MINIMIZING TEST LENGTH AND MAXIMIZING INFORMATION[1,2]

Rebecca Bryson, San Diego State University

There has been much interest in recent years in the possibility of administering shortened tests while retaining information contained in the full length test. This poses the important question of whether or not presently used full length paper-and-pencil tests can be shortened for computer terminal administration without excessive loss of information.

A major advantage of administering items on computer terminals is that the computer obviates the requirement that all persons be administered the same items. In other words, the test may employ a branching strategy in which sequentially administered items are contingent upon previous responses. However, the extent to which this branching capacity may be useful is largely unknown. The more traditional method of reducing testing time entails selecting (from the full length test) a limited subset of items to be administered to all examinees. Because short tests of this type (linear tests) have the advantage of being easily and expensively administered in paper and pencil form, their effectiveness should be used as a standard against which branching tests may be compared.

Most previous attempts to evaluate branching tests have been accomplished by simulated administration of items utilizing item response data banks. Although analyses of this type are of interest, they are not substitutes for subjecting shortened tests to actual tryout.

Considerations in Shortening Tests

In order to understand the reasons underlying the selection of methods of designing shortened tests used in the present study, it is necessary to consider the parameters involved in various methods of item selection, how these parameters have been previously used to construct shortened tests, and how they may be expected to affect results.

Parallelism between short and long tests is a function of the ways in which known item parameters are used in constructing shortened tests. The way in which these parameters, particularly item difficulty and item discriminating power, should be used to develop short tests has been the subject of some controversy. Approaches for developing branching tests have in the past relied primarily on item difficulty as a means of selecting items to comprise a branching paradigm. The branching rule, stated in its simplest form has been: If a question is answered correctly, administer a more difficult item; if incorrectly, administer an easier item. Using this approach, Bayroff and Seeley (1967) developed verbal and quantitative tests for computerized administration. Scores derived from these short tests correlated more highly with the respective long test scores than the expected value of the correlation of an equivalent number of randomly selected linearly administered items.

Lord (1970) and Stocking (1969) have considered in some detail the expected effects of various branching strategies on measurement of different levels of the ability range when the experimentally manipulated item parameter is item difficulty and item discriminating power is assumed to be constant. They have concluded that, theoretically, measurement in the extremes of the ability distribution should be improved by utilizing the branching capacity.

If the pool of items from which the shorter branching test is selected is scaleable in the Guttman sense, i.e., if passing a given item implies that all easier items will likewise be passed, then the only important consideration is item difficulty. To the extent that test content is not homogeneous in this sense (see Dubois, 1970, for implications of various indices of homogeneity), the likelihood of selecting items which provide maximum criterion discrimination is diminished by attending solely to item difficulty.

The item parameter generally given primary consideration when items are being selected to comprise a short linear test has been discriminating power. If a total test score criterion is used, this entails selecting items which correlate most highly with total test score. One risk in this approach is that highly redundant items might be selected at the expense of items involving important, but relatively unique components of criterion variance (see Loevinger, 1954).

As a remedy for this problem, Anastasi (1968) has advocated selecting items for a liner test according to "net effectiveness," i.e., their unique contribution to the prediction of total test score or some external criterion. She comments, however, that approaches of this type may be criticized on the basis of expected unreliability of partial regression weights when applied to single items. One net effectiveness approach developed by Moonan and Pooch (1966) partially circumvents the unreliability problem by selecting items in order of their contribution to a multiple R, then unit weighting each selected item.

Item analysis methods which select items with high discriminating power as well as those which select items showing discriminating power in a net effectiveness sense have been developed primarily for linear tests, but also may be applied to branching tests. Lord, Novick, and Birnbaum (1968) have considered the joint effect of discriminating power and item difficulty and their relationship to ability. Linn, Rock and Cleary (1969; see also Cleary, Linn and Rock, 1968), have attempted, with varying degrees of success, to incorporate discriminating power into item selection strategies when devising branching tests; however, their index of discriminating power was based on the total group item-test point biserial rather than on discriminating power for the group to whom the item would be administered.

