ED 205 594 TM 810 495

AUTHOR Mills, Craig N.: Simon, Robert
TITLE A Method for Determining the Length of Criterion-Referenced Tests Using Reliability and Validity Indices.
INSTITUTION Massachusetts Univ., Amherst. School of Education.
SPONS AGENCY Air Force Human Resources Lab., Brooks AFB, Texas.
PUB DATE Apr 81
NOTE 35p.: Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Computer Assisted Testing: *Criterion Referenced Tests: Cutting Scores: Error of Measurement: Item Banks: *Test Construction: Test Reliability: Test Validity
IDENTIFIERS Binomial Distribution: Sample Size: TESTLEN: *Test Length

ABSTRACT

When criterion-referenced tests are used to assign examinees to states reflecting their performance level on a test, the better known methods for determining test length, which consider relationships among domain scores and errors of measurement, have their limitations. The purpose of this paper is to present a computer system named TESTLEN, which allows test developers to determine optimal criterion-referenced test lengths via simulation according to user-specified conditions. Such conditions may include ability distribution, item statistics, cut-off score, advancement score, test model, number of examinees, and number of replications. Output includes item statistics and values of decision consistency, kappa, and decision accuracy for each replication. The mean, range, and standard deviation of decision consistency, kappa, and decision accuracy are reported across replications. Also reported is the proportion of the examinee group assigned to each mastery classification. A school district developing a test as a diagnostic examination is used as an example of TESTLEN's practical application. Recommendations and directions for the use of TESTLEN appear in the appendix. (Author/AEF)

1/19/81

A Method for Determining the Length of Criterion-Referenced Tests Using Reliability and Validity Indices[1,2,3]

Craig N. Mills and Robert Simon[4]
University of Massachusetts, Amherst

Criterion-referenced tests are used to determine an examinee's status with respect to some well-defined domain of behavior (Hambleton & Eignor, 1979; Popham, 1978). Construction of a criterion-referenced test (CRT) usually involves (among other things) drawing a representative sample of items from a pool of items which measures the domain of content of interest. Of central importance in the test development process is the determination of the number of items to be included. The length of the test (or subtests if several objectives are measured in a test) is directly related to the usefulness of the scores. In general, short tests lead to less reliable and valid scores than longer tests. Longer tests, however, while generally resulting in more precise estimates of ability, require more testing time and may cause examinee fatigue if they become very long. Also, since it is often the case that several objectives are assessed in a single CRT, practical

[1] A paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA, April, 1981.

[2] This project was performed pursuant to a contract from the U.S. Air Force Human Resources Laboratory. However, the opinions do not necessarily reflect their position or policy, and no official endorsement by the Air Force should be inferred.

[3] Laboratory of Psychometric and Evaluative Research Report No. 110, Amherst, MA: School of Education, University of Massachusetts, 1980.

[4] This project was performed under the direction of Dr. Ronald K. Hambleton. The authors wish to thank him for his direction, criticism, and advice throughout the project and the preparation of this paper.

considerations argue against a large number of items per objective. It is important, therefore, that criterion-referenced tests contain enough items to yield scores with desired levels of reliability and validity without requiring excessive amounts of testing time.

In many instances, the purpose of a CRT is to provide an estimate of an examinee's domain score with respect to an objective (or competency) of interest. In such a case, when the purpose is to estimate a domain score, the relationship among domain scores, errors of measurement, and test length can be used to determine an optimum test length (Lord & Novick, 1968).

The primary use of CRTs is, however, to assign examinees to categories or states reflecting levels of performance in relation to the objectives measured in a test. When mastery decisions are being made it is possible to determine test length in relation to the number of misclassification errors which can be tolerated.[1] The purpose of this paper is to describe a system, implemented with the aid of a computer, which can be used to determine test lengths which will lead to specified levels of classification errors. First, several procedures for determining test lengths will be reviewed. After the brief review, the computer-assisted system for determining test length will be presented.

Methods for Determining Test Length

Millman (1972, 1973) considered the relationship between test length, advancement scores, and the probability of misclassification of an individual with a known domain score by using the simple binomial test model. The assumptions of the model are well-known and can be found elsewhere (Millman, 1972, 1973; Hambleton & Eignor, 1979; Lord & Novick, 1968). Millman's tables provide the probability of incorrect classification of individuals with known domain scores for several test lengths, advancement scores, and cut-off scores on the domain score scale.

Wilcox (1976) related the work of Fhaner (1974) to Millman's (1973) work. An indifference zone is used in the Fhaner-Wilcox method for determining test length. An indifference zone is that distance around the cut-off score in which it is assumed that relatively little harm is done when examinees with domain scores on that interval are misclassified. Certainly in most instructional situations such misclassifications result in only short-term assignment to instructional sequences. Masters who are close to the cut-off score who are misclassified as non-masters may benefit from a short remediation sequence. Non-masters who are incorrectly classified as masters will, in all likelihood, be quickly identified. The more serious errors are those which misclassify individuals who are farther from the cut-off score. The binomial model is utilized in the work of Fhaner and Wilcox to determine test lengths which reduce, to specified levels, the likelihood of misclassification of individuals at the ends of the indifference zone.

