DOCUMENT RESUME

ED 110 442 SP 009 433

AUTHOR Holland, James G.
TITLE Variables in Adaptive Decisions in Individualized Instruction.
PUB DATE Mar 75
NOTE 64p.

EDRS PRICE MF-$0.76 HC-$3.32 PLUS POSTAGE
DESCRIPTORS *Criterion Referenced Tests; Curriculum Development; Diagnostic Tests; *Individualized Instruction; Individualized Programs; *Instructional Materials

ABSTRACT

This study attempts to elucidate some quantitative measures to assess the adequacy of adaptive decisions in individualized materials. The primary purpose of the study is to improve the curriculum developer's ability to generate better adaptive materials by improving his judgment of the quality of the diagnostic portions of his material in meeting the objectives of adaptive instruction. Three measures of variables reflecting the rationale of adapting to individual differences are presented. These measures are: (a) ratio of teaching time to total time, (b) predictive validity ratio, and (c) discriminability ratio. The use of these measures are demonstrated with seven widely diverse examples of adaptive programs. Each of the three measures yielded a considerable range of values over the seven programs, but none of the programs proved adequate on all three measures of the necessary conditions for adaptive decisions. Although adapting instruction with prescriptive tests may continue to be widely used, there is not yet an empirical basis for that use. (Author/RC)

Variables in Adaptive Decisions in Individualized Instruction

James G. Holland

Learning Research and Development Center
University of Pittsburgh

The present study attempts to elucidate some quantitative measures to assess the adequacy of adaptive decisions in individualized materials. The primary purpose of this effort is to sharpen the curriculum developer's ability to generate better adaptive materials by sharpening his judgment of the quality of the diagnostic portions of his material in meeting the objectives of adaptive instruction. Despite the heavy emphasis over the past decade on prescribing materials adaptive to individual needs by diagnosing these needs through criterion-referenced tests (Glaser, 1963), the principles involved in preparing good adaptive materials have been left implicit.

Moreover, those who decide on the use of new materials need descriptive tools for determining whether the materials reflect the rationale in adapting. In the large, carefully controlled field evaluation (Office of Economic Opportunity, 1972), products of the emerging theory of instruction have failed to prove the worth claimed in theory for them. Such embarrassing failures are jeopardizing the further development of the theory of instruction. There is, however, the strong possibility that many, or even most, of the materials and procedures used failed to reflect the theory. Objective and quantitative measures of the key variables involved in preparation of educational materials are important to evaluator and developer alike if curriculum materials are to really reflect the scientific base usually claimed for contemporary instructional procedures.

The measures to be described here are derived from the rationale for adapting and an attempt to formulate rather direct, simple indices of the important variables. The general strategy adopted is similar to that followed in developing the black-out ratio (Holland, 1967) as a measure for programmed teaching items. The resulting measures should do for diagnostic items what the black-out ratio does for teaching items. They should provide a basis for research in adaptive decisions while giving clear guidelines to the developer of adaptive materials. Effort was made to develop measures which (1) assess the adequacy of diagnostic items in meeting the aims of adaptation, (2) are simple and easy to apply, (3) discriminate among programs which differ in the adequacy of adaptive decisions, and (4) are objective in that different persons using the measures will obtain the same results.

"Individualization" or "adaptive education" have become vogue terms which have occasionally been used to describe quite different things (cf. Cronbach, 1971). To some the terms connote the unstructured curriculum of the open classroom; and for others, they can mean individual choice of objectives. In this paper, the terms are taken to mean that individual differences in needs are diagnosed in an attempt to present each student with only those teaching materials he needs to reach proficiency in the terminal objectives of the course. Hence, the course objectives are the same for all students, but the student who is able to pass many diagnostic items skips much unnecessary teaching material, while the student who misses many diagnostic items must, as a consequence, get additional material sometimes identified as remedial.

Thus, adaptive materials have two separate components: test items which diagnose the student's need and teaching materials which fill that need. In Individually Prescribed Instruction (IPI) and most other adaptive materials, the two different types of items are clearly designated by the developer.

The present paper is addressed to criteria for diagnostic items used in adapting, not criteria used for developing the teaching items. There is already a set of well established principles as to how teaching materials should be designed and by what criteria they might be judged. Teaching items have a quite different function, and a different, even incompatible, basis for evaluating their worth than diagnostic items. Generally, whether or not the teaching material is in the old familiar formats of early programmed instruction, the principles embodied in its design are programming principles. Usually, the student's tasks follow some form of gradual progression. (Although in individualized materials, some effort is made to tailor this progression to the individual.) Individual teaching items are expected to evoke the desired, to-be-learned behavior before the student reaches a correct answer. Thus, the items should provide a low black-out ratio (Holland, 1967). It is anticipated that when the student reaches a particular level, he will be able to give the required performance since his answer insures he has performed adequately. Hence, good teaching material generally has a low error rate. The teaching item does not trap the student into errors or attempt to diagnose his deficiencies. Instead, its purpose is to evoke the new behavior so that it may be reinforced and established.

By way of contrast, individualization requires a quite different type of item. Test items serve a diagnostic function. They serve to differentially predict; different performance on a diagnostic item is used to recommend different learning materials. Therefore, considerations in test design are appropriate for these diagnostic items. First, to be useful, a diagnostic item must discriminate among individuals. A zero error-rate item would be worthless. A good diagnostic item reveals differences in performance with some students answering correctly and others making one or more types of errors. Thus, a good diagnostic item meets criteria incompatible with those met by good teaching items. It is for the special properties of the diagnostic process that new measures are here proposed and demonstrated.

The Measures

Adaptive materials characteristically (1) save the student's time and effort by letting him skip unneeded teaching material, (2) test each student to determine his needs, and (3) reflect individual differences among the students. These considerations suggest three important measures for the merit of adaptive tests. One reflects the potential savings in the student's time compared with the cost to him in time for the diagnosis. Another reflects the validity of prediction of the need for the learning material, and a third reflects the discriminability of the test. Simple indices of these three factors are proposed in the form of three ratios. The three indices will be called consequence ratio, predictive validity ratio, and discriminability ratio.

Consequence Ratio

Adaptive tests are designed to give the student the teaching material he needs without wasting his time (and patience) with material he does not need. But testing itself comes at a cost in student time. It would be inefficient to spend a lot of time testing to enable a student to skip only a short teaching sequence. The appropriate index is a ratio of teaching time to total time. The total time is the combination of teaching time and testing time. If a unit of teaching material requires 30 minutes to complete but is preceded by a 30 minute pretest that would enable a passing student to skip the material, then the cost to the student of being tested is as great as the savings he stands to gain by passing the pretest. The consequence ratio for this test would be the 30 minute teaching period divided by the one-hour total (30 minute teaching plus 30 minute testing) or .50. Clearly, no matter what other merits this test may have, it would be unacceptable since the passing student can only break even. If, on the other hand, a 1-minute test could be used to prescribe this same 30-minute teaching unit, the cost is small compared to the potential gain in passing and the consequence ratio is 0.97. If other necessary conditions are met, this would be a very worthy instance of adapting.

The phrase "consequence ratio" is used to avoid implying anything about the merit of the teaching material that follows. The consequence of a test being evaluated might even include further testing; for example, placement tests place the student in units which may include test items that permit "looping" past subsections of the unit. The size of the consequence is everything in the catchment area under the test in question. It should be apparent that it is the potential consequence which is of concern here. That some students skip the material does not change the potential consequence of failing the test; it is, rather, the point. The consequence ratio addresses itself to the size of the cost in time (or amount of material) saved compared with the total amount of time (the test plus the consequence).

Predictive Validity Ratio

The validity of a diagnostic item is the extent to which the item correctly predicts the need or lack of need for some teaching material before a posttest is taken to measure the same competence. The adequacy of such a prediction can be measured simply by giving first the diagnostic test and then the criterion (or posttest) without giving any instruction between the two tests. Poor performance on the diagnostic test predicts poor performance on the criterion test unless instruction is received. Likewise, good performance on the diagnostic test predicts good performance on the criterion test even when instruction is omitted.

In form, this procedure may sound to the reader like a reliability measure because it involves prediction by one test of performance on another parallel test. However, because performance on the criterion test is the targeted performance, validity seems conceptually correct.

When no instruction is provided between tests, the diagnostic test predicts correctly or "hits" when comparable material is answered correctly both on the initial and subsequent tests or when comparable material is answered incorrectly on both (see Table 1). A student passing a diagnostic test is expected to be able to skip the teaching material and pass the criterion test while one who fails the diagnostic test needs the teaching material and without it should fail the criterion test. Failures of prediction, or "misses," occur when the student is correct on the diagnostic test but incorrect on the criterion test or incorrect on the diagnostic test but correct on the criterion test.

Insert Table 1 about here

With the test and retest procedure, with no intervening instruction, the predictive validity measure is based on the ratio of hits to total number of decisions. If everyone who passed a diagnostic test also passed the criterion test and all who failed the diagnostic test also failed the criterion test, then the ratio of number of hits to total number of decisions would be 1.0. If, on the other hand, tosses of a coin were used as the diagnostic test, these chance decisions would lead to half or a 0.5 ratio of hits to total number of decisions. Most tests will fall between these two extremes; for example, with a ratio of 0.75, one quarter of the students either were unnecessarily assigned teaching material or directed to skip material they in fact need. Such a low value for the validity ratio would presumably be acceptable to the developer or the user only if the consequence were very large compared with the time needed to complete the test. Ordinarily, one should expect validity ratios close to 0.90 or better.