Two methods of selecting items which should theoretically discriminate very accurately among the persons to whom they are administered are outlined below:

Wright and Panchapakesan Parameters (WRIPA). Wright and Panchapakesan (1969) have designed a program to obtain item difficulty and item discriminating power estimates based on item characteristic curves. The item difficulty (log easiness) estimate of an item is related to more conventionally obtained item difficulty estimates, based on the percentage passing, but tends to be stable across samples of varying ability. The item discriminating power estimate, however, refers to the discriminating power among persons whose ability level is such that half of them may be expected to pass the item and half to fail it. While no one has derived an optimal way of combining these parameters for use in developing branching tests, it should be possible to avoid complete reliance on item difficulty estimates by selecting within each difficulty level the item which shows the greatest discriminating power.

BRANCH Approach. A strategy and program for selecting items (when branching is permitted) that should maximize the prediction of total score was devised by Wolfe (1970). The program operates as follows: Point biserial correlations of all items with total test score are calculated. The most discriminating item for the total group is selected and the group is partitioned into those who pass the item and those who fail the item. Correlations of all remaining items with total score are then calculated for each of the two new groups. The most discriminating item for each group is then selected and the groups are split into those who pass and those who fail the second item--producing four groups. This process is continued until a specified number of items is selected or until the item selected fails to make a significant discrimination for the group for which it was selected. The maximum number of groups produced will be 2^n where n equals the number of items to be administered to each person. (This is, of course, true only where n is a constant.)

The WRIPA approach is analogous to selecting items which show maximal discriminating power for a linear test, the major difference being that there is some assurance that items selected have discriminating power for the particular subgroups to which they will be administered. The BRANCH approach is a net effectiveness approach in that maximally discriminating items are selected for subgroups which are homogeneous with respect to previously administered items. Interesting comparisons can be made between BRANCH and other item selection procedures. For example, it can be demonstrated that if items were perfectly Guttman scaleable, BRANCH would select items only on the basis of item difficulty. Or, if the same items were selected by BRANCH for all groups, then there is evidence that a linear test would suffice. (Comparisons between configural scoring and summed scoring would of course be necessary to determine whether or not the particular items missed make any difference.)

In general, it would appear that BRANCH should be an excellent means of item selection, irrespective of the nature of the total test from which items are selected. If the test is highly homogeneous but improved measurement is effected by administering items whose difficulties are reasonably compatible with Ss ability, then the WRIPA scaling parameters should provide useful means of selecting items.

Approaches Used in the Present Study

The present research compared the BRANCH and WRIPA modes of item selection, using a large pool of previously collected item response data for item selection and cross validation. The resulting short tests were then administered on computer terminals. To compare information provided by branching tests with that provided when short easily administered paper-and-pencil tests were used, two types of linear tests were devised. The first linear test included items showing the highest correlation with total test score and is referred to as the high validity (HI VAL) approach. Items included in the second short linear test were selected by SEQUIN to provide a linear net effectiveness comparison. These short linear tests were administered in paper-and-pencil version.

In order to compare all four approaches across tests having somewhat different characteristics, the General Classification Test (GCT) and the Mechanical Aptitude Test (MECH) were used. GCT is a verbal test with extremely high internal consistency (KR_{20} = .975). MECH contains items of two types, tool knowledge and mechanical reasoning. (The mechanical reasoning items are similar to those found on the Bennett Test of Mechanical Comprehension.) As might be expected, MECH is less homogeneous (KR_{20} = .928). Both tests show a fairly wide range of item difficulties (GCT .93 - .20, and MECH .97 - .08).

In general, the four ways of deriving the short tests may be depicted as in Table 1.

