Using estimates of item ease and item discrimination, procedures are provided for computing estimates of the reliability and percentage of failing scores for tests assembled from these items. Two assumptions are made: that the average item coefficient will be approximately equal to the average of the estimated coefficients and that the score distribution for the test will be approximately normal. The predicted mean test score is equal to the sum of item ease coefficients, and the predicted variance is equal to the square of the sum of item discrimination indices divided by 4.5. The fail point is always sixty percent of the number of items. Normal curve tables are used to estimate the percent of score below the fail point. A normal curve probability table is provided, as is a computer program in BASIC for using this method. The results of using this procedure with seventeen tests are presented and compared with the obtained score distribution statistics for samples of either 51 cases or 201 cases. Kuder-Richardson formula 21 reliability coefficients may also be obtained from this procedure. (CTM)
Extension Course
Institute
UNITED STATES AIR FORCE
-GUNTER AFS, ALABAMA

Professional Education is Paramount

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY
Jergil M. Smith"

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC) AND
USERS OF THE ERIC SYSTEM."
ESTIMATING EXAMINATION FAILURE RATES AND RELIABILITY PRIOR TO ADMINISTRATION

by

Vergil M. McIntosh

October 1975

Research and Evaluation Division, Extension Course Institute, Air University
The USAF Extension Course Institute, with hundreds of courses and thousands of examinations, is in an excellent position to apply sophisticated techniques in its evaluation program. One such technique is described here -- a program to estimate failure rates and reliability prior to test administration.

Since the field-testing and refinement of so many instruments is a luxury beyond our means, predictive measures of difficulty and reliability are necessary tools of test development and evaluation. Mr. Vergil McIntosh, of the ECI Evaluation and Research Division, has developed predictive measures that meet our needs admirably in this area.

This report on the programs he has developed has been published in the thought that other educational institutions, both military and civilian, can benefit from our findings. The comments of users would be appreciated.

HAROLD MARKOWITZ, JR., LtCol, USAF
Chief, Evaluation and Research Division
TABLE OF CONTENTS

SECTION                      PAGE
A  Introduction              1
B  Procedures
   Statistical Formulas      1
   Manual Computations
C  Computer Program for Computing Estimates     6
   ISE 2 Program for Computing Estimates
   Example of a Statistical Report
D  Conclusions
   Findings
   Summary

LIST OF FIGURES

FIGURE                      PAGE
1  Worksheet for Computing Estimates Manually 4
2  Normal Curve Probability Table           5
3  ISE Program to Compute Examination        7
   Statistical Estimates
4  Printout of a Statistical Report          8
Section A - Introduction

Problem:

Because of the requirement to place examinations in use before pre-testing, the Extension Course Institute (ECI) sometimes finds that examinations are too difficult or their reliability is not high enough to be acceptable. Therefore, a procedure is needed to accurately predict these test statistics before the test is activated.

To meet these needs a system has been devised and evaluated to estimate test statistics by making an estimate of the ease and discrimination index for each item. The procedure was tried, compared with actual statistical analyses, and found, in nearly all cases, to give close approximations.

The procedure was first computed manually using a worksheet and a normal curve probability table. A computer program was later developed which makes the computations and prints out a report in approximately one minute. Both the manual and the computer procedures are described in the following sections.

Section B - Procedures

Statistical Formulas:

In order to follow the rationale for the procedure it is necessary to consider the statistical formulas involved in the present statistical analysis of examinations. These formulas are:

Reliability - Kuder-Richardson Formula 21.

\[
R = \frac{n\sigma^2 - r M(n-M)}{\sigma^2(n-1)}
\]

Where: \( n \) = the number of items on the examination; \( \sigma \) = the standard deviation of scores; \( M \) = mean of examination scores.

1. Internal standards define an unacceptable examination as one having a failure rate in excess of 35% and/or a reliability coefficient of less than .75.
Standard Deviation ($\sigma$) = $\sqrt{\frac{\sum x^2}{n}}$

where: $x =$ any deviation from the mean; $x^2 =$ sum of the squared deviations; $N =$ number of cases.