Discriminability Ratio

In an exploratory effort applying the consequence ratio and the predictive validity ratio to several sets of curriculum material, the need for this third measure became apparent. There are instances in which virtually all students answer diagnostic items correctly and others in which all answer incorrectly. In either case, all students receive the same prescriptions; therefore, the programs are not adaptive because the tests detect no individual differences to accommodate. A simple ratio of those passing the diagnostic items could be used; both 0.0 and 1.00 (no one passing and no one failing, respectively) would represent the extremes in lack of discriminability (0.50 would be the optimum). But since the three proposed measures would usually be discussed together, this was rejected to avoid having a 1.00 as the ideal for both consequence ratio and the validity ratio but as the poorest possible value for the discrimination ratio. Therefore, it was decided to express the discriminability measure as the ratio of the number who either passed or failed, whichever is smaller, to the total number taking the test. The discriminability ratio, then, varies from 0.0 to 0.50. It is zero if either all students pass or all fail; half passing would give a ratio of 0.50; one quarter passing (or one quarter failing) would give a ratio of 0.25.

It is clear that when there is no discriminability, that is, when the ratio is 0.0, the materials are not adaptive to individual differences because the test reveals no differences. Beyond this, there is no absolute minimum acceptable value; but a ratio as low as 0.10 would presumably be useful only if the test is highly valid and the consequence very large. Ordinarily a ratio of approximately 0.25 would be adequate if both validity and consequence are at least fairly large.

Use of the Measures

These three indices are quantitative measures of three variables involved in goodness of adaptive decisions. Adequacy on each of the variables is a necessary condition to meet the rationale of adaptive instruction. Excellence in any one is not a sufficient condition. Complete inadequacy for validity, consequence, or discriminability renders the diagnostic procedure worthless for individualization regardless of the value of the other two. On the other hand, there are no fixed, all-purpose values that can be regarded as acceptable. The ideals are clear, as are the values indicating extreme failure. Between these extremes experience with the measures will be required.

It must be clearly understood that these measures do not evaluate the overall usefulness of any set of curriculum material. The technique for such evaluation is well known and involves measurement on criterion measures _after_ the students have used prescribed teaching material. The present three measures in no way evaluate the teaching material. They are, rather, measures of the adaptive characteristics of the program and not measures of characteristics of the teaching material. Neither are they measures of the achievement to be expected from using the materials. It is possible for the adaptive testing to be excellent and for the curriculum to fail to teach. It is even possible for the adapting characteristics to be poor and the overall usefulness of the material to be very good; although, in this case, the overall worth of the curriculum material would probably be improved by correcting the deficiencies in the diagnostic testing and adapting procedures.

Experimental Demonstration of the Measures

The actual use of these measures will demonstrate their utility in revealing to the developer and consumer strengths and weaknesses in the diagnostic and adapting procedures. Segments of seven different sets of adapting materials were evaluated using these measures in order to determine the practicality and usefulness of the measures. Instructional programs were chosen to cover a variety of adapting styles. Chosen materials included two LRDC courses (Science and Math), a program with large loops, one computer-assisted instruction program involving very fine adjustments, a remedial math program with only an overall placement test, one example of a Crowder-type "intrinsic" program, and a linear program with a binary decision tree enabling initial placement. Since only small segments were tested, the results should not be taken as necessarily indicating the quality of the adaptive test material through the entire program even though an effort was made to choose representative units.

Ideally, published programs would include sufficient data to estimate these three indices. Unfortunately, of the six programs receiving classroom use, none gave information of any kind on the validity of the test items; and only two gave data sufficient to estimate the discriminability of the items. These same two gave information necessary to make an informed estimate of the relative amount of time necessary for teaching components and for testing components (although all but one did give teaching time information). Therefore, it was necessary to gather data sufficient to calculate the three ratios using students from the populations appropriate to each program so far as was practical. The techniques of measurement are described for each of the seven course segments. Among the seven a variety of problems are revealed and recommendations emerge illustrating the val [illegible] th. measures.

Job Corps Advanced General Education Program

The Job Corps Advanced General Education Program (Office of Economic Opportunity, 1968) is a self-instructional program designed for Job Corpsmen who have high school reading skills but who have not finished high school. The course covers everything necessary to take and pass the GED test resulting in a high school diploma. The course as a whole consists of 124 lessons divided into three levels. The lessons are grouped into units consisting of 2 to 15 lessons per unit (see Figure 1).

Insert Figure 1 about here

Each of the 24 units is preceded by a screening test. A score of 85 percent or better on the screening test enables the student to skip all of the lessons in that unit. Unit II-2, a unit with the average number of lessons, was chosen as the portion of the program to demonstrate the three indices of adapting quality.

This particular program was of interest for several reasons. First, as shown in Figure 1, it has a classic adapting structure with unit tests that enable the student to 'loop' over some portions of the teaching material when test performance indicates that he does not need the material. Moreover.

this program was known to be of high overall quality. While this es not imply that the adaptive features are good, the tests seemed of adequate length to be valid and yet short enough to be efficient.

The six lessons required, according to the Teacher's Manual, 510 minutes. Thus, passing this unit's screening test saves a student 510 minutes of work. It required only 14 minutes on the average for the 28 college sophomores serving as subjects to complete the screening test. Therefore, the consequence ratio was an excellent 0.97. The lower educational level of the targeted high school dropout could make a difference, but even with the testing time doubled, the ratio would be 0.95.

In estimating the validity, these subjects first took the screening test and then, without intervening learning material, took the posttest. A failure on the screening test forecasts failure on the posttest if the student has not taken the prescribed teaching material. Similarly, a pass on the screening test predicts a pass on the posttest. A failure of prediction would occur if the subject passed the pretest and failed the posttest/or failed the pretest and passed the posttest. Table 2 shows the its and misses in prediction for these data. The predictive validity

Insert Table 2 about here

ratio is 0.36. In other words, the prediction is somewhat less accurate than simply flipping a coin to decide whether or not the student should skip the unit. This program is totally inadequate in the validity of the test. The cause of the problem is apparent on examination of the pretests and posttests. Throughout the 24 screening tests in the program, factual information is tested; and, in the case of our subjects, none was able to attain a passing score on these factual questions. However, the posttests throughout the program (and in the GED test itself) test not factual information, but reading comprehension in the designated subject area.

All subjects performed identically on the pretest; they all failed and presumably the high school drop out would do no better. Thus the diagnostic test is non-discriminating; the discrimination ratio is 0.0. Because all students would be given exactly the same prescription (i.e., take the six lessons), this material is not, in fact, individualized.

The present author still considers this program to be of high quality. The teaching material does the job it was designed for and graduates of the program are able to pass the GED test, as shown by the data of the developers and by this author's own direct experience. But the quality of the teaching material is not in question here. The present study deals only with measurement of variables in adapting to individual differences. If all of the unit screening tests perform as badly as the one tested in this study, then it would be clear that the adaptive feature is not doing it's job. Although it is excellent to have tests that are small and manageable in comparison to the size of the consequence, the pretests seem to test for something different than the terminal behavior reflected in the posttest. There is then no basis to continue to use the screening tests because of the low validity and low discrimination. If the developers had had at their disposal the indices recommended here, corrective measures could have been taken. It would certainly seem ill advised not to adapt a curriculum covering the totality of high school. Nevertheless, testing in the present form is a complete waste of time since the validity index shows that, at least for unit II-2, prediction is below chance in accuracy and the discrimination index shows that the pretest fails completely to discriminate.

These outcomes, if general over the whole program, suggest two recommendations. A revision of this program should include redevelopment of the screening tests to better predict the terminal behaviors shown in the posttest. Second, anyone now using the program should stop using the screening tests and either give all students the whole course or administer the present "posttests" before the unit as screening tests since the posttests and the GED both measure reading comprehension.

Programmed Reviews of Mathematics (Flexer & Flexer, 1967)

This is a program in remedial mathematics for college students who have had the typical mathematics background required of entering college students but who are now beginning a science course requiring use of math. Flexer and Flexer indicate that many students are unprepared to handle the mathematics in a typical basic science course. They prepared six short, remedial books each of which can usually be completed in one to three hours. Each book has a placement test to diagnose the student's need for remedial work in the area of mathematics covered by the book.

The Flexer and Flexer program is useful for the present study for several reasons. First, the problem they address seems especially likely to provide important advantages of adapting to individual differences. All of the students supposedly have learned all of the mathematics covered, but the experience of college science teachers is that a sizable percentage of their students lack the basic mathematics necessary for lab work. Abraham Flexer was motivated by the desire to avoid spending weeks of class time in a biology course teaching math to those students who need it and thereby depriving those who do not of the opportunity to proceed with the intended contents of the course.

Second, the present author was well acquainted with this program because it was developed as a project of an organization directed by him (the Harvard Committee on Programmed Instruction). To the author's knowledge, Flexer and Flexer were aware of the requirements of good adaptive test materials. They knew the need for correctly assessing the individual s need as efficiently as possible and the need to discriminate between students who did and did not need special work in math. In short, this program was chosen because it should be exemplary on all three variables.