TABLE 1

Characteristics of Four Item Selection Procedures

Maximization of Net Effectiveness (Reduce Redundancy)	Branching Permitted (Discriminate at Appropriate Level): Yes	Branching Permitted (Discriminate at Appropriate Level): No
Yes	BRANCH	SEQUIN
No	WRIPA	HI VAL

Each of the four methods was used to construct two short tests--one of items selected from GCT and the other, of items from MECH. Simulated item administration of the eight short tests (generally five items per person) was accomplished using item

response data banks. Tests requiring branching were then administered on computer terminals while the short linear tests were administered by the usual paper-and-pencil methods. Basically, then, the experiment included two phases, the first phase involving item selection and cross validation on large item data banks, and the second phase involving an experimental tryout of the shortened tests via traditional methods or computer.

Hypotheses

1. That extremely short tests (5-6 items) can be developed and administered via computer terminal with little loss of the information contained in the total (100 item) test.

2. That BRANCH is the best means of selecting items for a shortened test. (This was anticipated because BRANCH minimizes redundancy and assures that each item is maximally discriminating for the group to whom it is administered.)

3. That WRIPA is the second best item selection technique for constructing a short GCT test, but third best, for constructing a short MECH test. (The fact that WRIPA allows the administration of items varying in difficulty level but not necessarily contributing to the prediction of all components of criterion variance suggests that it should provide a useful item selection technique for constructing a short GCT test, but because of the somewhat greater heterogeneity of MECH, important information would be lost by the use of the WRIPA procedure.)

4. That SEQUIN is the second best item selection technique for constructing a short MECH test, but third best for constructing a short GCT test. (SEQUIN allows for representation of unique components of criterion variance which should be useful with MECH, but with GCT not so necessary as allowing persons to take items compatible with their ability level.)

5. That the traditional HI VAL approach is the poorest method of selecting items from both GCT and MECH. (The HI VAL approach maximizes neither discrimination at the appropriate ability level nor representation of unique components of criterion variance.)

METHODS

Samples

All samples were composed of men who went through recruit training at the Naval Training Center (NTC), San Diego. Specific samples used were as follows:

1. Item responses to GCT and MECH were obtained for a sample of 10,000 recruits and used to select items according to the four methods evaluated in the present study.

2. Item responses from two independent samples of 100 recruits were used for cross validation (simulated item administration). That is, although complete item response data were available for each of these groups, the data were used to obtain scores on each of the short tests constructed according to the four methods.

3. Two samples of 250 recruits in their third week of recruit training were administered short linear versions of GCT and MECH, one sample receiving items selected by SEQUIN, and the other items selected according to the HI VAL approach.

4. A total of 526 recruits between their third and fifth week of recruit training were administered items on computer terminals located at NTC San Diego. The Ss were randomly split into two groups (263 Ss in each group), one of which received WRIPA versions of GCT and MECH, and the other BRANCH versions of the same two tests.

Apparatus

BRANCH and WRIPA tests were administered on an IBM 1500 computer assisted instruction system superimposed on an IBM 1130 central processing unit. Thirteen individual test stations were used. Specific pieces of equipment used for administration of items and recording of responses included a 1510 cathode ray tube display unit with light pen and keyboard (located at each of the 13 test stations, and used to display items), a 2310 disc unit (for reading items onto the cathode ray tube), and a 2415 tape unit (for recording response data). An IBM 1518 typewriter was located at the proctor station to indicate when subjects began and completed the tests, as well as to indicate any malfunctions during the testing.

Procedure

Initial Item Selection and Scoring. Shortened versions of GCT and MECH were constructed according to the following four methods:

1. HI VAL. Item validities for predicting total test score were obtained for each of the 100 items in the GCT and the 100 items in MECH. The entire sample of 10,000 recruits was used for item analysis and selection. In each case, the five items showing the highest point biserial correlation with total test score (irrespective of item difficulty) were selected to comprise a short test. Hence, a five-item GCT test and a five-item MECH test were constructed. Scoring was accomplished by simply summing the correct responses.

2. SEQUIN. Five-item GCT and MECH tests were also constructed according to the SEQUIN procedure. Previous SEQUIN analyses (Swanson, 1968) based on samples of 1000 recruits were used for selecting items. This approach involves selecting in sequence a series of items each of which would contribute maximally to the multiple R, but unit weighting each selected item to obtain a score which is used in computing the shortened tests' correlation with the total test score.

