ED 171 720 TM 006 663

AUTHOR	Levine, Michael V.; Rubin, Donald E.
TITLE	Measuring the Appropriateness of Multiple-Choice Test Scores.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	College Entrance Examination Board, New Ycrk, N.Y.; Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RB-76-31
PUB DATE	Dec 76
NOTE	31p.
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Ability; *Aptitude Tests; College Entrance Examinations; Item Analysis; *Mathematical Formulas; *Multiple Choice Tests; Predictive Measurement; Probability; Response Style (Tests); Scores; Senior High Schools; Test Bias; *Test Interpretation; Test Reliability; Test Validity
IDENTIFIERS	Maximum Likelihood Estimation; Schclastic Aptitude Test

ABSTRACT

Appropriateness indexes (statistical formulas) for detecting suspiciously high cr low sccres cn aptitude tests were presented, based on a simulation of the Schclastic Aptitude Test (SAT) with 3,000 simulated scores--2,800 normal and 200 suspicious. The traditional index--marginal probability--uses a model for the normal examinee's test-taking behavior only, based cn item characteristic curve theory. The other twc indices use a generalization of the traditional index which allows ability to vary during testing. One uses the standard likelihood ratio to quantify the amount of improvement of fit achieved by permitting ability to vary across items. The cther index estimates the parameter values of the varying ability models, and uses estimated parameter values to indicate the degree of aberrance. Files cf candidates with 4%, 10%, 20%, and 40% aberrance were generated by modifying item scores of normal examinees. Results showed that 20% aberrance was surprisingly well detected for the suspiciously lcw grcup on all three indices. Suspiciously high candidates were even more easily detected. Results are significant because they suggest that inappropriately scoring candidates (such as low ability students whc cheat or high ability students who misinterpret instructions), can be detected without reference to background variables. (CP)

RESEARCH BULLETIN

RB-76-31

MEASURING THE APPROPRIATENESS OF MULTIPLE-CHOICE TEST SCORES

Michael V. Levine

and

Donald B. Rubin

This Bulletin is based on research equally supported by the College Entrance Examination Board, Educational Testing Service, and the Graduate Record Examinations Board. The points of view and opinions expressed do not represent official College Board, ETS, or GREB position or policy. The Bulletin is a draft for inter-office circulation. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
December 1976

MEASURING THE APPROPRIATENESS OF MULTIPLE-CHOICE TEST SCORES

Abstract

A student may be so atypical and unlike other students that his aptitude test score fails to be a completely appropriate measure of his relative ability. We consider the problem of using the student's pattern of multiple-choice aptitude test answers to decide whether his score is an appropriate ability measure. Several indicators of appropriateness are formulated and evaluated with a simulation of the *Scholastic Aptitude Test*.

MEASURING THE APPROPRIATENESS OF MULTIPLE-CHOICE TEST SCORES[1,2]

Multiple-choice aptitude test scores are intended to measure the relative abilities of students. But sometimes they fail. A student can be so unlike other examinees that his or her test score cannot be regarded as an appropriate ability measure. Two hypothetical examples are

> Example I (Spuriously high score): A low ability examinee copies answers to several difficult items from a much more able neighbor.
>
> Example II (Spuriously low score): A very able examinee, fluent in Spanish, but not yet fluent in English, misunderstands the wording of several relatively easy questions.

There are, of course, many other possible ways for scores to fail. We limit ourselves to cases in which a complicating process (e.g., selective copying or low English fluency) tends to produce an unusual proportion of easy items wrong and hard items right. Thus we do not expect to be able to recognize a high ability cheater who occasionally copies from another high ability examinee because he will not have many easy items wrong. Similarly, we do not expect to recognize a low ability, low fluency examinee.

Our goal is to design a practical method for using patterns of item scores to detect aberrant candidates. For this purpose we formulate appropriateness indices--statistics computed from the examinee's item scores that tend to be low when the test is an inappropriate measure of the examinee's ability and high otherwise. A very low index value opens the question of whether the test adequately measures the examinee.

An essential feature of our approach to testing problems is the use of only the test itself: Appropriateness indices are functions of the examinee's item scores.

In this paper three general types of appropriateness indices are formulated. A representative of each type is evaluated using Monte Carlo data in which most of the simulated examinees have responded according to the usual aptitude test model while a few aberrant ones have not.

It will be seen that all our indices perform quite well, at least for the test we are now using to evaluate our approach (the Scholastic Aptitude Test) and the types of aberrance we have considered. More specifically, suppose 10% of the examinees are aberrant and we consider the 5% of the examinees with the most extreme appropriateness scores. A random rule would yield 10% aberrant examinees and 90% normal in the extreme group. Using appropriateness indices, we have designed rules yielding 50% aberrant, 50% normal examinees in the extreme group.

We consider these results important because they suggest that examinees for whom a test is not appropriate can be detected without reference to additional background variables such as race, religion, gender, parents' occupation, etc. That is, they suggest there is internal evidence in the examinee's answer sheet indicating whether he or she approaches the test as do other candidates with the same ability.

THREE TYPES OF APPROPRIATENESS INDICES

In order to present the intuitions supporting our indices we return to Example I, the hypothetical low ability copier. He has an improbable pattern of responses for a low ability examinee because he has correctly answered several hard items. His pattern is also improbable for a high ability examinee because many easy items are wrong. His irregular pattern of item scores seems contrary to the customary psychometric assumption that ability is constant during testing. In fact his irregular response pattern may be much better described by a model in which ability is permitted to change somewhat during testing.

We have been investigating three basic types of indices. The reasoning leading to each will be presented now. Later a representative of each type will be formulated more precisely and evaluated.

Our simplest index type, marginal probability, uses a model for the normal examinee's test-taking behavior only. The usual model (reviewed in the next section) for the Scholastic Aptitude Test (SAT) specifies the conditional probability of an observed pattern of item responses, the probability that an examinee randomly chosen from all the examinees with a given ability produces the observed pattern of item responses. The marginal probability of a pattern is obtained by averaging over the distribution of ability in the population of examinees. The marginal probability of an aberrant examinee's pattern is expected to be relatively low because it is unlikely that a high ability person misses an easy item or a low ability person passes a hard item.

The other two index types are generalizations of the usual model that were formulated as mathematically tractable descriptions of the types of aberrance we are now studying. These models were suggested by the following reasoning. The aberrant examinee's "complicating process" leads us to expect evidence of both low ability (easy items failed) and high ability (hard items passed). In a sense soon to be made precise, the aberrant candidate behaves as if his ability were changing throughout the test. Thus we expect to obtain a much better fit of the aberrant examinee's data by using a generalization of the test model that allows ability to vary during testing.