Since we do not have all of the data available to substitute in the above formulas until a sample of student solutions has been received, it is obvious that we must make some estimates.

Ebel gives a formula which can be used to estimate the variance of the scores on a test. It is expressed as:

$$\sigma^2 = \frac{(\sum D)^2}{6}$$

where: $D =$ the sum of the indices of discrimination for a test.

In using this formula to predict the variance of a sample of ECI tests it was found that variance can be predicted best by using a divisor of about 4.5. The reason for this difference is not known, but Ebel may have used a different formula for computing discrimination indexes.

In order to estimate the failure rate, it is necessary to compute the area under the normal probability curve falling below the fail score. This can be computed by determining the difference in standard deviations between the mean and the fail point by the formula:

$$SD\ \text{diff} = \frac{x}{\sigma}$$

where: $x =$ the difference in score units between the mean and the fail score; and $\sigma =$ the standard deviation of the scores. By referring to a table of the fractional parts of the area under the normal probability curve, the percent of scores falling between the mean and fail point can be determined (e.g. Table A p. 458 in Garrett Statistics in Psychology and Education).

Subtracting this value from 50 percent results in the percent of estimated failures for the examination. This, of course assumes student scores approximate a normal distribution. In using this procedure with a group of ECI courses, it was found that the predictions were generally close to the actual failure rate.

Manual Computations:

The steps in estimating the examination statistics are as follows:

STEP 1: **ESTIMATE THE EASE INDEX AND DISCRIMINATION INDEX FOR EACH ITEM IN THE ITEM BANK.** This step is done by the test constructor as he checks the item pool. If the items have been used on previous examinations, the item analyses statistics can provide a good basis for estimating the expected performance of each item. Estimates for individual items may not have a high degree of accuracy; however, when averages for all items are computed, the estimated and actual performance ought not differ greatly. This generalization is drawn from the known fact that a number of estimates when averaged will be very close to the true value. This step can be refined and the accuracy improved through (a) preparing guidelines for making estimates, (b) collecting and analyzing data on estimates, and (c) holding in-service training on making estimates for test constructors.

STEP 2: **SELECT ITEMS FOR THE TWO PARALLEL COURSE EXAMINATIONS (CE) FORMS AND COMPUTE THE AVERAGES OF ITEM DISCRIMINATION INDEXES AND THE ITEM EASE INDEXES.** A worksheet (see figure 1) has been devised to assist in making the computations.

STEP 3: **COMPUTE THE VARIANCE (\(\sigma^2\)) AND STANDARD DEVIATION (\(\sigma\)).** See page 2.

STEP 4: **COMPUTE THE MEAN (M) OF THE RAW SCORES.** M equals the number of items on the examination times the average item ease.

STEP 5: **COMPUTE THE FAIL POINT.** Fail point = .60 x the number of items on the examination.4

STEP 6: **SUBTRACT THE FAIL POINT FROM THE MEAN AND DIVIDE THE DIFFERENCE BY THE STANDARD DEVIATION.** This gives the difference in terms of standard deviation units.

---

4. Internal standards mandate this fail point which is based on Air Training Command resident school standards.
WORKSHEET

for estimating test failure rates and reliability

A. Number of items on the examination
B. Sum of Discrimination Indexes
C. Mean of Discrimination Indexes
D. Sum of Item Base Indexes
E. Mean of Base Indexes
F. Mean of Raw Scores
G. Fail Score
H. Estimated Variance
I. Estimated Standard Deviation
J. Difference between Mean and Fail Score
K. Difference "J" in terms of Standard Deviations
L. Percent of Scores between Mean and Fail Point
M. Estimated Failure Rate

Estimate the test Reliability using Kuder-Richardson formula 21:

R = \frac{n \sigma^2 - M (n-M)}{\sigma^2 (n-1)}

N. \sigma^2 = \frac{A \times B}{A}
O. n - \sigma = \frac{A - F}{A}
P. M (B-\sigma) = \frac{F \times C}{A}
Q. The numerator = \frac{N - F}{A}
R. n - 1 = \frac{A - 1}{A}
S. The denominator = \frac{B \times A}{A}
T. Reliability = \frac{\sigma}{S}

Figure 1. Worksheet for Computing Estimates Manually.
Normal Curve

Fail Point
Area below fail point = 50.00 - 1.68 = 3.32% which is the percent of students expected to fail exam. $x$ is the difference in raw score points between the mean and the fail point.