Each of the six programs is published in a separate booklet and includes a considerable amount of data from the several test-runs of the material at Harvard University and Emmanuel College in biology, chemistry, psychology, and sociology courses. Much of the data is concerned with the teaching material and the gains produced by the course, which are, of course, the proper emphasis for program evaluation. They also include ample data on teaching times and at least enough data to estimate the consequence ratio for the program as they tested it. Discrimination ratios are also reported as the percentage of each class which passed each test item and hence was excused from using that portion of the program. Unfortunately, they did not test for the validity of the test items.

To obtain estimates of validity for the present study, a group of undergraduate psychology students were administered the test materials for one logarithm unit and for the three fraction units. The tests used were not the single items on which the programs were originally evaluated, but instead, were the tests provided in the introduction of the books which contained from five to eight items per decision for the four decisions evaluated. The criterion for a pass in each case allowed for one incorrect answer in each set.

Surprisingly, this was not the form of test used in their original testing of the program. They had tested the whole class at the beginning of the term with a placement test having a single item for each separate diagnostic decision. They gave no reason in the published version for changing from the single item to the several item test. Perhaps it was an effort to increase the validity or perhaps it was on the advice of the publisher who may have felt that a slightly longer test would have better face validity and thus be better for marketing. Nevertheless, it seemed the proper course to apply the measures of the adequacy of adapting to the final published long test form.

This decision led to a serendipitous result. The outcome for the longer test is considerably different than for the shorter test. Changing the test in a way that superficially would seem likely to improve it, instead, when empirically evaluated, is shown to have seriously flawed a previously excellent program.

In the first effort to apply the measures, 28 students took the eight-item pretest for the logarithm unit (the first of three decisions for the logarithm book) and one week later repeated this test without, of course, using the program. Similarly, 10 students took the three pretests for the three parts of the fractions program (the lengths of these were four, five, and six items) and retook the tests one week later. For all sets of tests, records were kept of the time required for each student to complete each test. In calculating the consequence ratio, the published times for the programmed materials were used. Table 3 and 4 indicates the results of these evaluations. Both show reasonably high validity ratios (0. 93 for logarithms and 0. 83 for fractions) and fair consequence ratios (0. 88 for logarithms and 0. 82 for fractions). Surprisingly, however, the tests for both programs showed poor discrimination. Of the 28 students taking the logarithm test, 26 failed and only 2 passed for a discriminability ratio of

---

Insert Tables 3 and 4 about here

---

only 0.07. Of the 30 decisions in the fractions program, only 5 were passed for a discriminability ratio of 0.17. These discriminability ratios were far from the values indicated by Flexer and Flexer for the percentage passing the various tests with one item per decision. Unlike the present results with the longer tests in which the bulk of the students failed, they found many single item tests were passed (62 percent for the fractions tests yielding a 0.38 discrimination index). The combination of the validity data measured in the present study which was unavailable in the Flexer and Flexer data and the consequence ratio and discriminability from Flexer and Flexer's data suggest that this program is excellent in its overall ad_pting characteristics. However, the very low discriminability obtained in this study indicates that the recommended long form of the test has largely ruined the adaptive feature of the program.

To determine whether or not the unexplained change in the tests had this effect, the fraction tests were administered to another set of ten psychology students. It was possible to identify a single test item in each of the three fractions tests which was like that used in the original single item test. Using these items, an evaluation was made for both the single item and the longer test form with the same subjects. It can be seen in

Insert Table 5 about here

Table 5A that the results with the long form of the tests replicate the first set of data reported in Table 3. For the long tests a good validity ratio (0.83) and a goód consequence ratio (0.83) is to little avail in view of the poor discrimination ratio (0.17) caused by the bulk of the outcomes being failures. On the other hand, as shown in Table 5B, using the single-item test, as Flexer and Flexer did originally, provided very good discriminability (0.40), (quite close to their reported 0.38 ratio) with many passes. The single-item test also, of course, increased the consequence ratio to an excellent 0.96. However, as one might expect, the short test does have a lower validity ratio (0.73).

These results dramatically illustrate the merit of gathering the data needed for these three indices of the goodness of adapting. An originally tested version adapted fairly well to individual differences with excellent discriminability and a sizable gain for passing the diagnostic test. Flexer and Flexer may have had some indication that the predictive validity was low, although in combination with the good values for the other two variables the original form was useful and acceptable. It may have seemed prudent and safe to lengthen the test somewhat to increase its validity. Surprisingly, this created a serious deficiency in the program which apparently went undetected. An adequate level is necessary on all three measures. Each is necessary, no two of them sufficient. Moreover, a step taken to improve one could cause deterioration in another. Lengthening a test could reasonably be expected to improve validity, but it will also take more student time and a high validity can be obtained by using pre-and posttests which almost all will fail (or everyone will pass), but such a test has low discriminability.

LRDC's Individualized Mathematics

The classic program in the field of individualized instruction is LRDC's Individually Prescribed Instruction in Mathematics, or IPI Math (Lindvall & Bolvin, 1967). This set of curriculum materials has served as a prototype for individdually prescribing instructional units through diagnostic testing. The program contains 359 instructional objectives from 10 learning areas. The objectives are subdivided into 7 graduated levels of difficulty called A through G, corresponding roughly to a conventional kindergarten through sixth grade math curriculum. There is at each level a placement test which diagnoses the need for each unit appropriate to that level. The placement test indicates which units may be skipped, and which units the student should be pretested on. The pretest for each unit identifies more specifically within the area which lessons or "skill booklets" the student should use. The skill booklets contain the educational material, but they also have additional testing material, the curriculum-embedded tests (CET's). These diagnose the subject's mastery of that lesson and indicate his readiness to take still another test, the unit posttest.

Besides the prototypical nature of the IPI Math program, a second important consideration suggested its use in the present study. The layering of different tests presents interesting problems in evaluating, separately and in combination, the various elements of a compound diagnostic system. For example, the consequence of passing one item on a placement test is to skip not only all teaching material in the catchment area under that item, but to skip pretest materials, curriculum-embedded tests and posttest materials, as well. Hence, placement test items may have a large consequence, although much of the consequence is additional testing. On the other hand, to evaluate the overall testing structure, the total set of tests could be set apart from the teaching material to determine the cost effectiveness of the total testing structure for a given unit of teaching.

A unit of level B was chosen for analysis. Twenty-two children from a second grade urban classroom were routinely given the placement test at the beginning of the school year in September. In this instance, data were collected on the time required to do the test (33 minutes to 2 hours, with a median of 1 hour, 20 minutes). The questions for each of the 10 areas were separately analyzed, and the multipliaation unit was chosen for further analysis based on the high degree of discriminability shown by the placement test items. (Eleven students passed and eleven failed this unit for a 0.50 discriminability ratio.) It should be noted that, although the most discriminating portion of the placement test was chosen, the average discriminability for all units represented on the placement test was rather good, with an overall index of 0.35.

All students were then given three test packages, the pretests, the CET's, and the posttests, for all four skills in the multiplication unit. These tests were administered before any additional learning material and regardless of whether they had passed or failed the multiplication items in the placement test. A comparison among these tests provides the index of validity for the tests. Completion times were measured for each of the tests. Later, when the children came to the designated units in the curriculum, estimates were made of the teaching time. The teaching time for each child was based on the number of days spent working on these math units and the length of the scheduled daily math time. Although it might be argued that this is a realistic way to measure the time because it is the way the material is used, potential distractions in classroom use make it crude. For this reason; an additional method was used to calculate "time". Since test items and teaching items in IPI Math are similar in form, content, and length, "time" was estimated simply by counting the number of items used for teaching and the number of items on each of the various tests. The results of these two methods of estimating the consequence and costs of testing corresponded closely.

The variables in adapting decisions can be measured separately for each different test or for combinations of the tests. Four separate evaluations seem particularly of interest:

(1) evaluation of the placement test in terms of the savings in passing the test for the student who would, by passing, escape all of the work under the catchment for given items in the placement test, including the teaching material, pretest, CET's and posttest.

(2) evaluation of the placement test and the pretest together in terms of the validity of the two tests combined and of all the consequences

below the combination of the two tests, namely, the teaching material, CET's and posttest,

(3) evaluation of the pretest alone.

(4) evaluation of all tests in combination.

Placement test alone. There are only five placement test items diagnosing the need for the multiplication unit. The combination of all the consequences of failing these five items, including pretest, teaching material, CET's and posttest, is 375 items, giving a consequence ratio of 0.99. Thus, the savings to the student of passing the placement test are considerable; the savings are not simply that he skips teaching material but that he skips additional test material as well.

Validity of the placement test items was evaluated by comparing how well the placement test predicted performance on the CET's for each of the four skills in the multiplication unit. The frequency of hits and misses is indicated in Table 6. Passing the multiplication items on the placement test enables the student to skip all four multiplication skills. Hence, four "decisions" are made. With CET's from the four skills for 22 students, there were 58 hits out of a total of 88 predictions for a ratio of 0.66. As indicated earlier, about half of the placement decisions were passes and half failures for a perfect discriminability ratio of 0.50.