Type II indices (likelihood ratios) use the standard likelihood ratio technique to quantify the amount of improvement of fit achieved by permitting ability to vary across items. Thus to compute a type II index both the usual model and a generalization of the usual model are fitted to the examinee's data by selecting parameter values that maximize the probability of the examinee's pattern of item responses. The ratio of the two probabilities indicates how much better the generalized model fits.

Type III indices (estimated ability variation) are obtained by estimating the parameter values of the varying ability models and using the estimated parameter values to indicate the degree of aberrance.

TEST THEORY

The observed pattern of right and wrong answers on a randomly chosen answer sheet will be treated as the outcome of a two stage experiment. In the first stage, an examinee with ability θ is sampled. In the second stage a sequence of independent dichotomous random variables $u_1, u_2, \ldots\ \ldots u_n$ is generated. These are the item scores, coded one for correct and zero for incorrect.

The usual model for the SAT is primarily concerned with the relation between ability and item scores. According to this model the conditional probability that u_i is one is a continuous, increasing function of ability, $P_i(\theta)$, called the *item characteristic function*. The conditional probability that a randomly selected examinee with ability θ produces the pattern of right and wrong answers corresponding to the vector of item responses $U = < u_1, \ldots u_i, \ldots u_n >$ is then

$$(1) \qquad f(U|\theta) = \prod_{i=1}^{n} P_i(\theta)^{u_i}[1 - P_i(\theta)]^{1-u_i} \quad .$$

For a discussion of item characteristic curve theory see Birnbaum (1968).

In this work each item characteristic function is assumed to have the "logistic" functional form

$$(2) \qquad \begin{array}{l} P_i(\theta) = c_i + (1 - c_i)\{1 + e^{-a_i(\theta - b_i)}\}^{-1} \\ 0 < a_i \quad , \quad -\infty < b_i < \infty \quad , \quad 0 \leq c_i < 1 \quad . \end{array}$$

This functional form is used regularly with multiple-choice aptitude tests. For evidence supporting its adequacy for the tests and population we wish to study, see Lord (1968) and Levine and Saxe (1976).

This basic model, in which examinees differ only in ability, will be called the standard model of item characteristic curve theory. Various generalizations will be used to describe aberrant examinees. The major one used in this paper is the Gaussian model in which we assume that a new ability Θ_i is sampled for each item. Thus the probability that the i -th item is correct becomes $P_i(\Theta_i)$ instead of $P_i(\Theta)$. In the Gaussian model, "item abilities" Θ_i are assumed to be independent normal random variables with mean Θ_o and variance σ^2 .

In the first stage of the standard model, an examinee with ability Θ is sampled. In the first stage of the Gaussian model, on the other hand, an examinee with "central ability" Θ_o and "ability variance" σ^2 is sampled. Thus the Gaussian model can accommodate two kinds of differences between examinees. The standard model can be seen as the limiting case of the Gaussian model with the ability variance σ^2 equal to zero.

The generalization of the conditional probability (1) used to define the standard model becomes

$$(3) \qquad f(U|\theta_o,\sigma^2) = \int\cdots\int \prod_{i=1}^{n} P_i(\theta_i)^{u_i} Q_i(\theta_i)^{1-u_i} \phi[(\theta_i - \theta_o)/\sigma]d\theta_1\cdots d\theta_n$$

$$= \prod_i \int P_i(t)^{u_i} Q_i(t)^{1-u_i} \phi[(t - \theta_o)/\sigma]dt$$

where $\phi(x)$ is the Gaussian density $(2\pi)^{-1/2}e^{-x^2/2}$.

In the discussion section we will wish to refer to other generalizations of the standard model. Like the Gaussian and standard model, each uses a vector of parameters Θ to characterize the examinee and assumes that a new ability θ_i is independently sampled for each item. The models differ in the specification of the distribution of the θ_i and are defined by a formula of form

$$(4) \qquad f(U|\Theta) = \prod_i \int P_i(t)^{u_i} Q_i(t)^{1-u_i} dF_{\Theta}(t)$$

where the definition of Θ differs from model to model. For example, we have the standard model with $\Theta = < \theta >$ and all the $\theta_i = \theta$, the Gaussian model with

$$\Theta = < \theta_o,\sigma^2 > \quad , \quad \theta_i \sim \underline{N}(\theta_o,\sigma^2) \quad .$$

And finally, as a limiting case, we have the unconstrained model in which the θ_i may be any value and

$$\Theta = < \theta_1,\theta_2,\ldots \theta_i,\ldots \theta_n > \quad \text{where} \quad -\infty < \theta_i < \infty \quad .$$

THE INDICES

Type I: Marginal Probabilities

If the (generally unknown) density for the θ 's is specified and denoted by g , then the formula

$$(5) \qquad \int_{-\infty}^{\infty} f(U^*|\theta)g(\theta)d\theta$$

can be used to obtain the marginal probability of a vector of item scores U^* . The standard model specifies a particular formula for the conditional probability $f(U^*|\theta)$. Our different marginal probability indices specify different ability densities $g(\theta)$.

The density $g(\theta)$ summarizes our information about a sampled examinee's ability before scoring the test. Suppose we choose to ignore that information and base our ability estimate only on the examinee's test performance. Mathematically this can be expressed by replacing $g(\theta)$ by a density $\tilde{g}(\theta)$ with a very small variance and centered about $\hat{\theta}$, the maximum likelihood estimate of ability obtained by maximizing $f(U^*|\theta)$. As the variance of $\tilde{g}(\theta)$ tends to zero, $\int f(U^*|\theta)\tilde{g}(\theta)d\theta$ converges to $f(U^*|\hat{\theta})$. The logarithm of the maximum

$$\ell_0(U^*) = \log f(U^*|\hat{\theta})$$

is our representative type I index. We use it basically because it is straightforward to calculate and works well, not because we believe the single point distribution for $g(\theta)$ is reasonable.

