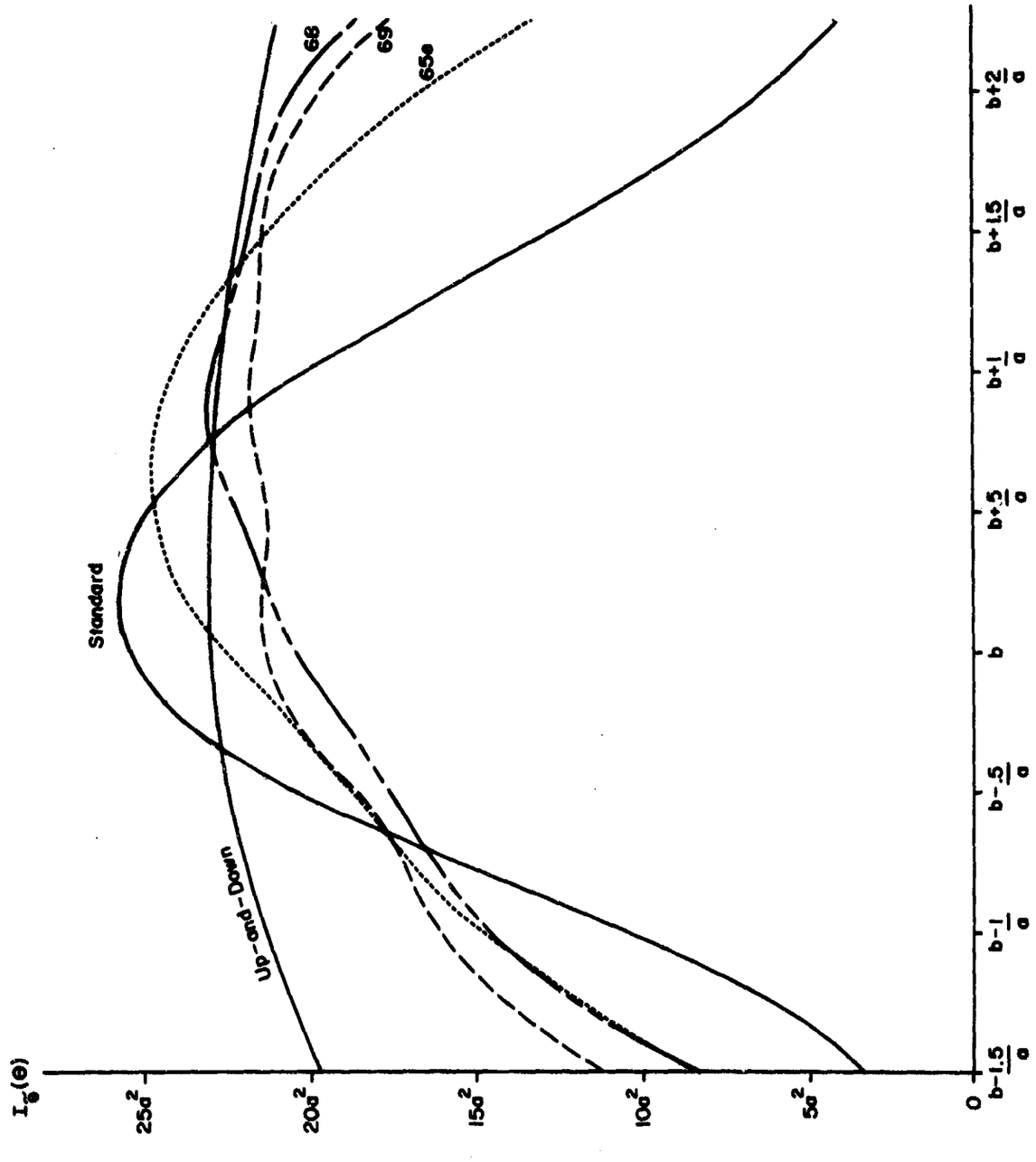

Fig. 3 Two-Stage Procedure n=60, c=.2.

When there is guessing, maximum information is likely to be obtained at an ability level higher than $\theta = b$, as is apparent from Figure 3. This means that the examiner will probably wish to choose a value of b (the difficulty level of the routing test) somewhat below the mean ability level of the group to be tested. If a value of b were chosen near μ_θ, the mean ability level of the group, as might well be done if there were no guessing, then the two-stage procedures shown in Figure 3 would provide good measurement for the top examinees (above $\theta = b + 1/a$) but quite poor measurement for the bottom examinees (below $\theta = b - 1/a$). If an examiner wants good measurement over two or three standard deviations on each side of the mean ability level of the group, he should choose the value of b for the two-stage procedures in Figure 3 so that μ_θ falls near $b + .75/a$. In this way, the ability levels of his examinees might be covered by the range from $\theta = b - .75/a$ to $\theta = b + 2.25/a$, for example.

