DOCUMENT RESUME

ED 048 343 TM 000 416

AUTHOR Brennan, Robert L.; Stolurow, Lawrence M.
TITLE An Elementary Decision Process for the Formative Evaluation of an Instructional System.
PUB DATE Feb 71
NOTE 41p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Computer Assisted Instruction, Criterion Referenced Tests, Decision Making, Discriminant Analysis, Educational Objectives, Error Patterns, *Evaluation Techniques, Instructional Improvement, *Item Analysis, Performance Factors, Post Testing, Pretesting, *Program Effectiveness, *Systems Approach, *Test Construction, Test Interpretation, Test Results

ABSTRACT

A replicable process for improving instruction through the consistent use of student data collected before, during, and after instruction is proposed. A rational analysis of different types of error rates (theoretical, base, posttest, instructional) and discrimination indices (base, posttest) leads to a set of rules for identifying test items and sections of instruction that require revision. The application of these rules is illustrated through the analysis of student responses to a subset of test items in a computer assisted instruction program in micro-economics. Finally, a discussion is presented that relates the proposed decision process to some theoretical issues in criterion-referenced testing and the formative evaluation of instruction. (Author/LR)

AN ELEMENTARY DECISION PROCESS FOR THE FORMATIVE EVALUATION OF AN INSTRUCTIONAL SYSTEM[1]

Robert L. Brennan

and

Lawrence M. Stolurow

Harvard University

During the past decade evaluators of programmed instruction and computer-aided instruction have recognized that it is very difficult, if not impossible, to determine subjectively the effectiveness of test items and instruction (see Rothkopf, 1963). In this paper we will specify a set of objective rules, based upon item performance date, for identifying those test items and sections of instruction that seem to require revision. This objective method should provide a more rational basis for decision-making than the subjective method of making decisions based upon some unidentified combination of subject matter knowledge, experience, and intuition.

The rationale for decision-making that we propose is basically an elaboration of a technique devised by Stolurow and Frase (1968). Their method is based upon a comparison of three different types of error rates for program frames: (a) the theoretical error rate (T), which is the error rate expected simply on the basis of random guessing; (b) the base error rate (B) which is the error rate obtained by students not exposed to the teaching material for the frame; and (c) the instructional error rate (I), which is the error rate obtained by students who have been exposed to the instruction.

In this paper we will treat not only program frames that are an integral part of instruction but also test items that occur both before and after instruction. In addition, we will use both error rates and discrimination indices as data for decision-making.

In order to put the decision process we propose into a conceptual context, let us assume that we have an instructional program teaching a set of terminal objectives. Chronologically, each terminal objective is tested by (a) a pretest item that occurs before the objective has been taught, (b) a terminal test item that occurs almost immediately after the objective has been taught, and (c) a posttest item that occurs "some time after" the objective has been taught.[2] Without loss of generality, we will assume (as is usually the case) that the set of pretest and posttest items form two tests that occur, respectively, prior to and following the instruction for all objectives. Furthermore, we will assume that all of the items testing any objective are either identical or "corresponding". (The concept of "corresponding" items will be treated in detail later; however, we can roughly define corresponding items as items that test the same content at the same level of difficulty.) In the final analysis, using item performance data, we want to identify those test items and sections of instruction (relevant to a given objective) that require revision. The decision process we propose will not necessarily tell the evaluator how to revise items and/or instruction, but the process will provide objective rules for deciding what to revise.

Types of Data and Decisions

Error rate is defined as the proportion of students getting an item incorrect, i.e.,

$$\text{Error Rate} = \frac{\text{Number of Incorrect Answers}}{\text{Total Number of Answers}} \qquad (1)$$

or

$$ER_i = \frac{N - \sum_{j=1}^{N} X_{ij}}{N} \qquad (2)$$

where ER_i means error rate for item i, N is the total number of students answering the item, $X_{ij} = 1$ if student j gets item i correct, and $X_{ij} = 0$ if student j gets item i wrong. We can also express Equation 2 as:

$$ER_i = 1 - \frac{\sum_{j=1}^{N} X_{ij}}{N} \qquad (3)$$

Since the last term on the right of Equation 3 is item difficulty level (DL_i), it is clear that

$$ER_i = 1 - DL_i \; ; \qquad (4)$$

i.e., error rate equals one minus difficulty level. Clearly, Equation 4 shows that from a theoretical viewpoint it is immaterial whether we use difficulty level or error rate; however, using error rate seems to facilitate an understanding of some of the decisions that will be proposed later.