Other type I (marginal probability) indices can be obtained by estimating the ability distribution $g(\theta)$ from the observed $\hat{\theta}$ distribution or by true score methods (Lord, 1970). The integration required to compute (5) can be intractable. A more easily computed type I index begins with the observation that the function of θ, $\log f(U^*|\theta)$, is ordinarily unimodal and roughly symmetric about $\theta = \hat{\theta}$. This suggests the second order approximation of $\log f(U^*|\theta)$

$$\ell_0 + \frac{1}{2}(\theta - \hat{\theta})^2 \ell_2$$

where ℓ_2 is the second derivative of log $f(U^*|\theta)$ evaluated at $\theta = \hat{\theta}$. If the ability density is given by the unit normal density, we then obtain the approximation of marginal probability

$$\frac{1}{\sqrt{2\pi}} \int e^{\ell_0} e^{\frac{1}{2}(\theta-\hat{\theta})^2 \ell_2} e^{-\frac{1}{2}\theta^2} d\theta$$

$$= e^{\ell_0} e^{\frac{1}{2}\hat{\theta}^2 \frac{\ell_2}{1-\ell_2}} (1-\ell_2)^{-\frac{1}{2}}$$

or equivalently

$$\ell_0 + \frac{1}{2}\hat{\theta}^2 \left(\frac{\ell_2}{1-\ell_2}\right) - \frac{1}{2}\log(1-\ell_2) \quad .$$

Type II: Likelihood Ratios

In order to use a likelihood ratio as an index of aberrance, we first maximize $f(U^*|\theta)$ given in formula (5) over θ. In logarithmic form, the likelihood ratio index is

$$\max_{\Theta} \log f(U^*|\Theta) - \ell_0 \quad .$$

Our representative of this type of index is obtained from the Gaussian model, where $f(U^*|\Theta) = f(U^*|\Theta_o,\sigma^2)$ as given in formula (4).

Type III: Degree of Aberrance Estimate

Our best index of this type was obtained from the Gaussian model by maximizing the probability $f(U^*|\Theta_o,\sigma^2)$. The index $\hat{\sigma}$ is the square root of the maximum likelihood estimate of the ability variance.

THE SIMULATION

The indices were evaluated with a simulation of the Scholastic Aptitude Test using Hambleton and Rovenelli's (1973) programs. To simulate a "normal" candidate, first an ability Θ was sampled from a normal, zero mean, unit variance population. Then the item scores for the examinee were simulated as a sequence of independent Bernoulli trials. The success probability on the i -th trial is $P_i(\Theta)$ as in formula (1) where the parameters a_i , b_i , c_i in the formula were obtained from Lord's (1968) fitting of an SAT-V administration.

Examinees with varying degrees of aberrance were generated by modifying the item scores of normal examinees. To simulate a spuriously high examinee cheating on, say, 20% of the test, first a normal examinee was simulated. Then 20% of the items were sampled without replacement. The sampled items were then scored correct whether they previously were correct or not. In this way files of candidates with 4%, 10%, 20%, and 40% aberrance were generated.

To generate a spuriously low examinee forced to guess on, say, 20% of the test we again begin by generating a normal examinee and sampling 20% of the items. Since the simulated test is a five-alternative multiple-choice test, we rescore the item as correct with probability 1/5 and incorrect with probability 4/5. In this way files of spuriously low-scoring candidates having 4%, 10%, 20%, and 40% aberrance were generated.

See Appendix I for details of the simulation and methods for finding maximum likelihood estimates. See the discussion section for comments on the test model and the modelling of aberrance.

RESULTS

The analogy between an observer in a psychophysics experiment trying to detect a faint signal and our problem of trying to detect aberrant candidates from equivocal patterns of item scores led us to use ROC curves (Green and Swets, 1966) for evaluating indices. To compute an empirical ROC curve for an index, say for concreteness ℓ_0 , and a given group of aberrant examinees, the index is evaluated for a sample of normal and aberrant examinees. The sampled examinees are then ordered from lowest to highest appropriateness score. The empirical ROC curve is the set of points $< x(t),y(t) >$ where

$x(t)$ = the proportion of normal examinees with $\ell_0 \leq t$,

$y(t)$ = the proportion of aberrant examinees with $\ell_0 \leq t$.

A random rule or a rule based on a poor appropriateness index will give an ROC curve close to the diagonal $x = y$. A good appropriateness index gives a curve well above the diagonal. The empirical curve provides an estimate of the probability that normal candidates will be incorrectly classified by a rule sufficiently stringent to detect a given percent of a particular kind of aberrant examinee. For example, suppose we choose t so that 5% of the population is classified as aberrant. Further suppose that 10% of the population is aberrant. Then the intersection of the curve with the line $.9x + .1y = .05$ gives the proportion of aberrant examinees correctly identified and normal examinees misclassified.

In Figure 1, marginal probability (ℓ_0) ROC curves are given for the various spuriously low groups. Each curve is based on 3,000 examinees: 200 examinees with the same percent aberrance and 2800 normal candidates. The same normal examinees are used for all ROC curves in this and the other figures.

Insert Figure 1 about here

Only the lower parts of the curves are relevant to our immediate purpose since a rule improperly classifying more than 30% of the normal candidates is not likely to be used in aptitude testing. The curves show that 20% aberrance is surprisingly well detected. They also show that marginal probability does only slightly better than chance for 4% aberrance. The expected net change in total test score for 4% aberrance turns out to be very small, although an occasional very bright and very unlucky candidate may be detected.

Figures 2 and 3 give ROC curves for the likelihood ratio test and the degree of aberrance index. These curves show the same pattern as the Figure 1 curves, at least over the lower part of the curves.

Insert Figures 2 and 3 about here

Figures 4, 5, 6 give the corresponding ROC curves for the spuriously high group. It can be seen that spuriously high aberrant candidates are more easily detected than spuriously low candidates. This is to be expected since the process generating spuriously low candidates necessarily contains a random component lacking in the spuriously high process. The spuriously low candidate is forced to guess, but the spuriously high candidate "knows" the right answer. Simulating high spuriousness typically results in changing more item scores than simulating low spuriousness.

Insert Figures 4, 5, and 6 about here

We recomputed the likelihood ratio ROC curve for the 20% spuriously low group using only those candidates with more than 10% of the item scores actually changed. The resulting curve, computed from 102 examinees, (Figure 7) appears comparable to the spuriously high curves.

Insert Figure 7 about here

The curious crossover in Figure 4 arises because according to the standard model the probability that a very able examinee answers all items correctly is nearly one. Thus if we begin with an able candidate with item score vector U^* and sample 40% of his items and make them correct, we obtain a new vector U^{**} which may have all or all but a few very hard items right. When this happens the probability $e^{\ell_0(U^{**})}$ will be very nearly one and frequently larger than $e^{\ell_0(U^*)}$. The larger the proportion of sampled items the more frequently $\ell_0(U^{**})$ will be abnormally large. In fact for some large proportion of sampled items, the ℓ_0 ROC curve should pass, as observed, beneath the diagonal.

Since rules that improperly classify large numbers of normal candidates cannot be used, the observed anomaly is inconsequential. Furthermore, it does not appear with the likelihood ratio test. This is probably attributable to the fact that the increment in $\ell_0(U^{**})$ is accompanied by a comparable increment in $\ell_n(U^{**})$, the likelihood under the Gaussian model.