The three two-stage tests shown in Figure 3 are as follows. Test 68 has an 11-item routing test with six score groups $x_1 = 0\text{-}3, 4, 5\text{-}6, 7\text{-}8, 9\text{-}10, 11$, corresponding to six alternative second-stage tests at difficulty levels b_2 where $a(b_2 - b) = -1.35, -.65, -.325, +.25, +.75$, and $+1.5$. Test 69 has a 17-item routing test with $x_1 = 0\text{-}5, 6\text{-}7, 8\text{-}10, 11\text{-}13, 14\text{-}15, 16\text{-}17$ and $a(b_2 - b) = -1.5, -.75, -.25, +.35, +.9, +1.5$. Test 65e has an 11-item routing test with $x_1 = 0\text{-}2, 3\text{-}4, 5\text{-}6, 7\text{-}8, 9\text{-}10, 11$ and $a(b_2 - b) = -1.5, -.9, -.3, +.2, +.6, +1.0$.

A table of numerical values would be bulky and will not be given here. Most of the conclusions apparent from such a table have already been stated.

No investigations have been carried out for shorter ($n < 60$) two-stage procedures with $c = .2$.

Summary

Various two-stage testing procedures were studied, using a mathematical model provided by mental test theory. When the test items cannot be answered correctly by guessing, certain two-stage procedures were found to be about as effective over the ability range of interest as were the "best" of the up-and-down tailored testing procedures studied previously (Lord, 1969). When low-ability examinees are able to answer all items correctly at least 20 percent of the time, however, no two-stage procedure was found that matched the effectiveness of the "best" up-and-down tailored procedures over this ability range.

This writer's feet-on-the-desk designs for two-stage procedures were found to produce comparatively poor results. Careful preliminary investigations may be required in order to obtain effective measurement over a wide range of ability.

REFERENCES

BIRNBAUM, ALLAN. "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In Frederic M. Lord and Melvin R. Novick, Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968. Chapter 18.

CRONBACH, LEE J. and GLESER, GOLDINE C. Psychological Tests and Personnel Decisions. (2nd ed.) Urbana, Ill.: University of Illinois Press, 1965. 347 pp.

LINN, ROBERT L., ROCK, DONALD A., and CLEARY, T. ANNE. "The Development and Evaluation of Several Programmed Testing Methods." Educational and Psychological Measurement. 29: 129-146; Spring 1969.

LORD, FREDERIC M. "Some Test Theory for Tailored Testing." Computer Assisted Instruction, Testing, and Guidance. (Edited by Wayne H. Holtzman). New York: Harper and Row, 1969. Chapter 7.

LORD, FREDERIC M. "Robbins-Monro Procedures for Tailored Testing." Educational and Psychological Measurement, 1971, in press.

STOCKING, MARTHA. "Short Tailored Tests." Research Bulletin 69-63 and ONR Technical Report, Contract N00014-69-C-0017. Princeton, N. J.: Educational Testing Service, July 1969. 13 pp.

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Educational Testing Service Princeton, New Jersey 08540	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

A THEORETICAL STUDY OF TWO-STAGE TESTING

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

5. AUTHOR(S) (First name, middle initial, last name)

Frederic M. Lord

6. REPORT DATE December 1969	7a. TOTAL NO. OF PAGES 26	7b. NO. OF REFS 6
8a. CONTRACT OR GRANT NO. N-00014-69-C-0017 b. PROJECT NO. NR 150-303 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) RB-69-95 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Office of Naval Research Navy Department Washington, D. C. 20360

13. ABSTRACT

When items cannot be answered correctly by guessing, certain two-stage testing procedures are about as effective over the ability range of interest as the "best" up-and-down procedures studied previously. When answers can be guessed correctly 20 percent of the time, no two-stage procedure is found to match the "best" up-and-down procedures over this ability range. Feet-on-the-desk designs for two-stage procedures may produce poor results.

DD FORM 1 NOV 68 1473 (PAGE 1)

S/N 0101.807.6801

Security Classification

AMD PPSD 13353

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
mental testing						

DD FORM NOV 65 1473 (BACK)
(PAGE 2)

Security Classification