3. WRIPA. In order to obtain estimates of item difficulty and of the discriminating power of items at various ability levels, a program written by Wright and Panchapakesan (1969) was used. The

specific parameters obtained were log easiness and the slope of the item characteristic curve at the median response (i.e., the point at which 50 percent of the people pass the item and 50 percent fail the item). Items were selected from GCT and MECH to construct approximately symmetric distributions of log easiness estimates with approximately equal intervals between final difficulty levels. Each selected item was the one which showed the largest slope within this context. Because the resulting paradigms for both GCT and MECH contained items which were rather consistently easier than desired, one additional difficult item was selected for persons who answered all five items correctly.[5] To obtain an estimate of final score for each terminal point, the sample of 10,000 recruits was used. The sample was successively sorted into groups passing and failing each of the indicated items. Mean total test score for each terminal point was then obtained for all persons in the sample of 10,000. These means were subsequently used as "scores" for persons terminating at the various points.

4. BRANCH. Program BRANCH (Wolfe, 1970) was used to select items. The general procedure has been described previously. For the present study, the program was used as follows: Validities (point biserial correlations with total score) were obtained for all items using the entire sample of 10,000 Ss. The most valid item was used to sort the sample into those who passed and those who failed. Validities were recomputed for each of the two groups. The most valid item for each of these groups was then chosen and groups were again sorted--producing four groups. This process was continued until five items had been chosen for each person--producing 2^5 or 32 groups of persons. Mean GCT scores were obtained for each of these groups. (Sample sizes for groups ranged from 41 to 1813.) These means were used as expected values of GCT score for each terminal point. (In BRANCH, each terminal point represents a unique pathway through the items.) Items were then selected from MECH in the same manner.

Tryout Using Item Response Data Bank. Item response data to full length GCT and MECH tests for two samples of persons (N=100 for each sample) were used to simulate branching. Items selected by each of the four methods from each of the two tests were "administered" to persons in the two samples. Scores were determined and the resulting short test scores were correlated with total length test scores to permit comparisons of efficiency in replicating total test score.

Administration of Shortened Tests. Five-item SEQUIN and HI VAL versions of GCT and MECH were administered at NTC San Diego. (Additional items were subsequently presented but are not relevant to the present study.) The two SEQUIN tests were administered to one sample of 250 recruits and the HI VAL tests, to another sample of the same size. Four minutes were allowed for administration of each test. Test scores were obtained by simply summing number of correct responses. These scores were then correlated with total test score. (The full length tests had been administered three weeks previously during routine classification testing.)

WRIPA and BRANCH tests were programmed for computer terminal administration. All instructions and sample items (the same sample items as given with SEQUIN and HI VAL tests) were administered via the terminals. WRIPA versions of GCT and MECH were administered to a sample of 263 persons, and BRANCH versions of the same tests to another sample of 263. Responses were made by subjects touching the spot beside the correct response with a light pen. Responses, response latencies, item scores and expected value of total test score were recorded for each subject. If the subject had not responded after spending 45 seconds on each item, a time warning was given. Ten more seconds were then allowed and if a response had not been made by then the response was considered incorrect and the next indicated item was administered.

Groups of 13 recruits were brought into the Computer Assisted Instruction Laboratory at 30-minute intervals. The actual amount of time spent on the terminal ranged between 7 and 20 minutes. Scores on full length GCT and MECH tests (administered 2-4 weeks previously) were obtained for all subjects. Expected values of total test scores obtained from administration of the short branching tests were correlated with actual total test scores.

RESULTS AND DISCUSSION

The major hypothesis, that extremely short tests (5-6 items) can be developed and administered via computer terminal with little loss of information contained in the total (100 item) test, appeared to be supported when short GCT tests were developed and item administration was simulated. Perhaps because of the greater heterogeneity of MECH, 5-6 item tests were not as good as short GCT tests. When the short branching tests were administered via computer and the short linear tests administered in paper-and-pencil form, the branching tests offered no advantage over the linear tests. Furthermore, all short-test approaches resulted in greater information loss than was incurred with the simulated runs.