DISCUSSION

We consider our work important because it demonstrates that in at least some cases there is internal evidence in an examinee's answer sheet for the appropriateness of a test. We do not, however, feel committed to our present indices or aberrance models. We might have just as well worked with the posterior mean of σ^2 from the Gaussian model as an aberrance index or an aberrance model in which the examinee fluctuates between two abilities. For example, there is the aberrance model in which the examinee has constant probability p of cheating on an item and performing as if he has infinite ability defined by the equation

$$(6) \qquad f(U| < p,\theta >) = \prod_i [(1 - p)P_i(\theta) + p]^{u_i}[(1 - p)Q_i(\theta)]^{1-u_i} \quad ,$$

$$0 \leq p \leq 1 \quad .$$

The observation that item characteristic curve theory--with its local independence assumption--may be too rudimentary to provide an adequate description of the stochastic structure of the SAT is by no means fatal to our main point, the point that answer sheets contain internal evidence of aberrance. In fact it can be argued that departures from a more specific model could be more easily detected.

In addition to studying other indicators and types of aberrance we feel that the following questions should be explored:

1. What is the effect (on aberrance indices) of using estimated item parameters?

2. What is the effect of estimating item parameters from samples containing aberrant examinees?

3. Can omitted and not reached items be used to increase the power of aberrance indices?

4. Can the interrelations between various items and subtests be incorporated in the test model and used to detect aberrance?

5. Do aberrance indices indentify a relatively large proportion of examinees in samples of candidates speaking English as a second language, in samples of candidates with moderately high test scores but very low socioeconomic status, in samples of known cheaters?

These questions form a rich and fertile area for future research.

References

Birnbaum, A. Some latent trait models and their use in infer an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968. Chapters 17-20.

Green, D. M. and Swets, J. A. *Signal detection theory and psychophysics*. New York: Wiley, 1966.

Hambleton, R. K. and Rovenelli, R. A. A FORTRAN IV program for generating response data from logistic test models. *Behavioral Science*, 1973, *18*, 74.

Levine, M. V. and Saxe, D. The use of periodic functions to measure the difficulty of aptitude test items. Research Bulletin 76-17. Princeton, N.J.: Educational Testing Service, 1976.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 1968, *28*, 989-1020.

Lord, F. M. Item characteristic curves estimated without knowledge of their mathematical form--a confrontation of Birnbaum's logistic model. *Psychometrika*, 1970, *35*, 43-50.

Footnotes

[1] This research has been partially supported by grants from the Graduate Record Examination Board and the College Entrance Examination Board. The points of view and opinions expressed do not represent official ETS, GREB or CEEB position or policy.

[2] We are grateful to Frederic M. Lord and Ingram Olkin for technical advice and for comments on an earlier draft of this paper and to Dorothy Thayer for the computations and a critical review of other aspects of the work.

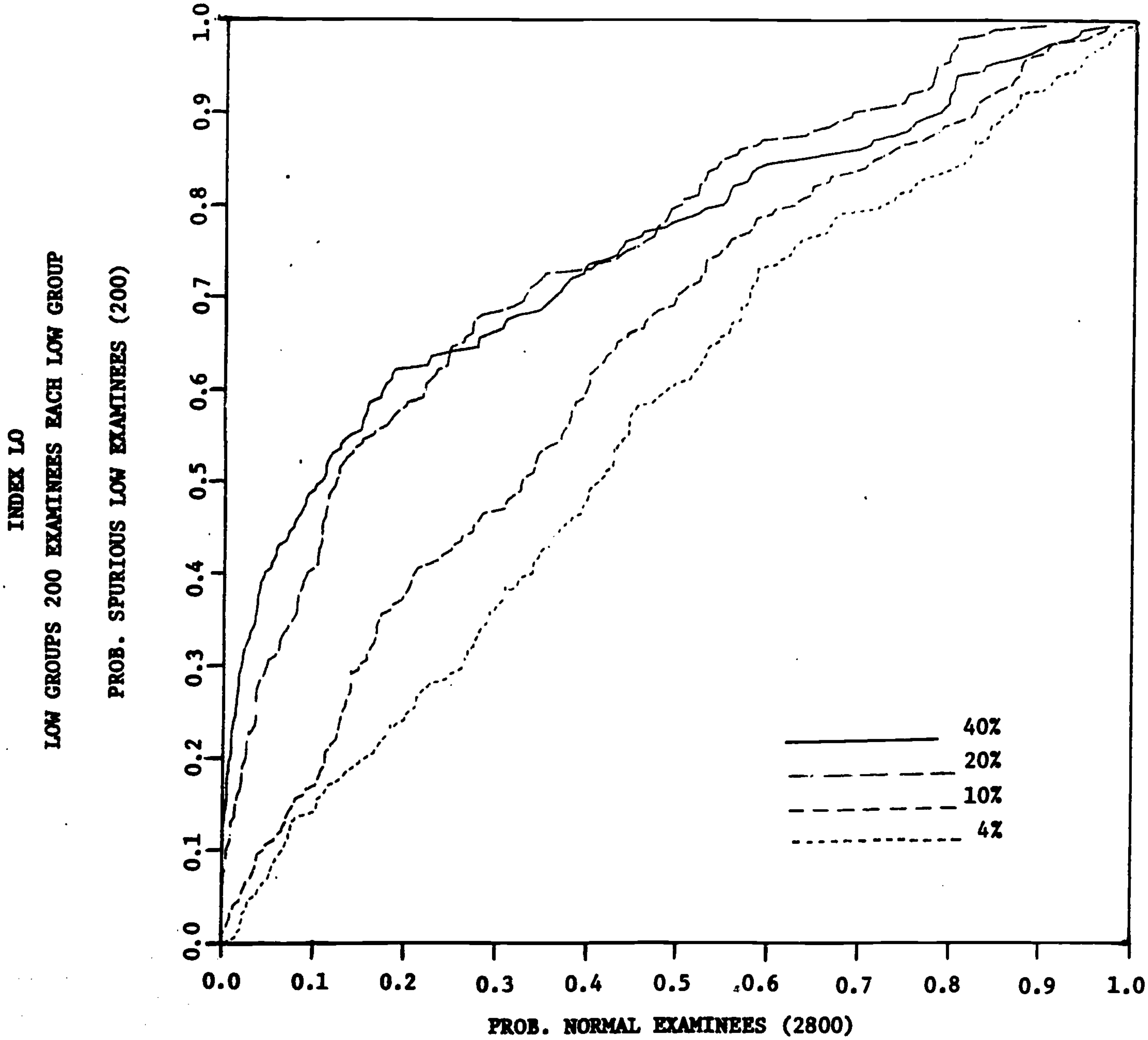

INDEX LO
LOW GROUPS 200 EXAMINEES EACH LOW GROUP
PROB. SPURIOUS LOW EXAMINEES (200)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
40%
20%
10%
4%
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
PROB. NORMAL EXAMINEES (2800)