Insert Table 6 about here

Placement plus pretests. The next question to be considered is the use of the placement test and the pretest in combination to predict whether the individual lessons are required. With the combination of the two tests discriminability remains high with a ratio of 0.40 (see Table 7) and the addition of the pretest lowers the excellent consequence ratio only slightly to 0.93, with a consequence of teaching material, CET's and posttests as a cost of placement test and pretests.

Insert Table 7 about here

Table 7 presents the possible combinations of passes and failures on the placement test and the pretest. For each of the combinations, a prescription is derived, and this prescription is evaluated as a hit or a miss based on the possible outcomes on the criterion test. In the last column of Table 7, the data obtained from our test subjects are presented according to outcomes on the placement test, the pretest and the criterion test. The predictive validity depends on the performance of each student on the combination of the two tests. In normal use of the materials, a student who passes the placement test will not receive the pretest, since he is passed out of the multiplication unit. Therefore, in this study, in which subjects take all the tests and no teaching material, if the placement test is passed predictive validity is calculated without regard to pretest outcome. As shown in Table 7, if the placement test is passed, there is a "hit" if the CET's are also passed, and a "miss" if the CET's are failed regardless of whether the pretest was passed or failed.

Normally, students who fail the placement test later take the pretest, and if they also fail a given pretest, they must use that skill booklet; in other words, it is predicted that without the teaching material they would fail the CET's.

If a test subject fails both the placement test and the pretest, the resulting decision is a "hit" if he also fails the CET's and a "miss" if he passes the CET's. If, on the other hand, a subject fails the placement test but then passes the pretest, the prediction is that he should pass the CET's since his classroom counterpart using the teaching material would not be prescribed the unit. Those test subjects taking all the tests but given no teaching materials who fail the placement test and pass the pretest yield "hits" of prediction if they pass the CET and "misses" if they fail the CET's.

The results of this double level of testing are surprising. Using the two tests in combination gave only 59 hits out of 88 decisions, for a ratio of 0.67. The combination of the placement and the pretest in this instance provides a negligible improvement of only 1 hit in prediction over use of the placement test alone. Neither the placement test alone, nor the combination of the placement test and the pretest provides a very adequate prediction when chance assignment would give 50% hits.

Pretests alone. Next the pretest alone was evaluated as though there were no placement test. The validity of the pretest was evaluated against two different criterion tests, the CET's and the posttest. The results of evaluating pretest validity against the two tests are given in Table 8. Almost identical validity ratios, 0.86 and 0.85 were obtained; these provide useful levels of validity. Moreover, with ~~37%~~ 35% of the students passing the pretest, the discriminability of the test is good. The consequence ratio for the pretest alone is 0.94; but the consequence in this instance includes not only the teaching material but the CET's and the posttests.

Insert Table 8 about here

Overview: All tests in IPI math. Either alone or in combination the placement test and the pretests give good consequence ratios. However, there is considerably more testing in IPI math. The CET's and posttests are part of the consequence of failing the placement test or the pretests. But with the complete program the question becomes: What is the ratio of the consequent teaching time alone compared to the total time for teaching plus all testing? This consequence ratio, evaluating the totality of the testing, is 289 teaching items compared to a total of 379 items for teaching and all testing, for a ratio of 0.76. Since the break-even point on testing and teaching is 0.50, this ratio of 0.76 is rather disappointing. Individually the tests in IPI Math have satisfactorily large consequences but, in combination, testing is overdone.

The CET's and the posttest are of little use in diagnosing individualized decisions. They follow the teaching material and, unless the teaching material is inadequate, these items should be very poor discriminators since most should be answered correctly. However, some form of testing after teaching is no doubt needed to avoid misuse of the teaching material in a classroom situation. It is doubtful, however, that both the CET's and the posttests are needed.

It is especially interesting that the combination of placement and pretest shows no improvement in validity over the placement test alone and; for the multiplication unit in level B at least, the pretest alone is the more valid test. Anyone revising IPI Math, or attempting a similar curriculum, should avoid the layering of test upon test. Given two tests of known validity the higher validity test should be used. Validities are not additive. With the combined placement and pretest procedure a pass on either test prescribes a skip for the unit, thus the lower validity test degrades the prediction. The false skips of the two tests combine to lower the validity below that of the more valid test.

Users of the present IPI Math would be well advised not to use all of the tests. It is interesting in this regard that Leinhardt (1974) found for the IPI Math course a negative correlation between the amount of testing done by various teachers and the student achievement at the end of the school year. Knowing the overall relative validity of the placement and pretest would help in choosing between pretests and placement tests if only one were to be used. If diagnosing the student's needs is the only consideration, it would seem reasonable to use pretests only, abandoning the placement test, CET's and posttests. But the most satisfactory solution would be a new effort by the developers of IPI Math to revise their tests to gain validity greater than that of the present pretest while greatly lowering the overall burden of testing.

Individualized Science

The Individualized Science Curriculum (IS) (Klopfer, Champagne, & Pittman, 1972) is another well-known LRDC individualized program considered by this author to be of high overall quality. IS is of special interest because, although it is individualized, diagnostic testing is much de-emphasized as compared with IPI Math. The only tests normally needed in Level B, for example, are the unit pretests. Passing a given section of a unit pretest permits the student to skip that lesson in the unit.

In this study one unit test (The Hooke Unit for Level B) was used with ten students from a school using IS. Since there is no posttest, predictive validity was measured by administering the pretest twice and predicting test performance of the second testing from the first testing. As always in these validity checks the teaching materials were not used between the two administrations of the test. The subjects were first tested in the summer a few weeks before the start of school and again about two months later when they took the unit pretest as a regular part of their classroom activity. Good performance on the Hooke pretest could exempt a student from six lessons, but two of these lessons were under the catchment of the same pretest questions. Therefore, five individual decisions were evaluated for each of the ten subjects. The six lessons required an average of 116 minutes and the pretests required an average of 11 minutes, giving an excellent consequence ratio of 0.91. The validity ratio was 0.98, reflecting the 49 hits and 1 miss summarized in the hit-miss chart in table 9. Table 9 indicates that the high predictive validity results from all fifty test decisions being failures on the first testing and all but one being failures on the second testing. Thus, the test failed completely to discriminate.

Insert Table 9 about here

Since this unit pretest lacks discriminability, the material does not adapt to diagnosed individual differences. Since all fail the pretest, all subjects would have received identical prescriptions. Despite a very high consequence ratio and a very high validity, the adapting procedure is inadequate because it fails to discriminate. Adequate values for each of the three variables are necessary for an adapting system; none is sufficient alone. If this failure of discrimination is characteristic of the whole science program, the user should abandon pretesting and either use it as a linear program or choose lessons based on student interest or teacher objectives.

The developers seem to have concentrated on the teaching material. In doing so, they produced an interesting and useful science curriculum. The failure of the diagnostic procedure, even if characteristic of the whole curriculum, does not negate the value of IS as teaching material. There is no empirical, logical, or compelling intuitive reason to believe that diagnostic testing and individual prescriptions will be particularly useful in all areas of instruction. Possibly the developers of IS had doubts about the importance of adapting in this curriculum and de-emphasized it. But without discriminating tests their exciting instructional material would be improved by dropping testing altogether. Otherwise, discriminating tests are needed.

Inductive Reasoning Program

The inductive reasoning program is a 256-item linear program constructed as an experimental demonstration of the teaching of a basic aptitude--Thurston's reasoning factor (Holland, 1962). Program items consisted of a row of "bottle-shaped" objects varying in color and direction which were arranged to provide patterns. The student picked from among five alternatives the object which would be next if the pattern were extended.

For an experiment on evaluation of branching effectiveness (Holland, Hoffman & Doran, 1972), a binary search sequence of items was added to provide a maximally efficient way to place a student at his proper beginning point in the otherwise linear sequence. The binary search procedure placed students by beginning with the middle item of the program and bisecting distances forward or backward after correct or incorrect responses. After the seventh choice, the student was considered to be at his correct beginning point (See Figure 2 for a graphic representation of the binary search procedure). All data needed to estimate consequence, validity,

----

Insert Figure 2 about here

----

and discriminability ratios are available from the study. For evaluating whether each individual decision in the branching sequence was a hit or a miss, similar program items in a pretest used in that study served as a criterion test.

Insert Table 10 about here

The consequence ratio for the seven-item test with a 256-item consequence is 0.97. With eleven subjects and seven decisions each, the total number of decisions in the program was 77. With twenty-five failures and fifty-two passes, the overall test is discriminating. However, only thirty-six of the total number of decisions were hits as compared to forty-one misses for a validity ratio of 0.47 (see Table 10). Thus, while the branching tree might be an efficient use of testing time, the use of single multiple-choice items failed to be better than a flip of the coin so far as validity was concerned.

This program illustrates the problem of test size. A short test may be very efficient in terms of time, but short tests tend to be less valid. Increasing test length can increase validity, but a longer test decreases the consequence ratio. The inverse relationship between test length and adequacy of adaptive testing poses a special dilemma for the curriculum developer.