In much of what follows we will assume that error rates are classified as either high (H) or low (L), and that the evaluator predetermines an appropriate cut-off point between high and low error rate. For any given objective, the cut-offs for TER, BER, IER, and PER must be identical in order to apply the rules that will be specified. Also, in most cases, the cut-offs chosen will probably be the same for all objectives; however, occasions can arise when certain objectives should have a higher (or lower) error rate cut-off than other objectives. For example, items testing very crucial objectives might be assigned a cut-off of 0.90, while other items might have a cut-off of 0.70.

Discrimination indices will be classified as either positive (+), negative (-), or non-discriminating (0). By positive and negative indices we mean indices that discriminate significantly (at some appropriate α - level) in the positive and negative directions, respectively. The discrimination index used should, of course, be appropriate for the data in question.

Before instruction we can obtain three types of data for each objective that has a pretest item:

(a) the Theoretical Error Rate (TER), which is the expected propostion of students getting a pretest item incorrect simply on the basis of random guessing; i.e., if K is the number of possible answers to an item, then

$$TER = \frac{K - 1}{K} . \qquad (5)$$

For example, if an item has five alternatives, we would expect 80 percent of the students to get the item incorrect simply by guessing, without any knowledge of the objective tested by the item[3];

(b) the Base Error Rate (BER), which is the observed proportion of students getting a pretest item incorrect; and

(c) the Base Discrimination Index (BDI), which is the discrimination index for a pretest item. (We will use total score on the pretest as the criterion variable for BDI.)

After instruction we can obtain two types of data for each objective that has a posttest item: (a) the Posttest Error Rate (PER), and (b) the Posttest Discrimination Index (PDI). (Total score on the posttest will be used as the criterion variable for PDI.)

Immediately following the instruction for any objective we can obtain the Instructional Error Rate (IER), which is the error rate on a terminal test item for a given objective[4]. Note that IER refers to the error rate on a terminal item, not the error rate on other questions associated with teaching the given objective. We will not consider Instructional Discrimination Index since, in our opinion, it does not seem to be very useful for making decisions beyond those that can be made with the other types of data.

In subsequent sections we will analyze the decisions that can be made on the basis of: (a) pretest data, alone; (b) posttest data, alone; (c) pretest and posttest data; and (d) pretest, posttest, and instructional test data. In this way, the contribution of the various types of data to the decision process should be evident. For each analysis, we will specify reasons for determining whether test items or instruction relevant to a given objective should be revised (R), questioned (?), or not revised (NR)[5]. Since we are assuming that all items testing a given objective are identical or "corresponding", a decision about item revision applies equally to all items testing the objective in question. For example, if on the basis of pretest data it is clear that an item should be revised, we must also revise the corresponding terminal test item and posttest item. Thus, when we say that an item should be revised, we mean that all items testing the given objective should be revised. Likewise, when we say that instruction should be revised, we mean that that part of the instructional system that attempts to teach the given objective should be revised.

Pretest Data

Prior to instruction we can collect three sets of data: Theoretical Error Rate (TER), Base Error Rate (BER), and Base Discrimination Index (BDI). Given these three sets of data, various reasonable rules can be formulated for making decisions about whether or not to revise test items. It is not likely that only pretest data would be used to make decisions about items, yet it is useful to consider the types of decisions that are appropriate on the basis of such data.

Rule 1. If TER and BER are both the same (i.e., H, H or L, L), then no necessity for revision is indicated. In this case, the observed error rate (BER) without benefit of instruction is approximately the same as the expected error rate (TER).