FIGURE 1

INDEX LN-LO METHOD 1

LOW GROUPS 200 EXAMINEES EACH LOW GROUP

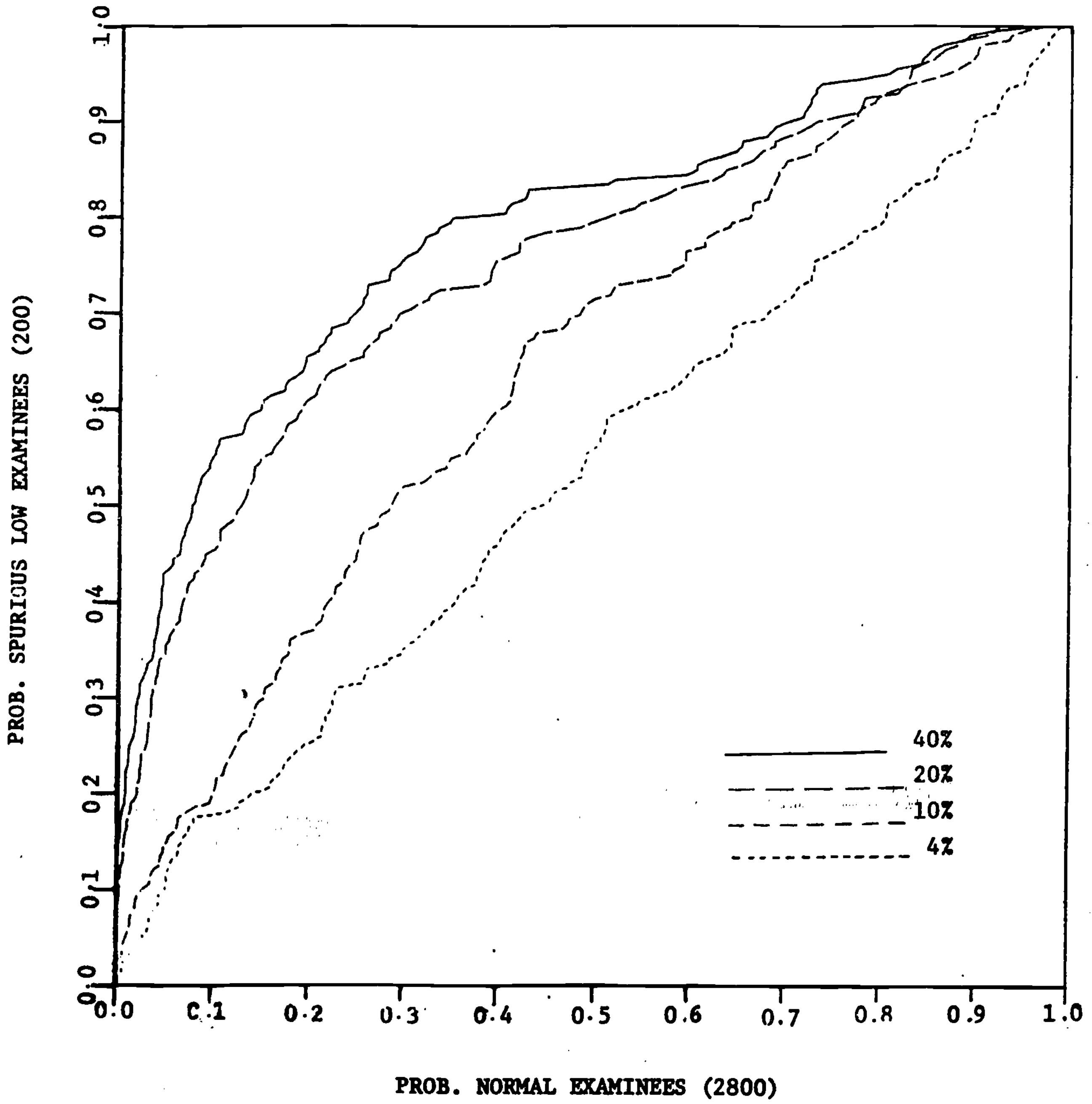


FIGURE 2

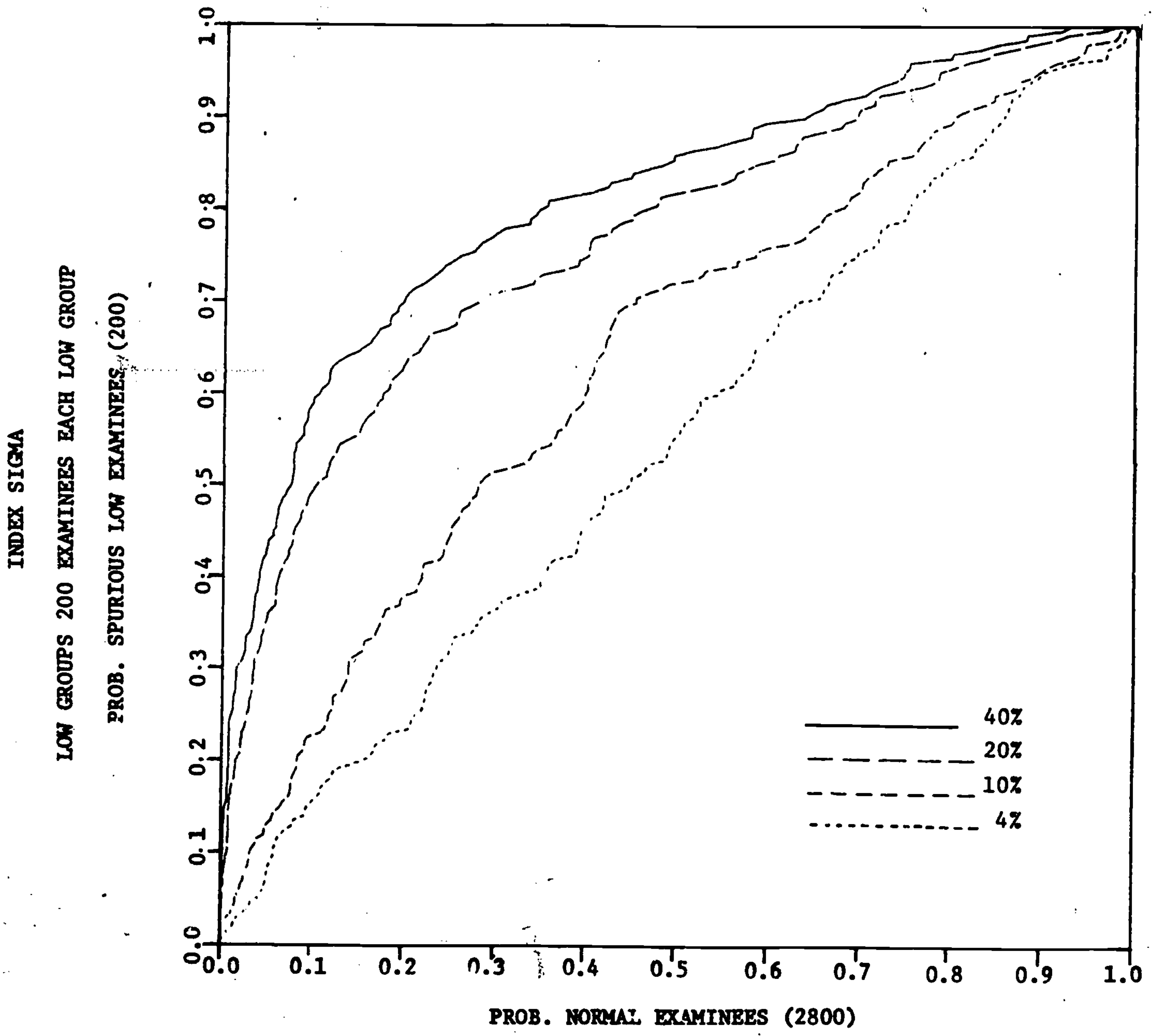

INDEX SIGMA
LOW GROUPS 200 EXAMINEES EACH LOW GROUP
PROB. SPURIOUS LOW EXAMINEES (200)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