Fractional parts of the total area under the normal probability curve, corresponding to distances on the baseline between the mean and successive points laid off from the mean in units of standard deviation

Example: between the mean and a point $1.51s (X = 1.51)$ are found 43.45% of the entire area under the curve.

<table>
<thead>
<tr>
<th>$z$</th>
<th>.00</th>
<th>.01</th>
<th>.02</th>
<th>.03</th>
<th>.04</th>
<th>.05</th>
<th>.06</th>
<th>.07</th>
<th>.08</th>
<th>.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.0000</td>
<td>0.0040</td>
<td>0.0080</td>
<td>0.0120</td>
<td>0.0160</td>
<td>0.0199</td>
<td>0.0239</td>
<td>0.0273</td>
<td>0.0299</td>
<td>0.0319</td>
</tr>
<tr>
<td>1.00</td>
<td>0.1359</td>
<td>0.1359</td>
<td>0.1363</td>
<td>0.1367</td>
<td>0.1371</td>
<td>0.1375</td>
<td>0.1379</td>
<td>0.1383</td>
<td>0.1387</td>
<td>0.1391</td>
</tr>
<tr>
<td>2.00</td>
<td>0.2567</td>
<td>0.2567</td>
<td>0.2570</td>
<td>0.2573</td>
<td>0.2576</td>
<td>0.2578</td>
<td>0.2581</td>
<td>0.2583</td>
<td>0.2586</td>
<td>0.2588</td>
</tr>
<tr>
<td>3.00</td>
<td>0.3793</td>
<td>0.3793</td>
<td>0.3796</td>
<td>0.3798</td>
<td>0.3801</td>
<td>0.3803</td>
<td>0.3805</td>
<td>0.3807</td>
<td>0.3809</td>
<td>0.3811</td>
</tr>
<tr>
<td>4.00</td>
<td>0.4987</td>
<td>0.4987</td>
<td>0.4989</td>
<td>0.4991</td>
<td>0.4993</td>
<td>0.4995</td>
<td>0.4996</td>
<td>0.4998</td>
<td>0.4999</td>
<td>0.5000</td>
</tr>
</tbody>
</table>

Figure 2. Normal Curve Probability Table.
STEP 7: DETERMINE THE AREA UNDER THE NORMAL DISTRIBUTION CURVE BETWEEN THE MEAN AND THE FAIL POINT using the table at Figure 2.

STEP 8: SUBTRACT THE VALUE IN STEP 7 FROM 50 PERCENT. This value is the estimated failure rate. It assumes the distribution of student scores approximates a normal distribution.

STEP 9: COMPUTE THE TEST RELIABILITY by substituting the appropriate values in the reliability formula.

SECTION C - COMPUTER PROGRAM FOR COMPUTING ESTIMATES

A computer program has been written in the BASIC language to expedite the computing process. The steps in the procedure are as follows:

STEP 1: Estimate the ease and discrimination indexes for each item in the item bank.

STEP 2: Input the item ease and discrimination indexes for the selected items into a disk file via a remote terminal. Do not use decimal points in inputting the data.

STEP 3: Use the ISE 2 computer program (see Figure 3) to compute the estimates and print out a report. In running the ISE 2 program, the file name for the data file should be entered in line 060. Line 070 should be checked (listed) to assure the read statement corresponds to data listed in the file. The value "y" will read the ease index, and "z" the discrimination index.