Rule 2. If TER is low (L) and BER is high (H), then no revision is indicated. This anomalous case could arise if the particular objective for the item involved concepts that are typically misunderstood. For example, many students (in the authors' opinion) believe that "inflammable" and "non-flammable" have different meanings. If an item were constructed testing whether or not "flammable" and "inflammable" have the same meaning, and if this item were given prior to instruction, it is quite possible that more students would get the item incorrect than we would expect on the basis of the theoretical error rate (TER). In this case, there is no reason to revise the item; rather, we expect that the instruction will correct the students' misconception.

Rule 3. If TER is high (H), and BER is low (L), then the item will probably need to be revised. In this case, students, without benefit of instruction, are performing considerably better than expected. It appears that the item itself may be teaching or that the distractors are so easy that most students can pick the correct answer by the process of elimination. In either case, the item should be revised.[6]

Rule 4. If an item is negatively discriminating before instruction, then the item is questionable in that it may need revision. If, however, the item is positively discriminating or non-discriminating, then no revision is indicated. A negatively discriminating item is questionable since it indicates that the worse students (on the basis of total test score) are out-performing the better students; however, a situation similar to that indicated in Rule 2 could be the cause of the negative discrimination index. A positively discriminating item is quite possible and reasonable prior to instruction simply because some good students are usually expected to perform better than chance on a pretest. A non-discriminating item is the best of all possibilities.

Rule 5. If an item is positively or negatively discriminating before instruction, then the prerequisites for the objective tested by the item should be checked. Clearly, whenever an item is discriminating (either positively or negatively) one group (upper or lower) is outperforming the other group (lower or upper). In such a case, it seems reasonable to check whether or not the group with the higher error rate does, in fact, possess the prerequisites necessary to achieve the given objective.

Insert Table 1 about here

These rules, as well as all other rules that will be discussed, are given in abbreviated form in Table 1.

Posttest Data

As a result of administering a posttest two types of data can be collected: the Posttest Error Rate (PER) and the Posttest Discrimination Index (PDI). Since these data are collected after instruction, theoretically decisions can be made about both items and instruction; however, it is very difficult to identify items and instruction that should be revised solely on the basis of posttest data. In almost every case, we can say whether or not there is something wrong, but we cannot pinpoint the problem.

Rule 6. If PER = L and PDI = 0, then neither the item nor the instruction need to be revised. This is the best possible situation, since the optimal conditions for both error rate and discrimination index are fulfilled; i.e., at the end of instruction we hope that most of the students get the posttest item correct (PER = L), and that the item is non-discriminating (PER = 0). (Later we will discuss our reasons for preferring non-discriminating items.)

Rule 7. If PER = L and PDI = + or -, then both the item and the instruction are questionable. The fact that PDI is clearly non-zero indicates a possible need for revision.

Rule 8. If PER = H and PDI = -, then both the item and instruction should be revised, since PER = H and PDI = - is the worst possible situation that can occur. It is possible that either the item or the instruction is at fault, but not both; however, we assume here that the most universally applicable decision is to check both the item and the instruction to see what revisions are needed.

Rule 9. If PER = H and PDI = + or -, then the instruction should be revised and the item should be questioned. Whenever error rate is high after instruction, something is wrong, but without additional information we do not know whether the fault definitely lies with the item or the instruction. However, the authors believe that evaluators are often more confident about the test items than they are about the instruction; it is also possible that the test items have been previously validated or partially validated. Therefore, in this case, it seems reasonable to place a less stringent decision on the item than on the instruction.

Rule 10. When PDI = + or PDI = -, then the prerequisites for the objective tested by the item should be examined. The reason for this decision is identical to that presented in Rule 5 in the previous section.

Pretest and Posttest Data

It is evident from Table 1 that neither the pretest data alone (see Rules 1-5) nor the posttest data alone (see Rules 6-10) give the evaluator much indication about which items and/or sections of instruction should be revised. Clearly, more meaningful decisions can be made by combining the two sets of data. When this is done all of the rules discussed in the last two sections are applicable, with the exception of Rule 5 which is superseded by Rule 10. In addition, one more rule can be specified.