A Tutor Text Program

Especially popular a decade ago, tutor texts have been prepared in a wide variety of topics from bridge to electronics. Often they were intended for popular consumption and sold in the trade market but some were prepared for college courses and technical training. In these programs each page has some material to be read followed by a multiple-choice item with two to four choices. Each potential choice directs the student to a particular "next" page which itself has some material to read and other multiple-choice items. A correct answer usually keeps the student on the mainline items while incorrect answers loop him through remedial material of one or a few items and eventually back to the mainline items.

A representative of this type of program was included in the present analysis for several reasons. First, it would be unthinkable not to have an example of Crowder's "intrinsic" programming technique which for some years was the leading example of adapting to each student's special needs. The approach is also of interest because it represents a rather fine-grain approach to teaching and testing. A single page, and often much less than a single page, includes one unit of teaching material and one test item. Hence, the student is diagnosed as to his ability to handle the next small mainline step or his need for a short remedial loop. In addition, this is an instance in which the teaching material requires no overt responding by the subject. All answers are to the diagnostic portions of the material. Since teaching and diagnostic material are so closely intertwined, it might seem to the casual observer that intrinsic programs could violate the assumption that teaching and testing can be separately identified. However, this intertwining offers no problem in practice and it would seem to offer no problem in theory in view of Crowder's description:

> Intrinsic programming assumes that the basic learning takes place during the student's exposure to the new material. The multiple-choice question is asked to find out whether the student has learned; it is not necessarily regarded as playing an active part in the primary learning process. (Crowder, 1962, p. 3)

An evaluation was made of the test elements in the seven mainline items in Chapter 5 of A tutor text on the arithmetic of computers, (Crowder, 1960). Testing for validity and discriminability required use of a test-retest procedure because there was no other appropriate criterion test. The test elements are spread through the program making it necessary for each of the nine college students serving as subjects to use the material leading up to each test element before answering the test item. After the first testing for each item, six hours to one day elapsed before the retest for that item and the use of the prescribed teaching material which included the next mainline item and the first testing for the next item. The cycle continued in this form until all seven mainline test elements were completed. The reader is reminded that no teaching material was taken between the test and retest for each element. The reason teaching material was necessary in this case, unlike any others in this study, was that the test elements were intended to predict the student's next need given exposure to the preceding teaching material which was itself preparatory for the test item

---

Insert Table 11 about here

The outcomes of the three indices are summarized in Table 11. The bulk of the test elements was passed on both test and retest giving a high 0.97 validity ratio. The retest procedure no doubt exaggerates validity as compared with use of a parallel test, but in this instance this validity problem is overshadowed by a discrimination problem. The discriminability ratio was an unsatisfactory 0.12 because usually the correct choice was made and consequently the mainline item prescribed. Used in the designated way, remedial pages would be used by few, if any, students. A paper shortage may correct this problem even if quantitative evaluation is ignored.

In failing a test element the student gets one or another remedial item and is then sent back through the mainline item again. Therefore, the consequence ratio is the average time for a single route through, including retaking the mainline teaching and testing. The consequence ratio is 0.75 which is very poor especially since part of the consequence is additional testing. With the extra testing removed, the ratio is only 0.65 merely 0.15 above the level in which test and consequence equal each other.

The poor consequence ratio seems endemic to the tutor text format with a little teaching and a little testing on each page. Moreover, a combination of adequate validity and discriminability is unlikely with one multiple-choice item. In this instance the high validity in this program resulted from the low discriminability, in that errors were so infrequent on either testing. On the other hand, if items are written so more students fail (providing better discriminability) chance choices would appear in the multiple choice format and with this guessing, validity would suffer. Considered as adaptive material, there is little to recommend Crowder's intrinsic programming. Nevertheless, these programs sometimes are fun. People even enjoy peeking at the error loops for errors they did not make. If so, these programs may be useful, but not because they are adaptive to individual differences.

Stanford CAI Reading

Atkinson's beginning reading program is one of the better known model programs in CAI. His article describing it and reporting data (Atkinson, 1968) added momentum to the use of computers for a fine-grained adaptation of teaching material to student needs. This program was chosen here to represent the extreme in detailed adapting. This is one of the frequently claimed possibilities offered by the use of computers. To explore the applicability of the present measures in this type of adapting it seemed reasonable to choose from among the most respected of models of CAI.

In the Atkinson reading program the child views a cathode ray tube which presents letters and words. A random-access audio device presents messages and the student places a light pen on the screen to indicate his choice among alternative answers presented by the cathode ray tube. One of three basic forms, "matrix construction," was used in this analysis. This form is the program's key format for teaching decoding of graphemes to phonemes. In a typical mainline item the child is presented a letter ("r") to the left of an empty cell and a vowel-consonant ending ("an") above the cell. Below the cell are four alternative words ("rat", "bat", "fan", and "ran"). He hears the automated message "touch and say the word that belongs in the empty cell." This mainline item is diagnostic, according to Atkinson, since "it is designed to identify three possible types of errors: (1) The initial unit is correct, butthe final unit is not ("rat"). (2) The final unit is correct, but the initial unit is not ("fan"). (3) Neither the initial unit nor the final unit is correctly identified ("bat")" (Atkinson, 1968, p. 228). For either of the first two errors the student gets a single frame which trains either the initial or final consonant and for the third type of error he receives both corrective frames. After any corrective frame the mainline item is repeated. After a correct choice ("ran") the student gets a confirmation frame and the next mainline item.

Enough data were presented to calculate consequence ratios and discriminability, but not enough to determine validity; although Atkinson's data for the percentage of each type of error does permit corroboration of the implications deriving from the validity determination made for this study. Atkinson indicated that the response rate was about four per minute for mainline and corrective items alike. The consequence of the first two types of errors is one corrective item and a repeat of the mainline item for a ratio of 0.67 for two types of error. The consequence for the third type of error (e.g., "bat" for "ran") is two corrective items and a repeat of the mainline item for a ratio of 0.75. Thus the average of the three possibilities is 0.70. This i s a disappointing validity ratio for a program sparking a multi-million dollar CAI movement.

To determine validity, nine mainline items were prepared for the matrix problem described in Atkinson's (1968) paper. These were prepared on sheets of typing paper in large letters as they appeared on the CRT. A group of nine students at the appropriate level in reading were presented these items one-by-one by their teacher who spoke the appropriate audio message. No corrective items were used; only the nine mainline items were used to diagnose trouble with the initial, final, or both consonants. The child touched the alternative with his finger and the teacher recorded the response. After completing the nine items the students were immediately retested with nine similar items that had the same words, but a different random arrangement of the response alternatives.

Care was taken to use children who would approximate the excellent discriminability obtained by Atkinson. He had 45% correct on initial contact with the mainline items. Subjects in the present study were 37% correct on the diagnostic testing (see Table 12). For the nine items and nine children there were a total of 81 decisions with 55 hits and 26 misses for a disappointing validity ratio of 0.68 or only 0.18 better than chance.

Insert Table 12 about here

It is revealing to consider the 17 hits which were instances of incorrect answers on both test and retest. The mainline item not only diagnoses that remedial material is needed, it determines whether training is needed on initial, final, or both consonants depending on the alternative chosen. Thus, when an error is made, it is supposed to indicate which particular alternative treatment is needed. But in this test-retest validity check did the students make the same errors on both testings if they fail both? No; of the 17 "hits" by double failures, 5 picked the same alternative and 12 picked one of the other two alternatives. Apparently, some children can and do read these words accounting for the above chance portion of the passes, but if they do not read, they just pick one regardless of the letters. If the criterion for hits and misses of prediction is whether or not the same alternative is chosen rather then simply passing or failing, then the validity ratio is much lower with 43 hits and 38 misses for a ratio of 0.53 with chance now being 0.25 for one four-way prediction.

This style of program seems beyond redemption. There appears to be no way to save the basic concept of the prescriptive aspect of this program. A single multiple-choice item predicting several alternatives intrinsically has problems with validity, unless the item is made unfailable which sacrifices discriminability for the gain in validity. In addition, using one item to predict the need for only two or three other items hardly provides adequate advantage to the student even if validity of prediction were perfect.

Should the program be used at all? Students using this course did better than a control group on standardized reading tests. What this means is not clear. First, "control" groups from so-called "standard" classrooms are not adequate bases for evaluation as Lumsdaine (1965) pointedly shows. It is possible, moreover, that the sizable exposure to the multiple-choice format gave the students some advantage on the standardized tests following this format. It is also likely that the peppy pace (four frames per minute) provided by good instrumentation exposed the kids to more material than the doldrums of a paper, pencil and blackboard method--but surely man should be able to give some help to the computer. If someone had the computer and the course,it would seem to do no great harm to use them since the technologically inadequate "control" class was not as good as the technologically inadequate CAI program.

It might be argued that the mainline items are principally teaching rather than diagnostic in function. The description of the purpose of these items, quoted earlier,would seem unequivocal in proclaiming their diagnostic function. But even if these items were considered teaching items, their one strength as test items, high discriminability, becomes a weakness. In teaching items, a high error rate indicates that the desired behavior has not occurred. There is a fundamental incompatibility of the two functions. Teaching items must get the students to do something new; testing items must detect the inability of a fair number of students to do it. All things considered, there is little to commend in this style of CAI program.