Simulated Administration

Obtained correlations of short test scores with long test scores (GCT and MECH) are listed in Table 2 for the two samples of 100 recruits. Base values for comparing these simulated runs with the expected value of the correlation of a random set of five items taken from the long test were established by using the Spearman-Brown formula. These were .66 for GCT and .40 for MECH. For both tests, all approaches represent substantial improvement over these base values.

TABLE 2

Cross-Validated Correlations of Simulated Short Test Scores With Total Test Scores, Using Four Item Selection Procedures and Two Tests

Item Selection Procedures	GCT		MECH	
	Sample 1	Sample 2	Sample 1	Sample 2
BRANCH	.95	.92	.83	.80
WRIPA	.90	.89	.69	.70
HI VAL	.87	.86	.67	.69
SEQUIN	.94	.89	.73	.70
N	100	100	100	100

The most consistent finding was that irrespective of method used for selecting items, correlations between short and long GCT tests are much higher than those between short and long MECH tests. It appears from these results that GCT could be substantially shortened to five or six item length without appreciable loss of information; but that MECH tests this short are not very satisfactory.

Of the four methods evaluated the HI VAL approach produced the poorest results on both GCT and MECH. Using HI VAL, the simulated short GCT test scores correlated .87 and .86 with total score and the short MECH test, .67 and .69 with total score.

Other comparisons among methods are less conclusive. It was expected that SEQUIN and WRIPA would be better item selection procedures than HI VAL and poorer than BRANCH. With respect to comparisons between SEQUIN and WRIPA, it was hypothesized that because GCT is an extremely homogeneous test and MECH is relatively heterogeneous, WRIPA would be better item selection procedure for constructing a short GCT test and SEQUIN for constructing a short MECH test. The actual findings suggest that SEQUIN is a slightly better procedure for constructing a shortened test even when test content is extremely homogeneous. That is, for both GCT and MECH short SEQUIN tests were slightly better than the respective short WRIPA tests.

Because program BRANCH successively selected maximally discriminating items for groups defined by patterns of previous item responses (i.e., utilized all available information in item selection) it was hypothesized that the short tests constructed by BRANCH would be superior to those constructed by any other method. For these simulated item administration cross-validations this was indeed the case. The correlations between BRANCH score and total score on GCT for the two cross-validation samples were .95 and .92 and, on MECH .83 and .80.

These simulated short test results suggest that if tests are to be shortened, very large samples are available for selecting items, and computer terminals are available for administering items, the BRANCH program should provide an excellent means of constructing tests to parallel the longer form.

Administration of Shortened Tests

Unfortunately, ambiguous results were obtained when the shortened tests were actually administered as such. The linear shortened tests were administered in paper-and-pencil form and branching tests were administered by computer terminal.

Correlations of short test scores with scores obtained on the total test (administered 2-4 weeks previously) are listed in Table 3. While SEQUIN was still a better item selection procedure than HI VAL and BRANCH better than WRIPA, previously demonstrated differences between BRANCH and SEQUIN were not maintained. In fact, results obtained with a five-item SEQUIN test were, for MECH, slightly better than those obtained using a five-item BRANCH test (r = .74 as opposed to r = .73).

For GCT, the results were equivalent (r - .83).

TABLE 3

Correlation of Short Tests
With Total Test Score

Item Selection Procedures	GCT	MECH	N
BRANCH[a]	.83	.73	263
WRIPA[a]	.79	.72	263
SEQUIN[b]	.83	.74	250
HI VAL[b]	.80	.73	250

[a]Administered on computer terminal 2-4 weeks after administration of total test.

[b]Administered in paper-and-pencil version three weeks subsequent to administration of total test.

These comparisons are critical, for the process of adapting tests for computer administration is a very expensive one which requires that definite advantages of this mode of administration be demonstrated. To the contrary, the present results suggest that as much information may be derived from a short linear paper-and-pencil test as from the more complex short branching test.