STEP 4: A report will be printed out on the remote terminal. A sample report is shown in Figure 4.
10 REM***THIS PROGRAM ESTIMATES MEANS, FAILURE RATES, AND RELIABILITY***
20 REM
30 REM ***DATA IS ENTERED FROM A FILE***
40 PRINT "ENTER COURSE AND FORM NUMBER"
50 INPUT C1,C2
60 FILES 6315058
70 FILES NRMCRV1
80 READ #1,X,Y,Z,W
90 N=N+1
100 K=Y
110 D=Z
120 IF MORE #1 THEN 80
130 REM***COMPUTE AVG EASE***
140 h=e/(n*100)
150 REM*** COMPUTE AVG ITEM DISC***
160 q=(d/100)/(n)
170 REM***COMPUTE MEAN OF RAWSCORES***
180 r=n*h
190 q=n*.60
200 REM***COMPUTE VARIANCE***
210 a=4.5
220 v=(d/100)^2
230 v=((d/100)^2)/(a)
240 s=v*.5
250 rem compute diff mean and fail pt in sd
260 o=(r-q)/s
270 PRINT "diff mean and fp in sd=", o
280 (0(=0*10)+.5 \0=INT(0)
290 READ #2,E,F
300 IF E<>0 GOT0290
310 I=.50 -F
320 k=(n*v-r*(n-r))/(v*(n-1))
330 PRINT(PRINT(PRINT
340 PRINT tab(14);"COURSE EXAMINATION STATISTICAL ESTIMATES"
350 PRINT TAB(16);"COURSE";C1  "FORM";C2;SPC(10);"DATE"; SPC(2);DATS
360 PRINT using 370, n
370: MR ITEMS=    ###
380 PRINT using 390, h
390: AVG EASE=    ###
400 PRINT using 410, q
410: AVG ITEM DISC=    ###
420 PRINT using 430, r
430: MEAN=   ###
440 PRINT using 450, s
450: STANDARD DEVIATION=    ###
460 PRINT using 470,q
470: PASS/FAIL POINT=    ###
480 PRINT using 490, t
490: EST FAILURE RATE=    ###
500 PRINT using 510, k
510: RELIABILITY=    ###
520 PRINT
530 END

FIGURE 3. ISE PROGRAM TO COMPUTE EXAMINATION STATISTICAL ESTIMATE
<table>
<thead>
<tr>
<th>ITEM</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nr. of Items</td>
<td>12</td>
</tr>
<tr>
<td>Avg. Ease</td>
<td>.71</td>
</tr>
<tr>
<td>Avg. Time Disc</td>
<td>.281</td>
</tr>
<tr>
<td>Mean</td>
<td>51.1</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>9.53</td>
</tr>
<tr>
<td>Pass/Fail Point</td>
<td>43.2</td>
</tr>
<tr>
<td>Test Failure Rate</td>
<td>.20</td>
</tr>
<tr>
<td>Reliability</td>
<td>.349</td>
</tr>
</tbody>
</table>

Figure 1. Printout of a Statistical Report.
Section D - Conclusions

Findings:

Comparisons were made between the estimates for several courses and item analyses based on samples of student test papers. The results showed generally close agreement. Differences were approximately of the same magnitude as differences found between two different analyses. Figure 5 is a table comparing estimates with student samples of 51 and 201. Zeros on the table indicate that data are not available. "CRSE" and "FM" indicate ECI course and examination form number.

![Table of Estimated Statistics]

Figure 5. Comparison of Estimated Statistics with Analysis of Student Samples.

Although estimates are generally close to the actual analyses, it is likely that some refinements can be made to the procedure, and guidelines can be prepared to assist test constructors in making estimates, and thus improve these estimates.
Significant advantages to be realized from using the estimating procedure are that it will (1) help assure that different forms of the CE's are equivalent, (2) reduce the number of CE's with excessive failure rates or low reliability and (3) require test constructors to carefully evaluate an item's function in a test. This will result in distinct improvement in test quality.

Summary:

A system has been developed to estimate examination statistics before an examination has been administered. The system requires the test constructor to make an estimate of the ease index and discrimination index for each item. These indexes are then used to compute test estimates using the worksheet or the computer program. Based on samples of actual student data the system has been found to provide relatively close estimates of test performance.
REFERENCES


Extension Course Institute, Hq Operating Instruction 11-12, Volume Review Exercises and Course Examinations, 20 October 1972.