Rule 11. If BDI = - and PDI = -, then the item should be revised. Both before and after instruction the item is negatively discriminating, which means that the upper group (based on total test score) has a proportionately higher error rate than the lower group. This clearly is an unfortunate circumstance indicating that the item should be revised.

Pretest, Posttest, and Terminal Item Data

Recall that Instructional Error Rate (IER) is the error rate on a terminal item immediately following instruction. If, in addition to pretest and posttest data, we also take into account IER, it is possible to make fairly definite statements about whether or not to revise most segments of instruction that are related to terminal objectives. The addition of IER does not, however, tell us much more about the revision of items than we already know from pretest and posttest data. All of the rules previously specified are applicable except for Rule 5 which is superseded by Rule 10. Also, we can specify four additional rules.

Rule 12. If Instructional Error Rate (IER) and Posttest Error Rate (PER) are low, then no revision (NR) of instruction is indicated. Both during instruction and after instruction most of the students seem to achieve the objective (tested by the instructional item and the posttest item); therefore, we have two indications that the instruction is adequate, and no revision is indicated.

Rule 13. If IER = L and PER = H, then the instruction should be revised. During instruction students seem to achieve the objective, but on the posttest the same students have a higher error rate for the same objective. Thus the data indicate a retention problem, and the instruction should be revised to correct this situation. Perhaps more review is needed.

Rule 14. If IER = H and PER = L, then the instruction should be questioned. This is probably an unlikely situation that would seldom occur in practice. However, the fact that students experience a high error rate on a terminal test item during instruction seems to indicate that something may be wrong with the instruction.[7]

Rule 15. If IER = H and PER = H, then the instruction definitely should be revised. Both during and after instruction students do not seem to achieve the objective under consideration. We, therefore, have two indications of a need for revising the instruction.

Decisions Based Upon Differences Between Error Rates

Most of the foregoing decision rules are dependent upon the evaluator's choice of a cut-off between high and low error rate. Dichotomizing error rate in this way clearly facilitates the identification of appropriate decision rules, and, in many cases, the simplicity of the technique will probably outweigh any loss of precision. However, we can also specify an additional set of four useful decisions rules that take into account quantitative differences between error rates. Three of these rules increase the power of previous decisions, the other provides essentially new information. We will call these error rates "derived" error rates in order to distinguish them from the "raw" error rates discussed in the previous sections.

Let us consider several limitations of the high/low classification procedure for error rates. Suppose that Theoretical Error Rate (TER) and Base Error Rate (BER) for a given objective are both classified as high (H), while Instructional Error Rate (IER) and Post-test Error Rate (PER) are both classified as low (L). Clearly, any actual arithmetic differences between TER and BER, as well as between IER and PER, will not affect the decisions we have thus far proposed. Also, since BER and IER are merely classified as high and low, respectively, we won't have a quantitative measure of how much learning has actually taken place.

Difference Error Rate

Rules 1-3 are useful for making decisions based upon categorical differences between BER and TER, but we can make more accurate decisions by actually computing the differences between these error rates. Let

$$DER = TER - BER, \quad (6)$$

where DER stands for "Difference Error Rate". If DER = 0, then the observed error rate (BER) on the pretest item in question is identical to the expected error rate (TER). If DER $<$ 0, then fewer students are getting the item correct than we would expect on the basis of random guessing. Finally, if DER $>$ 0, then more students are getting the item correct than we would expect. As discussed previously, the last possibility is often an unfavorable situation, since it can mean that the item somehow "gives away" the correct answer.

We can test the significance of a positive difference between BER and TER by computing

$$Z = \frac{DER - 1/2N}{\sqrt{TER(1 - TER)/N}}, \quad (7)$$

where N is the total number of students in the sample (see Snedecor & Cochran, 1967, p. 210).[8] The computed Z value is then compared with the normal curve standard score at an appropriate level of significance for a one-tailed test. (Note that we are interested only in positive values of DER.) We can now specify a more precise rule to replace Rules 1-3.

Rule 16. If the value of DER is significantly less than zero, then the item should be revised. In all other cases no revision is required.