40%
20%
10%
4%
PROB. NORMAL EXAMINEES (2800)

FIGURE 3

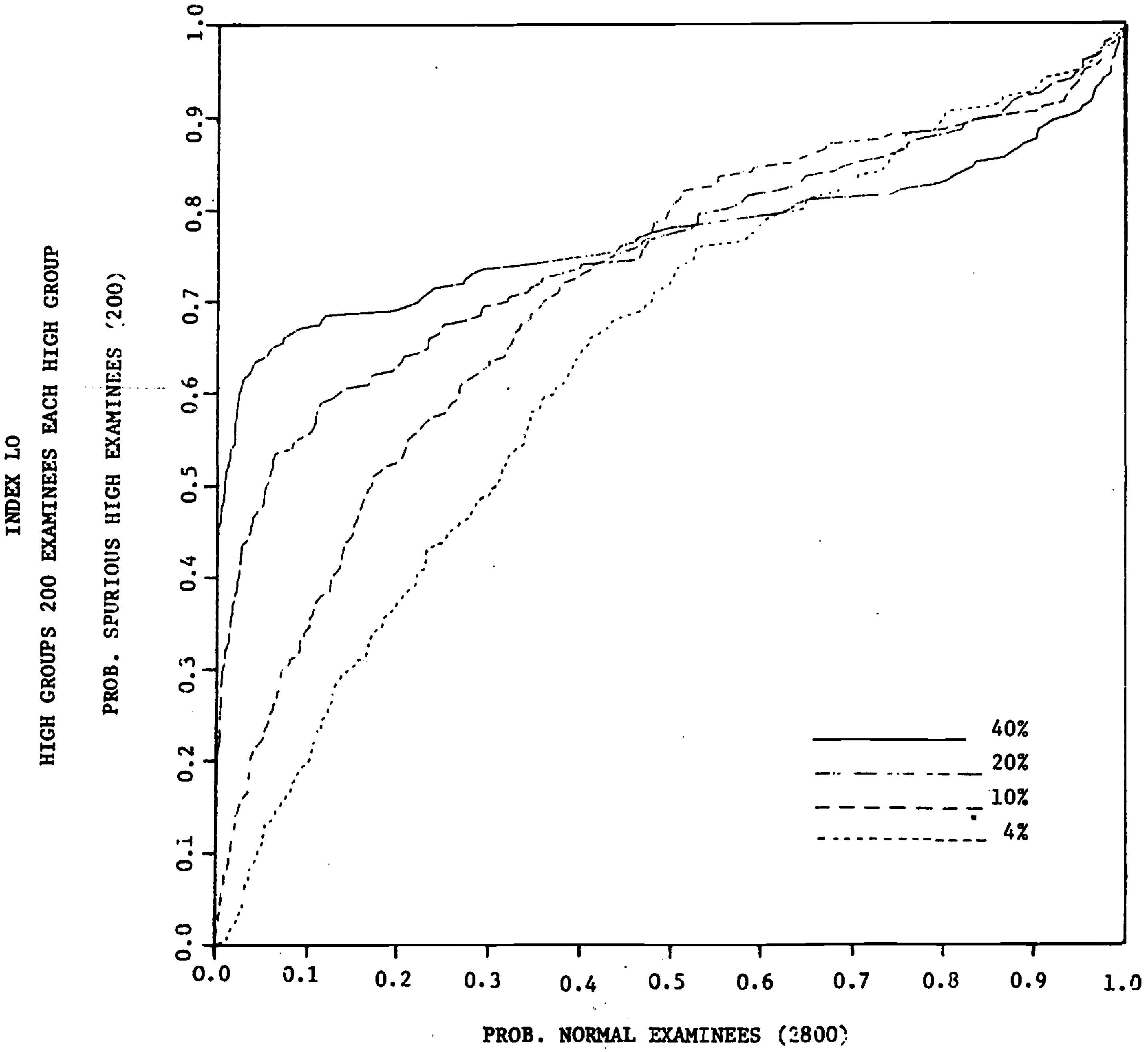

INDEX LO
HIGH GROUPS 200 EXAMINEES EACH HIGH GROUP
PROB. SPURIOUS HIGH EXAMINEES (200)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
PROB. NORMAL EXAMINEES (2800)
40%
20%
10%
4%