Summary

Summary of Demonstrations

For instruction to be adaptive, a range of validly measured differences among students must be accommodated by exposing the individual student only to those materials he needs. Do simple measures of the three aspects of adaptive instruction described and illustrated above permit user, developer, or evaluator to easily determine the adequacy of the adaptive features of almost any instructional material in meeting the assumptions involved in adapting? The use of the measures of cost effectiveness, validity, and discriminability were demonstrated with segments of seven different adaptive courses ranging from fine-grain adapting in CAI to pre-course placement testing. All of the materials tested attempt to apply modern instructional theory and most are well-known published courses.

Surprisingly none of the course segments tested proved adequate on all three measures. But in every case the application of the three measures prompted concrete suggestions as to what steps should be taken regarding the specific set of course material. One course, the Programmed Reviews of Mathematics (Flexer and Flexer, 1967) lacked adequate discriminability in its published form but performed well on all three measures when tested with the shorter test used by its authors in their pre-publication testing. The use of the suggested measures offers an easy remedy--use the original form of testing.

Similarly, applications of the measures prompt suggestions for remedies in other programs. The IPI math needs much less use of tests and improved validity. The Job Corps program for GED preparation requires new screening tests correcting the present lack of both discriminability and validity. In fact, most programs require cycles of test and revision to empirically increase the validity and discriminability of the tests while maintaining as short a test as possible.

On the other hand, in one instance of an excellent set of teaching material (Individualized Science), the complete absence of discriminability seems to be simply inherent in the nature of the course content. Children will seldom be adequately proficient in any unit and therefore would probably fail any valid pretest. Thus the course would be improved by committing the heresy of not attempting to adapt to individual differences in proficiency.

Examples of more fine-grained adapting (Tutor Text, and CAI Reading) do not seem to be capable of correction. The single-item multiple choice test seems unsuitable for gaining validity without sacrificing discriminability and the consequence of passing each test is too small to allow an increase in number of test items for each decision.

Distribution of Problems

Over half of the program segments had poor discriminability, with two being completely indiscriminate so that all students would normally be prescribed the same units.

Four of the seven segments were unsatisatisfactory in validity of the tests. The diagnostic tests did not correlate with the criterion tests when no teaching intervened. Therefore, many students would either be prescribed teaching material not needed to pass the criterion test or be allowed to skip teaching material even though they would not be able to pass the criterion test without it. Two of these four program segments actually diagnosed no better than a chance flip of the coin would have.

Three of the seven course segments were inadequate in cost-effectiveness to the student in that the potential saving in time for passing the test was too small given the amount of time required for the teaching or other consequent material. One of these nearly required as long in testing as would be required for the teaching material should it be prescribed.

While no program was adequate in all three measures, none was inadequate in all three either. The reason is probably that concentration by the developer on meeting one requirement can easily cause the sacrifice of another. Long tests are more valid but less cost efficient and a guessing game gives high discriminability but poor validity.

Implication: Is Adaptive Instruction a Myth?

This author believes, along with the vast majority who conduct research and development in educational technology, that, at least in some curricular areas, different students have different needs in reaching a given educational goal and that adaptive instruction will, therefore, be useful. But one looks in vain for compelling empirical support for this

this proposition. Webster's Seventh New Collegiate Dictionary defines myth as "an ill-founded belief held uncritically especially by an interested group." The interested group in educational R & D holds firmly the belief in the worth of adaptive instruction and their belief is amply rewarded by public funds. However, the present study shows that for the seven program segments evaluated, none met each of the necessary requirements for adaptive materials. Of course, some other set of material not tested may do so. This author believes that some material will prove adaptive; but, at present, this belief is still unfounded.

Nor does the experimental literature at present give foundation to the belief in the worth of adaptive instruction. A decade ago a review of studies of branching vs. linear programs failed to reveal advantages for branched programs (Holland, 1965). However, because of the lack of measurable variables for characteristics of "branching" these studies were unpersuasive.

Another line of research, reviewed by Bracht (1970), indicates a general lack of aptitude-by-treatment interaction. The preferred treatment seems not to change for subjects of different aptitude; but the myth-busting relevance of this is limited because adaptive instruction generally has not been concerned with adapting to aptitude differences based on norm-referenced tests but rather differences in achievement based on criterion-referenced tests.

Developers proceeded to produce adaptive materials despite the negative findings of nearly all research. This practice was at least excusable in view of the weakness of the research. The adaptive materials used in the research studies were also developed without explicit guidelines provided by measures for variables of adapting. Research findings which seem to fail to support the theory of adapting to differences could result merely from a lack of materials which truly reflect that theory.

Even though there is scant evidence for a specific advantage of an adaptive feature, many products of modern educational technology, which contain an adaptive feature, produce overall excellent results. IPI Math, IS, Job Corps Advanced General Education Program, and the Programmed Reviews of Mathematics are examples of such materials. Proof of whether or not the good performance of the better materials owes something to the adaptive feature awaits further evaluation; but the new products of educational technology involve many aspects which contrast with conventional practices. Usually the rigid, teacher-oriented classroom is gone; there is no more lock-step instruction whether or not diagnostic testing is used for prescribing instruction; there is more individual attention from teachers who stalk the classroom in search of praisable performance; and, perhaps most importantly, the teaching materials are often prepared following learning principles and behaviorally determined objectives. Therefore, diagnosing and adapting to individual differences is only one among many ways that these materials differ from those of a decade ago. But many of these factors, like adaptiveness, have rarely received clear scientific specifications and, as a consequence, the contributions of the various components have seldom been evaluated. Adaptive instruction could be important; some feature must be important; but as yet the faith in adaptive instruction is an unfounded belief.

A Solution: From Myth to Fact

Simple-to-apply measures of the necessary charcteristics for adaptive material should help the developer generate good adaptive instruction worthy of the, as.yet, unsubstantiated acclaim such instruction has received. With proper measurement and revision cycles there could soon be efficient adaptive materials using tests of proven validity and discriminability.

But, beyond the question of adaptive features, many proclaimed products of educational technology have lacked adequate testing of the underlying assumptions. Work on these products has proceeded without explicit definition or measurement methods for variables important to the preparation of curriculum materials. When measures of all key variables are developed, several educational myths may turn to facts as developers receive the tools for implementing the much touted offering of technology.

References

Atkinson, R. C. Computerized instruction and the learning process. *American Psychologist*, 1968, *23*, 225-239.

Bracht, G. H. Experimental factors related to aptitude-treatment interactions. *Review of Educational Research*, 1970, *40*, 627-645.

Cronbach, L. J. How can instruction be adapted to individual differences? In R. A. Weisgerber (Ed.), *Perspectives in individualized learning.* Itasca, Ill.: Peacock, 1971. Pp. 167-182.

Crowder, N. A. *A tutor test on the arithmetic of computers.* Santa Barbara, Calif.: USI Western Design, 1960.

Crowder, N. A. The rationale of intrinsic programming. *Programed Instruction*, 1962, *1*, 3-6.

Flexer, R. J. & Flexer, A. S. *Programmed reviews of mathematics.* New York: Harper & Row, 1967.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 1963, *17*, 519-521.

Holland, J. G. New directions in teaching machine research. In J. E. Coulson (Ed.), *Programmed learning and computer-based instruction*. New York: Wiley, 1962. Pp. 46-57.

Holland, J. G. Research on programing variables. In R. Glaser (Ed.), *Teaching machines and programed learning, II: Date and directions.* Washington, D. C.: National Education Association, 1965. Pp. 66-117.

Holland J. G. A quantitative measure for programmed instruction. *American Educational Research Journal*, 1967, *4*, 87-101.

Holland J. G., Hoffman, J. & Doran, J. The yoked control for assessing branching effects: Does individualization help? Paper presented at the meetings of the AERA, Chicago, March, 1972.

Klopfer, L., Champagne, A. & Pittman, J. *Individualized science.* Kankakee, Ill.: Imperial International Learning, 1972.

Leinhardt, G. Observation as a tool for evaluation of implementation. In M. Wang (Ed.), The use of direct observation to study instructional-learning behaviors in school settings. Pittsburgh: University of Pittsburgh, Learning Research and Development Center, 1974. 1974/9.

Lindvall, C. M. & Bolvin, J. Programmed instruction in the schools: An application of programming principles in "Individually Prescribed Instruction." In Sixty-sixth Yearbook of the National Society for the Study of Education, Part II. Chicago: NSSE, 1967. Pp. 217-254.

Lumsdaine, A. A. Assessing the effectiveness of instructional programs. In R. Glaser (Ed.), Teaching machines and programed learning, II: Data and directions. Washington, D. C.: National Education Association, 1965. Pp. 267-320.

Office of Economic Opportunity. An experiment in performance contracting. OEO Pamphlet 3400-5. Washington, D. C.: Office of Economic Opportunity, February, 1972.

Office of Economic Opportunity. Job Corps advanced general education program. Washington, D. C.: Office of Economic Opportunity. 1968.