Two problems limit the usefulness of the systems described above. First, a very good prior estimate of an examinee's domain score is required with the Millman method. Since the purpose of the test is to estimate the domain score, such a prior estimate will, in all likelihood, not be available or, if it is available, it may be imprecise. Second, Millman's work determines optimal test length for examinees at a single point on the domain score scale. (The system described by Wilcox [1976] considers only examinees at two points on the domain scale.) That is, the Millman and Fhaner-Wilcox methods determine test length for specific individuals; the two methods do not consider the case when a group of examinees is of interest. To the extent that a group of examinees with varying domain scores is to be tested, the two systems described above will result in test lengths which are not optimal for the group. Usually, the resultant tests will be longer than necessary to achieve adequate levels of decision accuracy.

Many test developers want to determine test lengths which will achieve desired levels of reliability and validity for a group of examinees. What is needed in that case is a system which incorporates group information into the decision regarding test length. Eignor and Hambleton (1979) and Eignor (1979) utilized group information when investigating the relationship between test length and several criterion-referenced measures of reliability and validity. Using the simple binomial model and the compound binomial model (often considered a more plausible model than the simple binomial model for explaining examinee performance), Eignor and Hambleton (1979) produced graphs of several reliability and validity indices for tests of various lengths for five substantially different domain score distributions (cut-off = .80). Several distributions were needed because measures of decision consistency and decision accuracy are dependent upon the location of the distribution of ability in relation to the cut-off score. Eignor and Hambleton clearly demonstrated that criterion-referenced measures of reliability and validity decrease as the distribution of domain scores moves toward (or centers over) the cut-off score. Eignor (1979) considered additional test lengths, domain score distributions, and advancement scores. The tables and graphs presented in the studies cited above should provide useful guidelines for practitioners who are concerned with determining optimal test lengths. At least three limitations exist, however, in the Eignor-Hambleton solution to the test length determination problem. If a test developer feels that the group of examinees of interest has a somewhat different distribution of domain scores than those considered in the two studies, the Eignor-Hambleton graphs will be of limited value. Similarly, the value of the information provided by the two studies is reduced considerably when test developers consider test lengths and advancement scores different than those reported. Third, if the item pool to be used in the test is not similar to one of the item pools used in the Eignor-Hambleton solution, the results will be limited in value. In summary, their tables are not sufficiently flexible to satisfy the requirements of many testing situations. A better system would be one in which test developers could more closely simulate local conditions by controlling the distribution of domain scores and the range of item statistics and then consider the consequences of various test lengths and advancement scores on the statistics of interest. In the next section of the paper such a system is described.

TESTLEN[1]: A System for Determining the Length of Criterion-Referenced Tests

One method by which optimal test lengths can be determined is to simulate local conditions on a computer. The FORTRAN program TESTLEN is designed to allow users to specify local conditions and to simulate test performance. By simulating several possible test lengths and cut-off scores users can obtain estimates of various statistics of interest. The values obtained may then be used to make decisions regarding optimal test lengths. As a result, requirements for test development or item selection are clarified.

TESTLEN will simulate parallel administrations of several criterion-referenced tests. Characteristics of the tests (test length, cut-off, distribution of item parameters) and characteristics of the examinee pool (number of examinees, distribution of domain scores) are under user control. Also, under user control is the number of replications of each parallel form administration to be simulated. Multiple replications allow users to determine the stability of the results. A brief description of the options available in the program is provided in this section of the paper. Detailed instructions for using the program can be found in Appendix A.

Using TESTLEN with Item Response Data

If users have field tested a set of items, it may be desirable to know the effects on decision consistency and decision accuracy of forming parallel tests by choosing specific subsets of the items. If examinee reponses have been scored (1 if correct and 0 otherwise) the data may be used as input for the program. If data on an external criterion

[1] Source debks may be obtained by writing to the authors at the Laboratory of Psychometric and Evaluative Research, School of Education, University of Massachusetts, Amherst, MA 01003. In order to cover costs of duplication of source decks, computer cards and mailing, checks in the amount of $25.00 made payable to the University of Massachusetts should accompany requests.

is also available it may be input as well. If an external measure is not available, decision accuracy is not computed.

Using TESTLEN With Only a Description of the Examinees

It may be that a user has an idea of the relative position of the students to be tested in relation to the content of interest, but does not have information about item characteristics. For example, it might be known that mathematical computation has always been a weak area in a course, but standard norm-referenced tests have always been purchased and item statistics have not been collected. In this case a simulation which utilizes the binomial test model might be chosen.

If a simulation which utilizes the binomial model is to be used, the user may specify the number of examinees out of 100 thought to be in each tenth of the domain score scale. Alternately, the user may choose values to describe a beta distribution which approximates the local domain score distribution. Beta distributions are defined on the interval [0, 1] which is the scale on which domain scores are located. Examples of distributions obtained for five different beta distributions are located in Table 1. Other examples of beta distributions and the statistics which describe them can be found in Novick and Jackson (1974, 112-113).

Using TESTLEN with Description of Both the Examinees and the Items

If users have information pertaining to item statistics as well as an indication of the distribution of examinee domain scores the compound binomial model may be used for the simulations. Pertinent

Table 1

Domain Score Distributions Obtained in the Program TESTLEN from Specification of Five Beta Distributions

Beta Parameters		Theoretical Results		Score Interval Midpoint									
P	Q	Mean	Standard Deviation	.05	.15	.25	.35	.45	.55	.65	.75	.85	.95
4	1	.80	.027	0	0	1	1	4	7	11	17	25	34
8	2	.80	.015	0	0	0	0	2	5	12	23	34	24
7	2	.78	.017	0	0	0	1	3	7	14	24	31	20
5	1	.83	.020	0	0	0	1	2	5	9	16	26	41
2	2	.50	.050	3	8	11	13	15	15	14	11	7	3

item statistics would be the range of difficulties and item-test correlations. TESTLEN simulates the compound binomial model via the one-, two-, or three-parameter logistic latent trait models, however, knowledge of latent trait theory is not required to use the program. If the compound binomial model is chosen for the simulation, the user must provide a description of the item pool to be used in simulating performance. Most methods for determining test length do not consider characteristics of the item pool. These characteristics are, however, very important. "Other things being equal, heterogeneous item pools require longer tests than do homogeneous pools to obtain similar levels of reliability and validity." In some cases, the increase in number of items required can be substantial (Hambleton, Mills, & Simon, 1980).