FIGURE 4

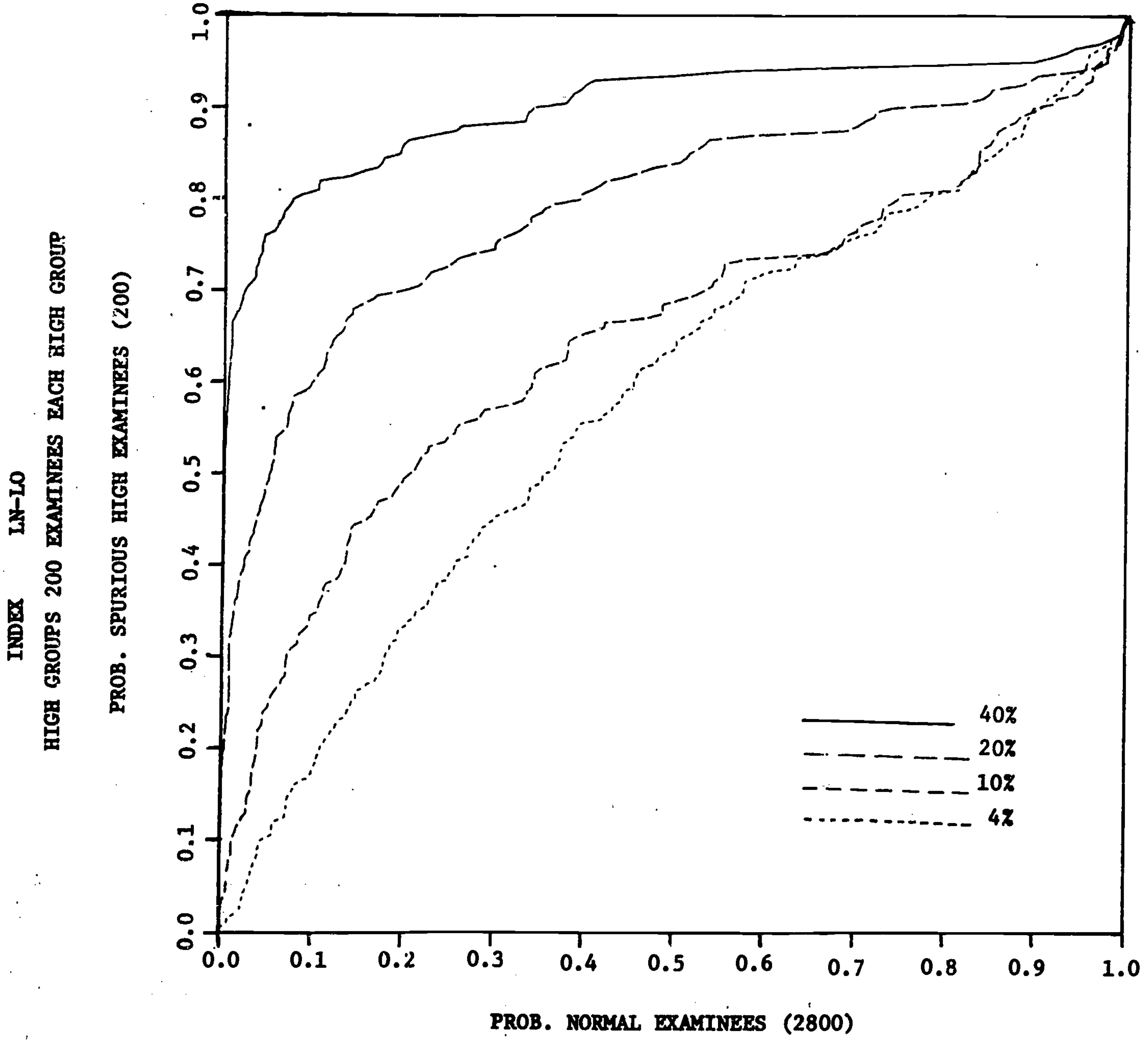

INDEX LN-LO
HIGH GROUPS 200 EXAMINEES EACH HIGH GROUP
PROB. SPURIOUS HIGH EXAMINEES (200)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
40%
20%
10%
4%
PROB. NORMAL EXAMINEES (2800)