Table 1

Sample Decision Table

		criterion test	
		pass	fail
diagnostic test	pass	hit	miss
	fail	miss	hit

$$\text{Predictive Validity Ratio} = \frac{\text{hits}}{\text{hits + misses}} = \frac{\text{hits}}{\text{total decisions}}$$

Table 2

Job Corps Advanced General Education Program

Validity and Discriminability

(1 Decision x 28 Subjects = 28 Total Decisions)

		criterion	
		P	F
diagnostic	P	0	0
	F	18	10

$$\text{Validity} = \frac{\text{Hits}}{\text{Hits \& Misses}} = \frac{10}{28} = .36$$

$$\text{Discriminability} = \frac{\text{Passes}}{\text{Passes \& Failures}} = \frac{0}{28} = .0$$

Consequence

$$\frac{\text{Teaching Time in Minutes}}{\text{Teaching Time \& Testing Time}} = \frac{510}{524} = .97$$

Table 3

Flexer & Flexer, Fractions

First Evaluation - Long Form Tests

Validity and Discriminability

(3 Decisions x 10 Subjects = 30 Total Decisions)

		criterion	
		P	F
diagnostic	P	5	0
	F	5	20

$$\text{Validity} = \frac{\text{Hits}}{\text{Hits \& Misses}} = \frac{25}{30} = .83$$

$$\text{Discriminability} = \frac{\text{Passes}}{\text{Passes \& Failures}} = \frac{5}{30} = .17$$

Consequence

$$\frac{\text{Teaching Time in Minutes}}{\text{Teaching Time \& Testing Time}} = \frac{150}{184} = .82$$

Table 4

Flexer & Flexer, Logarithms

Validity and Discriminability

(1 Decision x 28 Subjects = 28 Total Decisions)

		criterion	
		P	F
diagnostic	P	2	0
	F	2	24

$$\text{Validity} = \frac{\text{Hits}}{\text{Hits \& Misses}} = \frac{26}{28} = .93$$

$$\text{Discriminability} = \frac{\text{Passes}}{\text{Passes \& Failures}} = \frac{2}{28} = .07$$

Consequence

$$\frac{\text{Teaching Time in Minutes}}{\text{Teaching Time \& Testing Time}} = \frac{67}{76} = .88$$

Table 5

Flexer & Flexer, Fractions

Second Evaluation

A. Long-form tests

Validity and Discriminability

(3 Decisions x 10 Subjects = 30 Total Decisions)

		criterion	
		P	F
diagnostic	P	4	1
	F	4	21

$$\text{Validity} = \frac{\text{Hits}}{\text{Hits \& Misses}} = \frac{25}{30} = .83$$

$$\text{Discriminability} = \frac{\text{Passes}}{\text{Passes \& Failures}} = \frac{5}{30} = .17$$

Consequence

$$\frac{\text{Teaching Time in Minutes}}{\text{Teaching Time \& Testing Time}} = \frac{150}{181} = .83$$

B. Single-item tests

Validity and Discriminability

(3 Decisions x 10 Subjects = 30 Total Decisions)

		criterion	
		P	F
diagnostic	P	15	3
	F	5	7

$$\text{Validity} = \frac{\text{Hits}}{\text{Hits \& Misses}} = \frac{22}{30} = .73$$

$$\text{Discriminability} = \frac{\text{Failures}}{\text{Passes \& Failures}} = \frac{12}{30} = .40$$

Consequence

$$\frac{\text{Teaching Time in Minutes}}{\text{Teaching Time \& Testing Time}} = \frac{150}{156} = .96$$

Table 6

IPI Math, Multiplication (Level B)

Placement → CET's

Validity and Discriminability

(4 Decisions x 22 Subjects = 88 Total Decisions)

		(CET's) criterion	
		P	F
(Placement) diagnostic	P	20	24
	F	6	38

$$\text{Validity} = \frac{\text{Hits}}{\text{Hits \& Misses}} = \frac{58}{88} = .66$$

$$\text{Discriminability} = \frac{\text{Failures}}{\text{Passes \& Failures}} = \frac{44}{88} = .50$$

Consequence

$$\frac{\text{Consequence items (Teaching \& Pretest \& CET's \& Posttest items)}}{\text{Consequence items \& Placement test items}} = \frac{375}{380} = .99$$

Table 7

IPI Math, Multiplication (Level B)

Placement + Pretest → CET's

Validity and Discriminability

Diagnostic			Prescription	Criterion	Hit or Miss	Frequency
Placement		Pre-test		CET's		
Pass	+	Pass*	skip	Pass	hit	17
Pass	+	Pass*	skip	Fail	miss	5
Pass	+	Fail*	skip	Pass	hit	3
Pass	+	Fail*	skip	Fail	miss	19
Fail	+	Pass	skip	Pass	hit	5
Fail	+	Pass	skip	Fail	miss	4
Fail	+	Fail	take	Pass	miss	1
Fail	+	Fail	take	Fail	hit	34
					Total Decisions	88

$$\text{Validity} = \frac{\text{Hits}}{\text{Hits + Misses}} = \frac{59}{88} = .67$$

$$\text{Discriminability} = \frac{\text{Failures}}{\text{Passes + Failures}} = \frac{35}{88} = .40$$

Consequence

$$\frac{\text{Consequence items (CET's + Posttest + teaching)}}{\text{Consequence items + Placement + Pretests}} = \frac{362}{379} = .93$$

*In normal test procedure these tests would not be given because the placement test directed the units to be skipped.

Table 8

IPI Math, Multiplication (Level B)

Pretests ⟶ CET's
Pretests ⟶ Posttest

Validity and Discriminability

(4 Decisions x 22 Subjects = 88 Total Decisions)

(CET's)
criterion

(Pretest) diagnostic	P	F
P	22	9
F	4	53

$$\text{Validity} = \frac{\text{Hits}}{\text{Hits \& Misses}} = \frac{75}{88} = .85$$

(Posttests)
criterion

(Pretest) diagnostic	P	F
P	25	6
F	6	51

$$\text{Validity} = \frac{\text{Hits}}{\text{Hits \& Misses}} = \frac{76}{88} = .86$$

$$\text{Discriminability} = \frac{\text{Passes}}{\text{Passes \& Failures}} = \frac{\overset{31}{\cancel{32}}}{88} = \cancel{.32}\ ,35$$

Consequence

$$\frac{\text{Consequence items (Teaching \& CET's \& Posttest)}}{\text{Consequence items \& Pretest items}} = \frac{353}{375} = .94$$

Table 9

Individualized Science, Hooke Unit

Validity and Discriminability

(5 Decisions x 10 Subjects = 50 Total Decisions)

		criterion	
		P	F
diagnostic	P	0	0
	F	1	49

$$\text{Validity} = \frac{\text{Hits}}{\text{Hits \& Misses}} = \frac{49}{50} = .98$$

$$\text{Discriminability} = \frac{\text{Passes}}{\text{Passes \& Failures}} = \frac{0}{50} = .0$$

Consequence

$$\frac{\text{Teaching Time in Minutes}}{\text{Teaching Time \& Testing Time}} = \frac{116}{127} = .91$$

Table 10

Inductive Reasoning

Validity and Discriminability

(7 Decisions x 11 Subjects = 77 Total Decisions)

		criterion	
		P	F
diagnostic	P	26	26
	F	15	10

$$\text{Validity} = \frac{\text{Hits}}{\text{Hits \& Misses}} = \frac{36}{77} = .47$$

$$\text{Discriminability} = \frac{\text{Failures}}{\text{Passes \& Failures}} = \frac{25}{77} = .32$$

Consequence

$$\frac{\text{Teaching Items}}{\text{Teaching Items \& Testing Items}} = \frac{256}{263} = .97$$

Table 11

A Tutor Text on . . .

the Arithmetic of Computers

Validity and Discriminability

(9 Subjects x 7 Decisions = 63 Total Decisions*)

		criterion	
		P	F
diagnostic	P	52	0
	F	2	5

$$\text{Validity} = \frac{\text{Hits}}{\text{Hits \& Misses}} = \frac{57}{59} = .97$$

$$\text{Discriminability} = \frac{\text{Failures}}{\text{Passes \& Failures}} = \frac{7}{59} = .12$$

Consequence

$$\text{Consequence ratio} = \frac{\text{Consequence}}{\text{Total}} = \frac{1163}{1559} = .75$$

$$\text{Consequence ratio} = \frac{\text{Teaching}}{\text{Total}} = \frac{729}{1125} = .65$$

*Four incompleted items lowered the actual number to 59 total decisions

Table 12

Atkinson, Reading Program

Validity and Discriminability

(9 Decisions x 9 Subjects = 81 Total Decisions)

		criterion	
		P	F
diagnostic	P	38	10
	F	16	17

$$\text{Validity} = \frac{\text{Hits}}{\text{Hits \& Misses}} = \frac{55}{81} = .68$$

$$\text{Discriminability} = \frac{\text{Failures}}{\text{Passes \& Failures}} = \frac{33}{81} = .37$$

Consequence

$$\frac{\text{Teaching Items}}{\text{Teaching Items \& Testing Items}} = \frac{21}{30} = .70$$

Figure Captions

Figure 1. Adapting structure used in the Job Corps Advanced General Education Program. A score of 85% or better on a unit screening test enables the student to skip all of the lessons under the catchment area of the screening test. In Level II, the number of lessons per unit ranges from 2-12. In the figure, unit screening tests are depicted as rectangles, lessons as squares, and unit posttests as diamonds.

Figure 2. Branching tree for the binary search procedure showing the branching sequences which determine the various entry points into the linear sequence shown in the column at the extreme right of the figure. Each sequence begins with item 128 shown at the extreme left of the figure, proceeds upward to item 192 after a correct response or downward to item 64 after an incorrect response. This procedure repeats six times bisecting successively smaller intervals until an entry point in the last column is reached.