If the user desires, traditional item statistics (p's and r's) may be used. TESTLEN will transform these statistics into appropriate latent trait parameters. In this case, the distribution of domain scores is specified in the same way as when the binomial simulations are performed. That is, the user may read in the number of examinees out of 100 in each tenth of the domain score scale or a beta distribution may be specified.

If latent trait theory is familiar to the user, latent-trait parameters for items and examinees may be used. In this case, each parameter (difficulty, discrimination, pseudo chance, ability) may be distributed either normally with a specified mean and standard deviation or uniformly in a specified range.

Output from TESTLEN

Output from the program includes the following information for each replication:

1. item difficulties (p-values)
2. item-test (subtest) correlations
3. item b, a, and c values (latent trait simulations only)
4. the number of examinees
5. decision consistency
6. kappa
7. decision accuracy
8. chance agreement
9. proportion of examinees in each mastery classification

In each situation the mean, range, and standard deviation of decision consistency, kappa, and decision accuracy across the replications are reported.

Figure 4 contains a portion of an output from the program. The simulation utilized the three-parameter logistic latent trait model to generate the responses of 100 examinees to randomly parallel forms of a ten-item test. The 100 examinees were distributed normally on the latent ability scale with mean 0.0 and standard deviation 1.0. Item difficulties were specified to range from -2.00 to +2.00; discrimination ranged from +0.40 to +2.00; pseudochance values ranged from +0.15 to +0.25. The cut-off score was set at 0.00 (the center of the ability distribution) and the advancement score was 5 items correct.

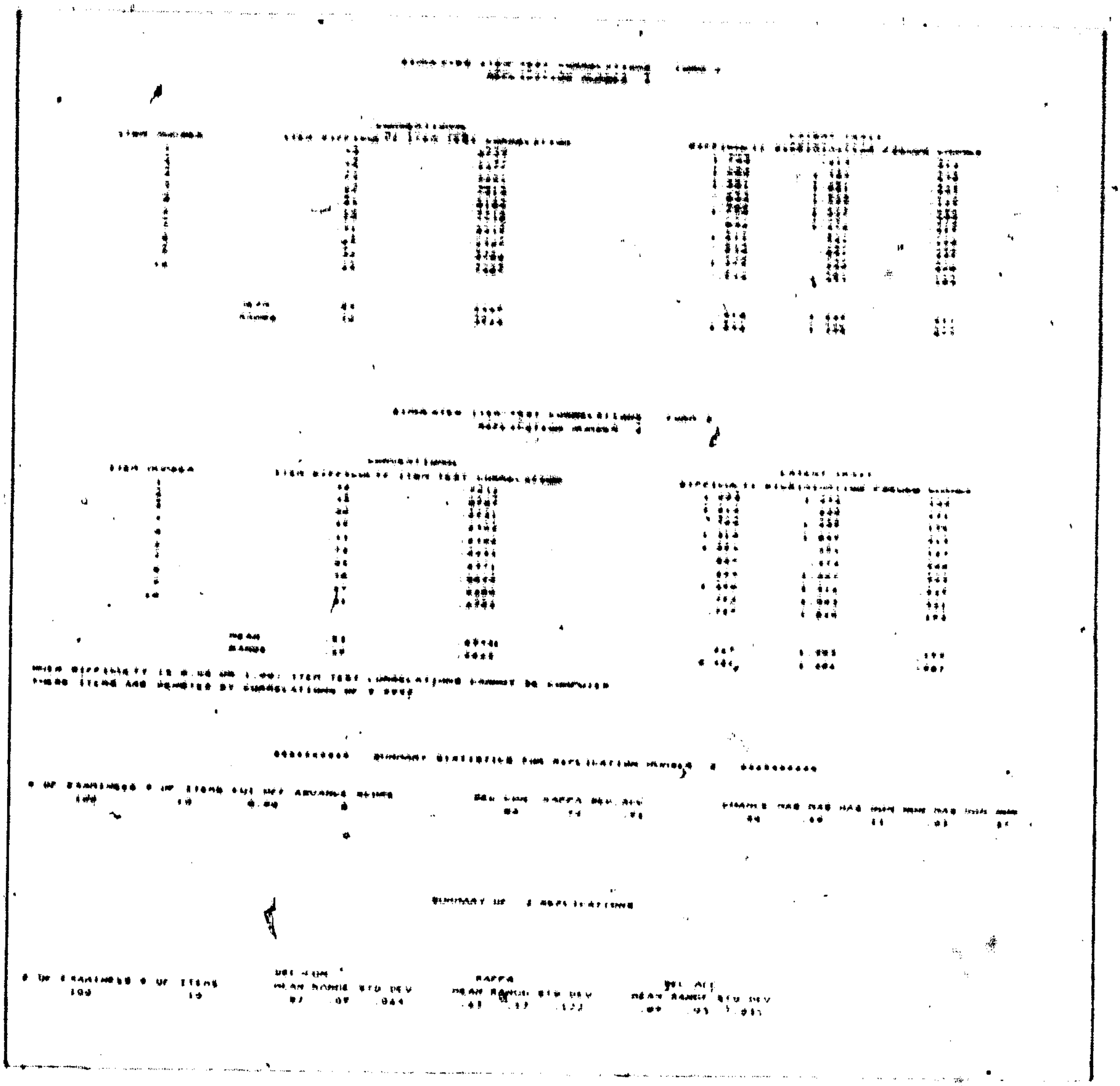

Figure 1. Sample output from TESTLEN. This is the second of two replications in which latent trait theory was used to simulate the performance of 100 examinees on randomly parallel forms of a 10-item test.