FIGURE 5

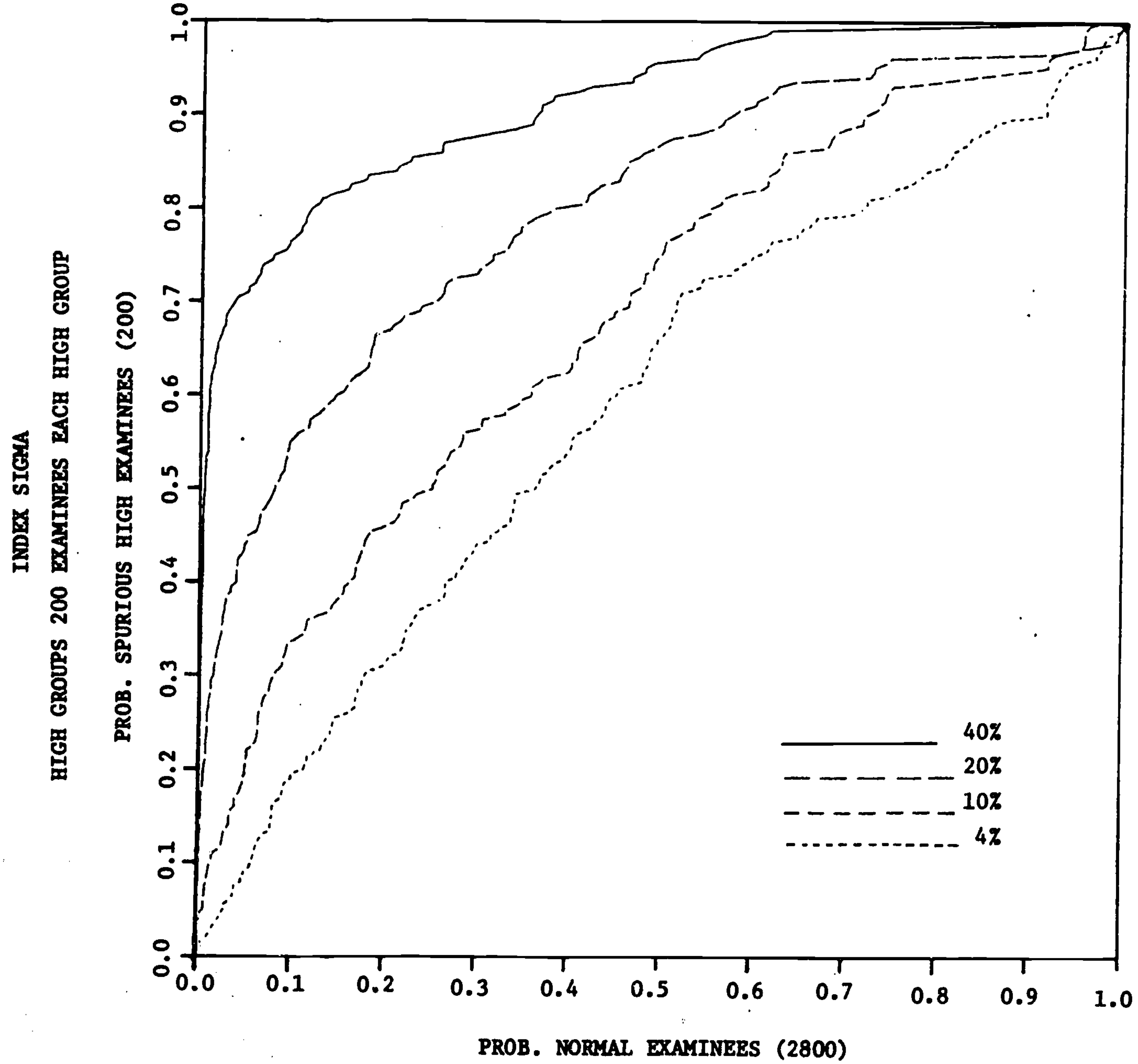
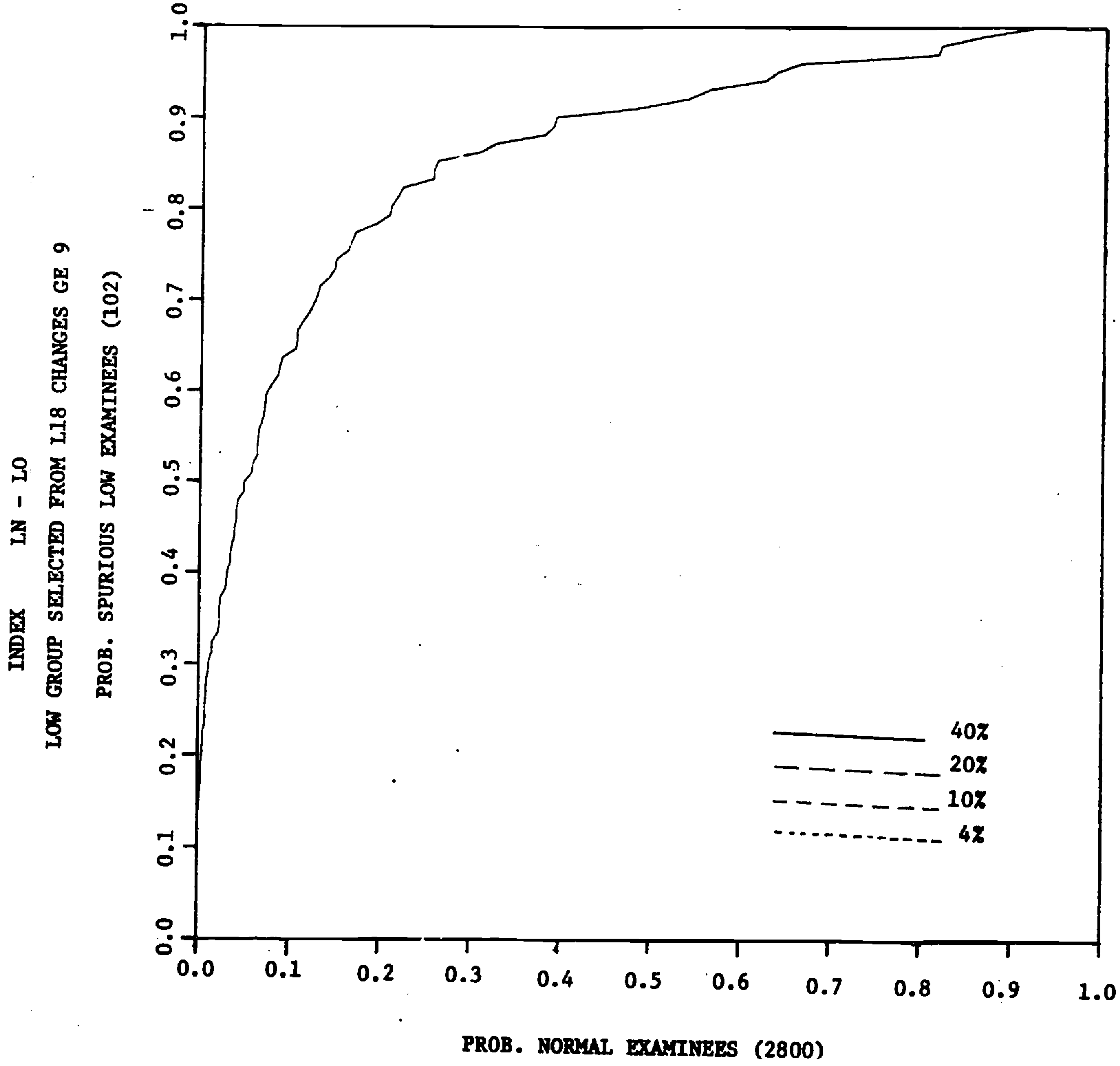

INDEX SIGMA
HIGH GROUPS 200 EXAMINEES EACH HIGH GROUP
PROB. SPURIOUS HIGH EXAMINEES (200)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
40%
20%
10%
4%
PROB. NORMAL EXAMINEES (2800)

FIGURE 6


INDEX LN - LO
LOW GROUP SELECTED FROM L18 CHANGES GE 9
PROB. SPURIOUS LOW EXAMINEES (102)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
PROB. NORMAL EXAMINEES (2800)
40%
20%
10%
4%

Appendix

Technical details on the computations are collected and listed below:

1. During the simulation of normal examinees a Tausworthe generator (Whittlesey, 1968) was used to generate item scores. To obtain Gaussian distributed abilities Pike's (1965) algorithm was applied to numbers obtained from the Tausworthe generator.
2. During the simulation aberrant examinees Learmonth and Lewis's (1973) algorithm was used to generate numbers uniformly distributed on the unit interval. To sample a proportion of items without replacement, 1 + (number of items) x (uniformly distributed number) was truncated to obtain an integer. This process was repeated (with new uniformly distributed numbers) until the desired number of items was selected. The uniformly distributed numbers were also used to modify the item scores of the sampled items for the spuriously low scoring aberrant candidates. A sample item was scored "correct" if a uniformly distributed number was $\leq .2$.
3. To compute L_0 , θ was first estimated with LOGIST (Wood, Wingersky and Lord, 1976). Estimated θ 's less than -5 were set equal to -5.
4. To compute L_n and σ , the steepest descent method in Gruvaeus and Jöreskog (1970) was used to maximize the likelihood function for the Gaussian model. The starting point was θ = LOGIST estimated θ and $\sigma = .1$, Only the steepest descent

References

Gruvaeus, G. T. and Jöreskog, K. G. A computer program for minimizing a function of several variables. Research Bulletin 70-14. Princeton, N.J.: Educational Testing Service, 1970.

Learmonth, G. P. and Lewis, P. A. W. Random number generator package LLRANDOM. Report no. NPS55LW73061A. Monterey, Calif.: Naval Postgraduate School, 1973.

Pike, M. C. Algorithm 267. Random normal deviate. Communications of the ACM, 1965, 8, 606.

Whittlesey, John R. B. A comparison of the correlational behavior of random number generators for the IBM 360. Communications of the ACM, 1968, 11, 641-644.

Wood, R L, Wingersky, M. S., and Lord, F. M. LOGIST - A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, N.J.: Educational Testing Service, 1976.