AN APPLICATION OF CONFIDENCE INTERVALS
AND OF MAXIMUM LIKELIHOOD TO THE
ESTIMATION OF AN EXAMINEE'S ABILITY*

FREDERIC M. LORD
EDUCATIONAL TESTING SERVICE

A mathematical definition of the theoretical relation between the examinee's actual responses to the test items and his "true ability" is selected. A maximum-likelihood solution is obtained for estimating the examinee's "true ability" from his responses to the items. The standard error of the maximum-likelihood estimate is obtained, its relation to the discriminating power of the test is pointed out, and some generalizations are drawn as to the optimum level of item difficulty. The Neyman-Pearson power function is applied to determine which of two psychological tests is the most powerful for the selection of "successful" examinees.

When we use the usual type of mental test score to measure the ability of the examinees in a group, the metric supplied by the test scores cannot be considered satisfactory. The inadequate nature of this metric is apparent when we consider that two tests measuring the same ability, administered to the same group of examinees, may yield two score distributions of entirely different shapes. The metric of "true" scores obtained from very long tests is subject to this objection as is the metric of fallible scores obtained from short tests. We here propose to use a more adequate metric for measuring the ability underlying the test score—a metric that will remain invariant from test to test—and to investigate what may be learned from a maximum-likelihood approach to the problem of estimating the examinee's ability, as defined by this metric, and from certain other related approaches of modern statistical theory. The reader is warned that, in view of the heavy (but not insuperable) computational difficulties in the way of any practical application, the present discussion is directed chiefly towards determining what conclusions of general theoretical significance can be drawn from a consideration of the proposed metric.

If the response of an examinee to each of \( n \) test items may be scored either 0 or 1, we may denote the score of examinee \( a \) on item \( i \) by \( x_{ia} (i = 1, 2, \ldots, n; a = 1, 2, \ldots, m) \), where \( x_{ia} \) equals either 0 or 1. For any effective test, the logic of the practical situation implies some relation, in general, between the probability that \( x_{ia} = 1 \), which we will denote by \( \text{Prob} (x_{ia} = 1) \), and \( \text{Prob} (x_{ja} = 1) \), where \( j \) denotes some item other than item \( i \); and also some

*The author is indebted to Dr. John W. Tukey for helpful comments on a draft of the present manuscript.
relation between \( \text{Prob} (x_{ia} = 1) \) and \( \text{Prob} (x_{ib} = 1) \), where \( b \) denotes some examinee other than examinee \( a \). If these relations can be suitably specified mathematically, we can apply Fisher's method of maximum likelihood \((25, 133-142; 16, 152-161)\) so as to obtain from the data on any actual set of examinees' answer sheets maximum-likelihood estimates of the parameters describing the test items and of the parameter describing the ability of each individual examinee.

Once the parameters to be estimated have been satisfactorily specified and maximum-likelihood estimates for these parameters derived, a large body of standard mathematical theory can be brought to bear on many unresolved problems in testing. The maximum-likelihood estimate of the examinee's ability itself constitutes an answer to the question of how the items should be weighted in obtaining the examinee's total score. The discriminating power of the test at different ability levels can then, in large sample theory, be measured by the usual standard error of the maximum-likelihood estimate. The examinee's responses to the test items may be used to set up a confidence interval \((16, \text{Ch. 11})\) within which the examinee's true ability may be assumed to lie. The length of this confidence interval provides a measure of the test's discriminating power at a given level of actual test score. If it is desired to build a test that will have maximum discriminating power at a given cutting score, the problem of the optimum distribution of item difficulties in such a test can be reduced, in large-sample theory, to a question of determining for what values of the item difficulties the standard error of the test score is a minimum. Finally, if it is desired to test some hypothesis—for example that a given examinee's true ability is above rather than below a given value—the Neyman-Pearson theory of testing hypotheses \((25, 152 \text{ ff; 16, Ch. 12})\) can be brought to bear; for example, the power function of any test score used to check this hypothesis can be determined and compared with the power functions of scores on other psychological tests in order to determine what sort of psychological test would be best for this purpose.

\section*{I. The Case of Free-Response Items}

We will here assume, as have Guilford \((6)\), Richardson \((19)\), Mosier \((17, 18)\), Ferguson \((3)\), Lawley \((9, 10)\), Lorr \((15)\), Tucker \((23, 24)\), Lord \((13, 14)\), Cronbach, and Warrington \((2)\), and others, that the probability \( P_{ia} \) that an examinee will answer an item correctly is a normal ogive function of his "true ability" \( c \) in the area measured by the test:

\[ P_{ia} = \int_{h_{i; \frac{R_{iz}}{N_i}}}^{m} N(c) \, dc. \]  

(1)

Here \( c_a \) is a population parameter measuring the true ability of examinee \( a \);