As can be seen in the figure, this was the second replication of the situation. The data at the bottom of the figure provide information regarding the mean, range, and standard deviation of the three statistics of primary interest (decision consistency, kappa, and decision accuracy) for the two replications.

An Example of a Practical Application of Program TESTLEN

TESTLEN can be used early in the test development process to provide useful data for decision-making. By simulating performance at several test lengths with cut-off and advancement scores of interest, developers can obtain estimates of the effect of these factors on consistency and accuracy of the test results. Estimates of the proportion of examinees who will need remediation are also obtained.

In order to illustrate an application of the program, suppose a school district is developing a test which will be used as a diagnostic examination. Results will be used to place students into an individualized curriculum. Fifteen objectives have been identified as indicators by the instructors of the course. All objectives are to be tested with as many items as needed to reach consistent and accurate classifications at least 70 percent of the time. The test must not, however, require more than 100 minutes to administer including distribution and collection of materials. Randomly parallel forms are to be developed and administered to approximately 300 students each year.

For the first objective, it is desired to classify individuals as having achieved the objective if they have domain scores equal to or greater than 0.80. Past experience would indicate that students entering the course generally have domain scores greater than 0.50 and that they are distributed uniformly between 0.50 and 1.00. Unit tests have indicated that items range from easy to moderate in difficulty (p-values range from about 0.50 to about 0.90) and that discrimination indices are all around 0.40. There appears to be little or no guessing on the items.

It can be seen from the description above that although the number of items used for each objective will vary, it is important to use as few items as possible for each objective in order to meet the time constraints. Table 2 shows the results obtained from TESTLEN for 11 possible test lengths and advancement scores for the first objective. The domain scores for the 300 examinees were distributed uniformly between 0.50 and 1.00. Five replications of each test were simulated. Means, ranges, and standard deviations of decision consistency, kappa, and decision accuracy for each test length and advancement score are included in the table. It can be seen that 8 items with an advancement score of 6 correct would be needed for this objective if desired levels of both decision consistency and decision accuracy are to be obtained.

Table 2

Measures of Decision Consistency, Kappa, and Decision Accuracy
Obtained from TESTLEN for 11 Test Lengths
and Advancement Scores
(π_o = .80, N=300, 5 replications)

Test Characteristics		Decision Consistency			Kappa			Decision Accuracy		
Number of Items	Advancement Scores	Mean	Range	Standard Deviation	Mean	Range	Standard Deviation	Mean	Range	Standard Deviation
2	2	.58	.07	.026	.16	.14	.055	.65	.02	.007
3	2	.72	.08	.033	.17	.23	.084	.55	.02	.009
3	3	.59	.04	.018	.20	.08	.032	.71	.05	.019
4	3	.68	.07	.031	.27	.12	.057	.64	.02	.012
5	4	.66	.03	.012	.33	.08	.030	.71	.05	.023
6	5	.67	.06	.023	.34	.11	.047	.74	.05	.020
7	5	.71	.06	.023	.35	.13	.051	.66	.06	.022
7	6	.67	.08	.032	.35	.16	.063	.78	.07	.028
8	6	.72	.09	.035	.39	.21	.084	.70	.04	.014
9	7	.70	.05	.019	.38	.10	.042	.75	.06	.029
10	8	.73	.06	.027	.47	.11	.050	.80	.04	.016

Recommendations for Use

The purpose of TESTLEN is to allow test developers to determine optimal criterion-referenced test lengths via simulation. In this section a few general recommendations regarding use of the program are provided.

It is not always possible to accurately specify characteristics of examinees and item pools. In such cases test developers will probably want to err on the side of conservatism since it may be better to have a few extra items than to err on the short side and have an unacceptable number of classification errors. The following recommendations are intended to provide guidelines for producing conservative test lengths. First, use sample sizes similar to the number of examinees to be tested. Larger samples will yield more stable estimates of reliability and validity, but test developers need to know the expected range of these statistics in their situation. Second, when in doubt about the distribution of domain scores, it is better to center the distribution close to the cut-off score. The closer the distribution is to the cut-off, the more classification errors will result. Thus, more items will be required to reach acceptable levels of decision consistency. Third, if characteristics of the item pool are not established, specify heterogeneous pools. This will lead to more conservative estimates of test length.

TESTLEN simulates parallel-form administrations of criterion-referenced tests. Some options of the program allow the user to choose between randomly or statistically parallel tests. If two tests are to be developed by randomly selecting items from an item pool, the user should specify randomly parallel tests. If, however, the tests are to be matched on item statistics, the user should choose the option for statistical parallelism. This option would also be chosen if only one test form is to be developed. Choosing statistical parallelism for the simulation would be akin to investigating a test-retest situation with one form.

Most of the options included in TESTLEN rely on a random number generator. Users will possibly have to modify the program to conform to the random number generator at their facility. Users should also determine the type of seed which produces best results with the random number generator.