Progress Flow Chart Level II

CORPSMEMBER'S NAME ______________________ DATE STARTED: ______________ DATE COMPLETED ______________

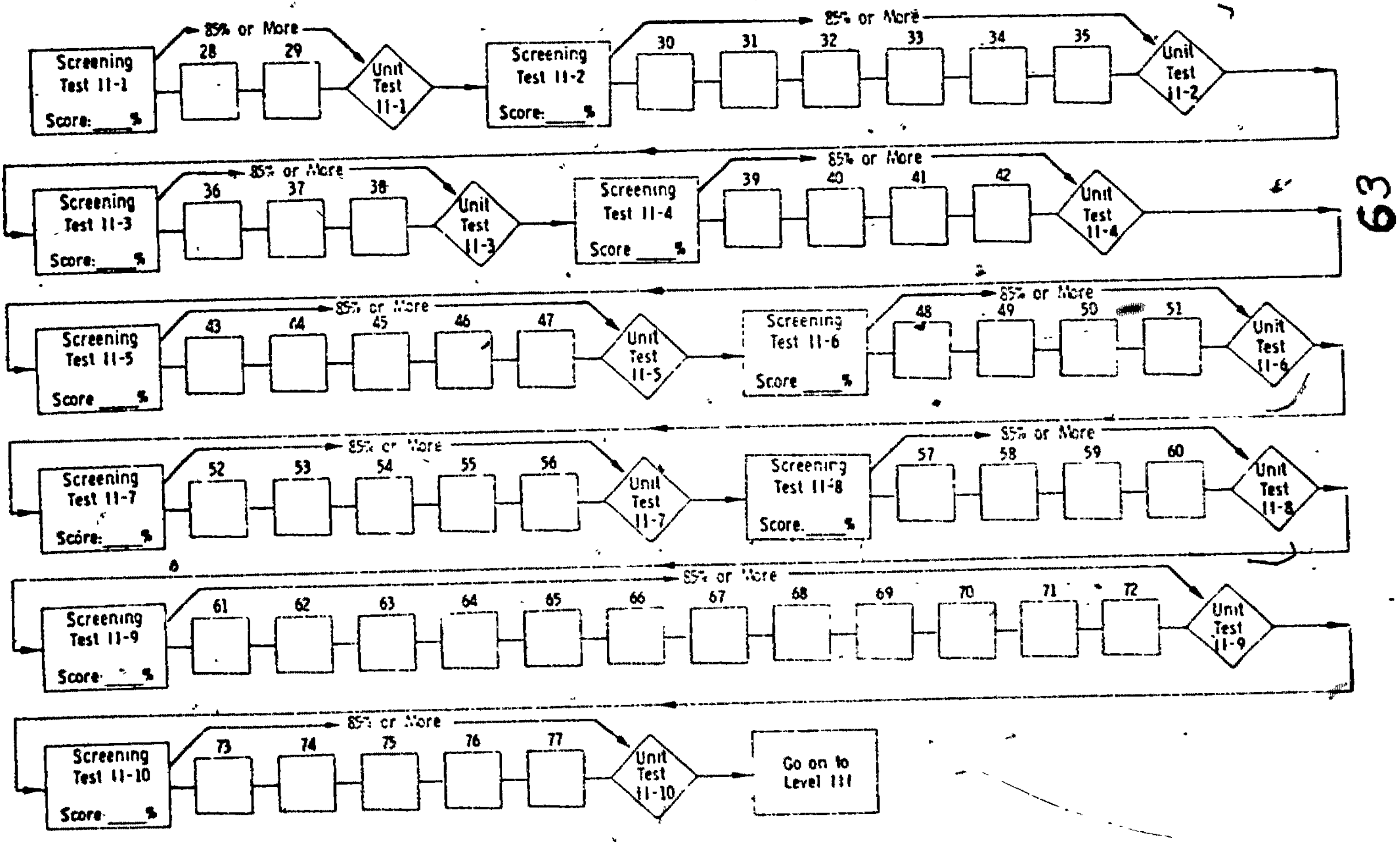

Fig. 1

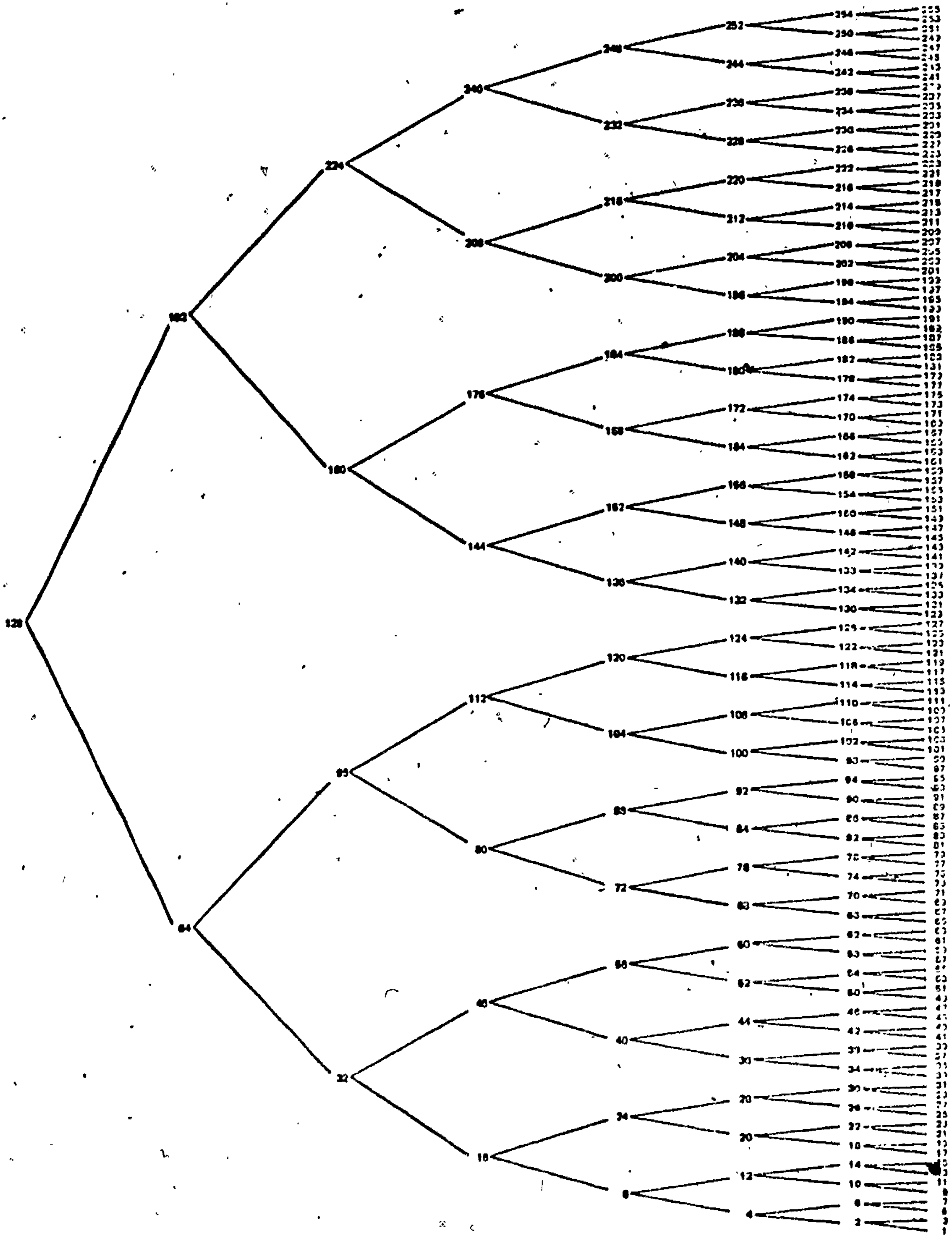
128
192
64
224
160
96
32
240
208
176
144
112
80
48
16
248
232
216
200
184
168
152
136
120
104
88
72
56
40
24
8
252
244
236
228
220
212
204
196
188
180
172
164
156
148
140
132
124
116
108
100
92
84
76
68
60
52
44
36
28
20
12
4
254
250
246
242
238
234
230
226
222
218
214
210
206
202
198
194
190
186
182
178
174
170
166
162
158
154
150
146
142
138
134
130
126
122
118
114
110
106
102
98
94
90
86
82
78
74
70
66
62
58
54
50
46
42
38
34
30
26
22
18
14
10
6
2
255
253
251
249
247
245
243
241
239
237
235
233
231
229
227
225
223
221
219
217
215
213
211
209
207
205
203
201
199
197
195
193
191
189
187
185
183
181
179
177
175
173
171
169
167
165
163
161
159
157
155
153
151
149
147
145
143
141
139
137
135
133
131
129
127
125
123
121
119
117
115
113
111
109
107
105
103
101
99
97
95
93
91
89
87
85
83
81
79
77
75
73
71
69
67
65
63
61
59
57
55
53
51
49
47
45
43
41
39
37
35
33
31
29
27
25
23
21
19
17
15
13
11
9
7
5
3
1

Fig. 2