Retention Error Rate

Rules 12 - 15 are useful for making decisions based upon categorical differences between IER and PER, but we can supplement these decisions by calculating the actual difference between IER and PER and comparing this value to some preassigned cut-off. Let

$$RER = PER - IER, \tag{8}$$

where RER stands for "Retention Error Rate". If RER = 0, then the number of errors on the posttest item and the related terminal item is identical, and no retention problem is evident. If RER $>$ 0 then students make more errors on the posttest item than on the terminal item. The latter situation can be serious if RER is considerably greater than zero; however, it is not clear how to define "considerably greater than zero".

We can, of course, test the statistical significance of RER if certain distributional assumptions can be made, but such a test would not, in our opinion, provide a meaningful basis for decision. What is needed is a cut-off above which the amount of forgetting is great enough to justify revision of instruction. Such a cut-off must take into account the criticality of forgetting which in turn is dependent upon many factors including the content matter of the instructional system and the population for which the system is being developed. Furthermore, there is no theoretical rationale for specifying the same cut-off for all items. Thus, in our opinion only the evaluator can make an appropriate choice of a useful cut-off. It, therefore, seems reasonable to specify the following rule as a more powerful version of Rule 13.

<u>Rule 17</u>. If RER $> c_1$, where c_1 is a cut-off specified by the evaluator, then the instruction should be revised, since the data indicate a retention problem. If $0 \leq$ RER $\leq c_1$, then no revision is required. The cut-off, c_1, need not be the same for all objectives.

The one possibility that we did not consider above is RER < 0; i.e., students make fewer errors on the posttest item than on the terminal item. We stated previously, in the discussion of Rule 14 that this is an unlikely occurrence; however, the evaluator may want to specify a cut-off below which he considers this problem to be serious enough to merit a closer examination of the instruction.

<u>Rule 18</u>. If RER $< -{}_2$, when c_2 is a cut-off specified by the evaluator, then the instruction should be questioned. If $-c_2 \leq$ RER ≤ 0, then no revision is required. As before, the cut-off c_2 need not be the same for all objectives.

<u>Percentage of Maximum Possible Gain</u>

None of the decisions discussed up to this point has made use of any measure of gain in knowledge relevant to a given objective that results from the instructional system. It is probably true that gain is not as important as final performance on the posttest, in most instructional systems; however, if students experience relatively little gain as a result of experiencing instruction, one can legitimately question the value of the instructional system itself. Thus, measures of gain have long been a subject of considerable interest in the fields of programmed instruction, computer-aided instruction, and multimedia instruction (see Lumsdaine, 1965).

The simplest measure of gain for an objective is the difference between error rate on a pretest item (BER) and error rate on the corresponding terminal item (IER)[9]. Such a measure would, however, mean that a gain of 0.50 resulting from BER = 1.00 and IER = 0.50 would be indistinguishable from a gain of the same magnitude resulting from BER = 0.50 and IER = 0.00. In the former case, the instructional system has failed to produce 50 percent of the gain in performance that could be achieved, while in the latter case, the instructional system has produced as much gain as possible given the entry level of the students. Thus, in the former case, some revision of the instruction may be desirable, while in the latter case, no revision in the instructional system is required on the basis of this particular data.

This above rather trivial example illustrates that simple gain does not provide a very meaningful basis for revising instruction. A better measure is percent of maximum possible gain for an objective, defined as:

$$PMPG = \frac{BER - IER}{BER} \quad . \tag{9}$$

In order to make use of this measure the evaluator must specify a cutoff that determines whether or not a given value for PMPG indicates a need for revision; i.e., Rule 19. If PMPG $< c_3$, where c_3 is a cut-off specified by the evaluator, then the instruction should be revised. The cut-off c_3 need not be the same for all objectives.

The literature contains many in-depth discussions and debates about the problems and pit-falls associated with measures of gain (see, for example, Cronbach and Furby, 1970). Most of this literature, however, treats measures of gain in the context of their use in inferential statistics or correlational analysis. While we appreciate the importance of these issues, we hasten to add that measures of gain, merely as descriptive statistics, can provide useful information to evaluators. We believe that the use of PMPG, as data for evaluation purposes, is a case in point.