Finally, the test length determination problem must be solved for each objective (or competency) on the test for which mastery decisions will be made.

Summary

In this paper, several methods for determining the length of criterion-referenced tests which are used to make mastery decisions were reviewed. For various reasons, the methods were considered less than ideal. A method which utilizes reliability and validity data to determine test length, implemented via the use of a computer, was presented. The method utilizes data which are relevant to the local situation to determine test lengths. Among the variables under user control are test model, number of items, number of examinees, ability distribution, cut-off score, advancement score, and number of replications (parallel administrations) to be conducted. Options also exist to allow the utilization of actual, rather than simulated response data.

References

Eignor, D. R. Psychometric and methodological contributions to eriterion-referenced testing technology. Unpublished doctoral dissertation, University of Massachusetts, 1979.

Eignor, D. R., & Hambleton, R. K. Effects of test length and advancement score on several criterion-referenced test reliability and validity indices. Laboratory of Psychometric and Evaluative Research Report No. 86. Amherst, MA: School of Education, University of Massachusetts, 1979.

Fhaner, S. Item sampling and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-175.

Hambleton, R. K., & Eignor, D. R. A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70. Amherst, MA: School of Education, University of Massachusetts, 1979.

Hambleton, R. K., Mills, C. N., & Simon, R. Determining the optimal length of a criterion-referenced test. Laboratory of Psychometric and Evaluative Research Report No. 111. Amherst, MA: School of Education, University of Massachusetts, 1980.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Millman, J. Tables for determining number of items needed on domain-referenced tests and number of students to be tested. Instructional Objectives Exchange. Los Angeles, CA, 1972.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Novick, M. R., & Jackson, P. H. Statistical methods for educational and psychological research. New York: McGraw Hill, 1974.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978.

Wilcox, R. A note on the length and passing score of a mastery test. Journal of Educational Statistics, 1976, 1, 359-364.

APPENDIX A

Directions for Using Program TESTLEN

The purpose of this handbook is to provide step-by-step instructions for using Program TESTLEN. It is assumed that the user has knowledge of format statements. That is, the user understands that a variable which is specified to have a format of F5.2 is a real number with two decimal places. The format I5 refers to a five place integer. With the exception of a random number generator, the program is pretty much machine independent. Although all options have worked satisfactorily no claim is made that the program is error free.

In order to use this handbook, the user must answer certain questions about the simulation which is desired. Based on the answers to each question, the user is referred to certain sections of the handbook where detailed instructions for setting up input are provided. After the instructions an example is provided.

Input parameter cards are to be located on Unit 6. Output is written to Unit 1. The first card of any TESTLEN run contains only one variable. This variable (NJOBS, Format = I5) directs the program as to how many different simulations or data sets are to be processed. The user should go through the directions in this handbook once for each job specified.

Directions for Assembling Input Decks

I. What type of simulation is desired?

If currently available item response data is to be used, go to II. For example, data from a pilot administration might be available. Subsets of items can be organized into parallel tests and results calculated.

If a binomial simulation is desired, go to III. The description of the examinee population is under user control.

If a compound binomial simulation is desired, go to IV. The description of both the examinee population and the item pool are under user control.

II. Utilization of Item Response Data

A. Is an external criterion measure available? An external criterion is a measure other than the test of interest which can be used to separate examinees into mastery categories. Another test or course grades might be used. The agreement between the classification of examinees on the test of interest and on the external criterion can be an indication of the validity of the test.

If there is not an external criterion, go to A.1.

If there is an external criterion, go to A.2.

A.1. The item responses should be located on Unit 11. The input deck should be set up as follows:

CARD 2: INTYPE, NITEST, NEX, CUT, CUTS, NREP, N, IPAR

INTYPE(I5) = 1

NITEST(I5) = The number of items to be included in each form. For example, if responses are to be organized into two parallel tests of 20 items, NITEST=20. NITEST cannot exceed 50.

NEX(I5) = The number of examinees (cannot exceed 1000).

CUT(F5.2) = Set to 0.00. This variable is not used when an external criterion is not available.

CUTS(I5) = The advancement score. This is the number of items an examinee must answer correctly to be classified as a master on the test.

NREP(I5) = The number of parallel administrations to be included in the current job.

N(I7) = Seed for the random number generator.

IPAR(I5) = Set to 0.

CARD 3: FMTIR

FMTIR(15A1) = The format by which the responses are to be read from Unit 11.

Example: Suppose data is available on 20 items and an instructor wants to separate the items into two parallel tests of 10 items each. 200 examinees took the items. The advancement score being considered is 7. The items are in fields of 2 on the data tape. The deck would be as follows:

```
Card 2:  ----1---10--200-0.00----7----19834513----0
Card 3:  (2012)
```

A.2. <u>Is the external criterion on the same file as the item responses or is it on a different file?</u>

If the external measure is on the same file, go to A.2.a.

If the external measure is on a different file, go to A.2.b.

A.2.a. The item responses and the criterion measure should be on Unit 11. The criterion measure should follow the last response. The input deck should be set up as follows:

CARD 2: INTYPE, NITEST, NEX, CUT, CUTS, NREP, N, IPAR

INTYPE(I5) = 2

NITEST(I5) = The number of items to be included in each form. For example, if responses are to be organized into two parallel tests of 20 items, NITEST=20. NITEST cannot exceed 50.

NEX(I5) = The number of examinees (cannot exceed 1000).

CUT(F5.2) = The cut-off score on the external criterion. If, for example, the external criterion were grade point average, the cut-off might be 3.00.