The fact that expected results were obtained with the simulated runs, but not with the on-line runs may possibly be due to either or both of the following factors:

Because of its use of successive sample splits for determining the sequential items to administer, BRANCH may capitalize on error to a much greater extent than SEQUIN. However, if this were the case, the discrepancy would also be expected to be apparent in the simulated cross-validation runs. Furthermore, it should be recalled that a very large sample--10,000 recruits--was used to select items for the BRANCH procedure, thus reducing the likelihood of capitalizing on chance.

A more plausible explanation is that the BRANCH correlations were substantially lowered because of the switch in mode of item administration.

This possibility is supported by the fact that when WRIPA tests were administered via computer terminal much poorer results were obtained than had been obtained by the simulated runs. While the Ss appeared extremely interested in taking the tests on computer terminals and there were no complaints about clarity of instructions, etc., the procedures represented a marked deviation from standard testing conditions.

In addition to the novelty of the equipment used to project and record responses, test content was rather subtlely altered. Each item was timed separately (very few items were unanswered) and no provision was made for returning to previously answered items. Furthermore, several items which appear toward the end of the long tests (both GCT and MECH) were selected for BRANCH. These may have been items which, more than anything else, discriminated between those who completed the long test and those who did not. With items timed separately for computer administration, all persons were exposed to all items; hence these items were probably less affected by speed factor.

Taken as a whole, the present study indicates that credence cannot be placed in results obtained from simulated item administration strategies if the purpose is to eventually produce tests to be administered on computer terminals. While lower correlations with total score were obtained with actual short linear tests than with the simulated linear tests, some decrement was to be expected because of the time span between the two administrations of the items and because of the changed context in which the items were presented. However, the finding that on-line administration branching tests are not better than their short linear (paper-and-pencil administered) counterparts suggests a large effect at least partially attributable to mode of administration.

REFERENCES

Anastasi, Anne. Psychological testing. New York: Macmillan, 1968, Chapter 7.

Bayroff, A. B., & Seeley, L. C. An exploratory study of branching tests, Military Selection Research Division, Behavioral Sciences Research Laboratory, U. S. Army, Technical Notice 188, 1967.

Cleary, T. Anne, Linn, R. L., & Rock, D. A. Reproduction of total test score through the use of sequential programmed tests. Journal of Educational Measurement, 1968, 5(3), 183-187. (a).

Cleary, T. Anne, Linn, R. L., & Rock, D. A. An exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-360. (b)

Dubois, P. H. Varieties of psychological test homogeneity. American Psychologist, 1970, 25(6), 532-536.

Linn, R. L., Rock, D. A., & Cleary, T. Anne. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29(1), 128-146.

Loevinger, J. The attenuation paradox in test theory. Psychological Bulletin, 1954, 51, 493-504.

Lord, F. Some test theory for tailored testing in Computer Assisted Instruction Testing and Guidance. (Ed. Wayne H. Holtzman) New York: Harper & Row, 1970, 139-183.

Lord, F. M., & Novick, M. R. (With contributions by Birnbaum, A.) Statistical theories of Mental Test Scores. Reading, Mass.: Addison Wesley, 1968.

Moonan, W. J., & Pooch, CPL. U. W. (USMC). SEQUIN: A computerized item selection procedure. San Diego: Naval Personnel and Training Research Laboratory, October 1966. (Research Memorandum SRM 67-8)

Stocking, Martha. Short tailored tests. Educational Testing Service, July 1969, RB-69-63.

Swanson, L. Unpublished SEQUIN analyses, 1968.

Wolfe, J. H. Specification for program BRANCH, unpublished Memorandum, July 1970.

Wright, B., & Parchapekesan, N. A procedure for sample free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

FOOTNOTES

[1]This research was supported by a grant from Office of Naval Research, Contract Number RR 006-04-01, Task Number 150-325, and conducted while the author was at U. S. Naval Personnel and Training Research Laboratory, San Diego, California.

[2]This paper does not reflect the official policy of the U. S. Navy.

[3]Figures depicting the branching paradigms for each method and characteristics of the selected items may be obtained from the author.