CUTS(I5) = The advancement score. This is the number of items an examinee must answer correctly to be classified as a master on the test.

NREP(I5) = The number of parallel administrations to be included in the current job.

N(I7) = Seed for the random number generator.

IPAR(I5) = Set to 0.

CARD 3: FMTIR

FMTIR(15A1) = The format by which the responses and the external criterion are to be read from Unit 11.

<u>Example</u>: Suppose data is available on 30 items and a district wants to separate the items into two parallel tests of 15 items each. 300 examinees took all of the items on two different occasions. The external criterion is previous course grades and the cut-off is 2.75. An advancement score of 10 is being considered. The item responses are in fields of 1 on the data tape with GPA following in a field of 4. The deck would be as follows:

```
Card 2:  ----2---15--300-2.75---10----28763547----0
Card 3:  (30I1, F4.2)
```

A.2.b. The item responses should be located on Unit 11. The external criterion should be located on Unit 12. The input deck should be set up as follows:

CARD 2: INTYPE, NITEST, NEX, CUT, CUTS, NREP, N, IPAR

INTYPE(I5) = 3

NITEST(I5) = The number of items to be included in each form. For example, if responses are to be organized into two parallel tests of 20 items, NITEST=20. NITEST cannot exceed 50.

NEX(I5) = The number of examinees (cannot exceed 1000)

CUT(F5.2) = The cut-off score on the external criterion. If, for example, the external criterion were grade point average, the cut-off might be 3.00.

CUTS(I5) = The advancement score. This is the number of items an examinee must answer correctly to be classified as a master on the test.

NREP(I5) = The number of parallel administrations to be included in the current job.

N(I7) = Seed for the random number generator.

IPAR(I5) = Set to 0.

CARD 3: FMTIR

FMTIR(15A1) = The format by which the responses are to be read from Unit 11.

CARD 4: FMTEX

FMTEX(15A1) = The format by which the external criterion is to be read from Unit 12.

Example: Suppose data is available on 16 items and a district wants to separate the items into two parallel tests of 8 items. 1000 examinees took the items. The advancement score being considered is 5. The external criterion is teacher ratings; 1.0=master, 0.0=non-master. The deck would be as below:

```
Card 2:  ----3----8-1000-1.00----5----11234567----0
Card 3:  (16I1)
Card 4:  (F3.1)
```

III. Simulations Utilizing the Binomial Model

A. Are the percent of examinees in each tenth of the domain score scale to be read in or will a beta distribution be used to describe the population?

If the user wants to read in the number of people in each tenth of the scale, go to A.1.

If a beta distribution is to be used, go to A.2.

A.1. The deck should be set up as follows:

CARD 2: INTYPE, NITEST, NEX, CUT, CUTS, NREP, N, IPAR

INTYPE(I5) = 4

NITEST(I5) = The number of items to be included in each form. For example, if responses are to be organized into two parallel tests of 20 items, NITEST=20. NITEST cannot exceed 50.

NEX(I5) = The number of examinees (cannot exceed 1000).

CUT(F5.2) = The cut-off score on the domain score scale. The cut-off score is a number between 0.00 and 1.00 which represents the domain score at which examinees are considered to be masters.

CUTS(I5) = The advancement score. This is the number of items an examinee must answer correctly to be classified as a master on the test.

NREP(I5) = The number of parallel administrations to be included in the current job.

N(I7) = Seed for the random number generator.

IPAR(I5) = Set to 0.

CARD 3: AREA(I), I = 1, 10

AREA(1)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.00, 0.10].

AREA(2)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.11, 0.20].

AREA(3)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.21, 0.30].

AREA(4)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.31, 0.40].

AREA(5)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.41, 0.50].

AREA(6)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.51, 0.60].

AREA(7)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.61, 0.70].

AREA(8)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.71, 0.80].

AREA(9)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.81, 0.90].

AREA(10)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.91, 1.00].

These 10 numbers must total 100.

Example: Suppose an instructor plans to test 500 examinees on a 10 item test. The cut-off score is to be .75 and the instructor wants to investigate the effects of an advancement score of 7. There is a small group of students (about 10%) who are definitely very low performers. The rest seem to be fairly evenly distributed in the top forty percent of the domain score scale. Five replications of the simulation are desired in order to get a feeling for the range of possible values, the input deck might be as follows:

```
Card 2:  ----4---10--500-0.75----7----54395183----0
Card 3:  ----0----5----5----0----0----0---25---25---20---20
```

A.2. The deck should be set up as follows:

CARD 2: INTYPE, NITEST, NEX, CUT, CUTS, NREP, N, IPAR

INTYPE(I5) = 5

NITEST(I5) = The number of items to be included in each form. For example, if responses are to be organized into two parallel tests of 20 items, NITEST=20. NITEST cannot exceed 50.

NEX(I5) = The number of examinees (cannot exceed 1000).

CUT(F5.2) = The cut-off score on the domain score scale. The cut-off score is a number bewteen 0.00 and 1.00 which represents the domain score at which examinees are considered to be masters.

CUTS(I5) = The advancement score. This is the number of items an examinee must answer correctly to be classified as a master on the test.

NREP(I5) = The number of parallel administrations to be included in the current job.

N(I7) = Seed for the random number generator.

IPAR(I5) = Set to 0.

CARD 3: IP, IQ

IP(I5) = First descriptor of beta distribution.

IQ(I5) = Second descriptor of beta distribution.

Example: Suppose a test developer wants to determine the effects of using a 5 item test with a cut-off of 0.80 and an advancement score of 4. Large numbers of examinees will take the test. Past experience has shown the bulk of the examinees to be located in the region of the cut-off score with a few in the region .40 to .60. Ten replications are to be conducted. The deck may be set up as follows:

```
Card 2:  ----5----5-1000-0.80----4---109812375----0

Card 3:  ----8----2
```

IV. Simulations Utilizing the Compound Binomial Model

A. Are latent trait parameters to be used or will classical statistics be read in and converted to latent trait values?

If latent trait parameters are to be read in, go to A.1.

If classical statistics (p-values, etc.) are to be converted, go to A.2.

A.1. The user must specify distributions which are desired for item difficulty, discrimination, pseudochance, and ability (b, a, c, and θ, respectively). Two options are available. Each variable may be distributed (1) normally with a specified mean and standard deviation or (2) uniformly across a specified range. The input deck should be set up as follows:

CARD 2: INTYPE, NITEST, NEX, CUT, CUTS, NREP, N, IPAR

INTYPE(I5) = 6

NITEST(I5) = The number of items to be included in each form. For example, if responses are to be organized into two parallel tests of 20 items, NITEST=20. NITEST cannot exceed 50.

NEX(I5) = The number of examinees (cannot exceed 1000).

CUT(F5.2) = The cut-off score on the domain score scale. The cut-off score is a number between 0.00 and 1.00 which represents the domain score at which examinees are considered to be masters. TESTLEN converts this value to a cut-off on the ability scale for the test which is generated.

CUTS(I5) = The advancement score. This is the number of items an examinee must answer correctly to be classified as a master on the test.

NREP(I5) = The number of parallel administrations to be included in the current job.

N(I7) = Seed for the random number generator.

IPAR(I5) = 0 if the two tests are to be randomly parallel.

= 1 if the two tests are to be statistically parallel (identical item parameters). This option would be chosen if the item pool is large enough to permit building identical forms or if only one form is actually to be developed and the second form is used as a hypothetical test for simulation purposes only.

CARD 3: IB, BBOT, BTOP

IB(I5) = 1 if a normal distribution of item difficulty parameters (b values) is desired.

= 2 if a uniform distribution of item difficulty parameters (b values) is desired.

BBOT(F5.2) = If IB=1, desired mean of item difficulties.

= If IB=2, lower limit of range of item difficulties.

BTOP(F5.2) = If IB=1, desired standard deviation of item difficulties.

If IB=2, upper limit of range of item difficulties.

CARD 4: IA, ABOT, ATOP

IA(I5) = 1 if a normal distribution of item discrimination parameters (a values) is desired.

= 2 if a uniform distribution of item discrimination parameters (a values) is desired.

ABOT(F5.2) = If IA=1, desired mean of item discrimination values.

= If IA=2, lower limit of range of item discrimination values.

ATOP(F5.2) = If IA=1, desired standard deviation of item discrimination values.

= If IA=2, upper limit of range of item discrimination values.

CARD 5: IC, CBOT, CTOP

IC(I5) = 1 if a normal distribution of item pseudochance (c values) is desired.

= 2 if a uniform distribution of item pseudochance (c values) is desired.

CBOT(F5.2) = If IC=1, desired mean of item pseudochance values.

= If IC=2, lower limit of range of item pseudochance values.

CTOP(F5.2) = If IC=1, desired standard deviation of item pseudochance values.

= If IC=2, upper limit of range of item pseudochance values.

CARD 6: ITHET, THTOP, THBOT

ITHET(I5) = 1 if a normal distribution of ability (θ values) is desired.

= 2 if a uniform distribution of ability (θ values) is desired.

THBOT(F5.2) = If ITHET=1, desired mean of the ability distribution.

= If ITHET=2, lower limit of range of the ability distribution.

THTOP(F5.2) = If ITHET=1, desired standard deviation of the ability distribution.

= If ITHET=2, upper limit of the range of the ability distribution.

Example: Suppose a 7 item test with a cut-off score of .70 and an advancement score of 5 is being considered for use where 150 students will be tested on the objective. There is one form of the test. The range of item difficulties is -2.00 to 2.00; item discriminations range from 0.50 to 1.75 and guessing ranges from 0.15 to 0.25. Ability of students is expected to be normally distributed with a mean of 0.00 and a standard deviation of 1.00. The data would be arranged as follows:

```
Card 2:  ----6----7--150-0.70----5----51287937----1
Card 3:  ----2-2.00+2.00
Card 4:  ----2+0.50+1.75
Card 5:  ----2+0.15+0.25
Card 6:  ----1+0.00+1.00
```

A.2. Are the percent of examinees in each interval of the domain score scale to be read in or will a beta distribution be used to describe the population?

If the number of people in each interval is to be read, go to A.2.a.

If a beta distribution is to be used, go to A.2.b.

A.2.a. The deck should be set up as follows:

CARD 2: INTYPE, NITEST, NEX, CUT, CUTS, NREP, N, IPAR

INTYPE(I5) = 7

NITEST(I5) = The number of items to be included in each form. For example, if responses are to be organized into two parallel tests of 20 items, NITEST=20. NITEST cannot exceed 50.

NEX(I5) = The number of examinees (cannot exceed 1000).

CUT(F5.2) = The cut-off score on the domain score scale. The cut-off score is a number between 0.00 and 1.00 which represents the domain score at which examinees are considered to be masters. TESTLEN converts this value to a cut-off on the ability scale for the test which is generated.

CUTS(I5) = The advancement score. This is the number of items an examinee must answer correctly to be classified as a master on the test.

NREP(I5) = The number of parallel administrations to be included in the current job.

N(I7) = Seed for the random number generator.

IPAR(I5) = 0 if the two tests are to be randomly parallel.

= 1 if the two tests are to be statistically parallel (identical item parameters). This option would be chosen if the item pool is large enough to permit building identical forms or if only one form is actually to be developed and the second form is used as a hypothetical test for simulation purposes only.

CARD 3: LTM, NCH, PBOT, PTOP, RBOT, RTOP

LTM(I5) = 1 if item difficulties vary, but all item discrimination indices are very similar in value and guessing is not thought to be a factor in test performance.

= 2 if item difficulites and discrimination vary, but guessing is not thought to be a factor in test performance.

= 3 if item difficulty and discrimination vary and guessing is thought to affect test performance.

NCH(I5) = The number of options per item.

PBOT(F5.2) = The lower limit of the range of item difficulties (p values) to be included in the test.

PTOP(F5.2) = The upper limit of the range of item difficulties (p values) to be included in the test.

RBOT(F5.2) = The lower limit of the range of item discrimination indices (r values) to be included in the test.

RTOP(F5.2) = The upper limit of the range of item discrimination indices (r values) to be included in the test (RBOT=RTOP if LTM=1).

CARD 4: ITHET, THBOT, THTOP

ITHET(I5) = 4

THTOP(F5.2) = set to 0.00.

THBOT(F5.2) = set to 0.00.

CARD 5: AREA(I), I=1,10

AERA(1)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.00, 0.10].

AREA(2)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.11, 0.20].

AREA(3)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.21, 0.30].

AREA(4)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.31, 0.40].

AREA(5)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.41, 0.50].

AREA(6)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.51, 0.60].

AREA(7)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.61, 0.70].

AREA(8)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.71, 0.80].

AREA(9)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.81, 0.90].

AREA(10)(F5.0) = The number of people out of 100 who are expected to have domain scores on the interval [0.91, 1.00].

<u>Example</u>: Assume an instructor is considering testing an objective with randomly parallel tests of 8 items. Items will be four option multiple-choice items. The cut-off score is 0.75 and the advancement score is 6. A large group of students are performing at high levels and another group is performing at a moderate level. The remaining students are evenly distributed between the extremes. There is a wide range expected in both difficulty and discrimination indices and guessing will probably be a factor. Five replications are desired on a sample of 120 students. The data could be arranged as follows:

```
Card 2:  ----7---18--120-0.75----6----58712367----0
Card 3:  ----3----4-0.30-0.80-0.25-0.65
Card 4:  ----4-0.00-0.00
Card 5:  ----0----0----0----0---20---15----5----5---15---20
```

A.2.b. The deck should be arranged as follows:

CARD 2: INTYPE, NITEST, NEX, CUT, CUTS, NREP, N, IPAR

INTYPE(I5) = 7

NITEST(I5) = The number of items to be included in each form. For example, if responses are to be organized into two parallel tests of 20 items, NITEST=20. NITEST cannot exceed 50.

NEX(I5) = The number of examinees (cannot exceed 1000).

CUT(F5.2) = The cut-off score on the domain score scale. The cut-off score is a number between 0.00 and 1.00 which represents the domain score at which examinees are considered to be masters. TESTLEN converts this value to a cut-off on the ability scale for the test which is generated.

CUTS(I5) = The advancement score. This is the number of items an examinee must answer correctly to be classified as a master on the test.

NREP(I5) = The number of parallel administrations to be included in the current job.

N(I7) = Seed for the random number generator.

IPAR(I5) = 0 if the two tests are to be randomly parallel.

= 1 if the two tests are to be statistically parallel (identical item parameters). This option would be chosen if the item pool is large enough to permit building identical forms or if only one form is actually to be developed and the second form is used as a hypothetical test for simulation purposes only.

CARD 3: LTM, NCH, PBOT, PTOP, RBOT, RTOP

LTM(I5) = 1 if item difficulties vary, but all item discrimination indices are very similar in value and guessing is not thought to be a factor in test performance.

= 2 if item difficulites and discrimination vary, but guessing is not thought to be a factor in test performance.

= 3 if item difficulty and discrimination vary and guessing is thought to affect test performance.

NCH(I5) = The number of options per item.

PBOT(F5.2) = The lower limit of the range of item difficulties (p values) to be included in the test.

PTOP(F5.2) = The upper limit of the range of item difficulties (p values) to be included in the test.

RBOT(F5.2) = The lower limit of the range of item discrimination indices (r values) to be included in the test.

RTOP(F5.2) = The upper limit of the range of item discrimination indices (r values) to be included in the test (RBOT=RTOP if LTM=1).

CARD 4: ITHET, THBOT, THTOP

ITHET(I5) = Set to 3

THBOT(F5.2) = First descriptor of beta distribution.

THTOP(F5.2) = Second descriptor of beta distribution.

Example: Suppose the results of an administration of one form of a 4-item multiple-choice test (5 options) with a cut-off score of 0.90 and an advancement score of 4 are of interest for a group of 50 examinees. Items range from moderate to easy, in difficulty, discriminations are all around .45 and guessing is not thought to be a factor. The average domain score is probably around 0.85, but a few examinees may be at or below 0.50. Only one replication of the simulation is requested. The data might be arranged as below:

```
Card 2:  ----7----4---50-0.90----4----11239867----1
Card 3:  ----1----5-0.50-0.90-0.45-0.45
Card 4:  ----3-6.00-1.00
```