When data of the type discussed in this section are used along with the basic pretest, posttest, and terminal item data, then the appropriate decision rules are: 6-11 and 16-19. If only pretest and posttest data are available, then Rule 16 can be used to replace Rules 1-3.

An Example

The data reported in Table 2 are based upon the responses of 28 students to a subset of test questions in an interactive CAI program in micro-economics developed at the Harvard Computer-Aided Instruction Laboratory.[10]

Insert Table 2 about here

The discrimination index used for both BDI and PDI is the phi-coefficient. In the case of BDI, all students with scores of four or more items correct on the pretest were classified into the upper criterion group, and all other students were classified into the lower criterion group. In the case of PDI, all students with scores of 15 or more items correct on the posttest were classified into the upper group, and all students with scores of 12 or fewer items correct were classified into the lower group. Both BDI and PDI were tested using a correction for discontinuity (see Edwards, 1967, p.333) and two-tailed probability levels.

Insert Table 3 about here

The categorical error rates and discrimination indices given in Table 3 are based upon the cut-off values indicated in the footnotes to that table. The cut-offs used were selected primarily for illustrative purposes, and are not necessarily intended to be optimal cut-offs from a theoretical standpoint. Note that the cut-offs are the same for all items.

Insert Table 4 about here

Table 4 lists the decisions that result from applying the various decision rules to three different subsets of the data reported in Table 3. When two rules indicate a need for revision, both are given; in most other cases, only one rule is applicable. Occasions do arise, however, when two or more different decisions are applicable to the same item or segment of instruction. For example, objective number five has IER = H, PER = L and PDI = 0. According to Rule 6 the instruction does not need revision, but Rule 12 indicates that the instruction is questionable. We have chosen to resolve such conflicts by selecting the decision that has the most serious implications for revision; i.e., "questionable" (?) has more serious implications for revision than "do not revise" (NR), and "revise" (R) has more serious implications for revision than either "questionable" (?) or "do not revise" (NR). Thus, for objective number five we have labelled the instruction "questionable" in the second set of decisions.

In Table 4 the first set of decisions uses more data than the second which, in turn, uses more data than the third. One possible effect of decreasing the amount of data used is illustrated by the decisions with regard to instruction for objective number five. Using all of the data for objective five in Table 4, Rule 19 indicates that the instruction should be revised. When, however, derived error rates are eliminated, Rule 19 becomes inapplicable, and Rule 14 indicates that the instruction should be examined, but not necessarily revised. Finally, when both derived error rates and IER are eliminated, both Rules 19 and 14 become inapplicable, and Rule 6 indicates that no revision is required. This situation is an empirical demonstration of the desirability of obtaining as much data as possible in order to strengthen decisions about the adequacy of instruction.

This statement does not, however, imply that an increase in the amount of available data will necessarily increase the number of decisions involving the revision (R) of items or instruction. Consider, for example, the decisions involving instruction, given in Table 4, for objective number 11. Using only pretest and posttest data, no revision is required according to Rule 6. When IER is included as data for decision making, the second set of decisions indicate that the instruction is questionable according to Rule 14. When, however, all available data are used (i.e., pretest and posttest data, IER, and the derived error rates), we again arrive at the decision "no revision" according to Rules 6 and 18[11]. Clearly, in the case of objective 11, an increase in the amount of available data ultimately confirms our initial judgment that no revision of instruction is required.

For this particular instructional system, Table 4 indicates that the availability of derived error rates increases the number of decisions that involve revision of items and instruction. Furthermore, in general, revision is most often necessitated by relatively poor performance on the posttest (note the many times Rules 8 and 9 are employed) and relatively poor retention (note the many times Rules 13 and 17 are employed). Also, the instruction seems to be in more need of revision than the test items. These general observations do, in fact, coincide with the predictions of the person responsible for developing this particular instructional program.[12]

Discussion

It is certainly reasonable to expect that some readers may feel that certain decisions we have proposed are not appropriate for their particular programs, or that other decision rules should be added. We have tried to specify those decisions that we feel are the most universally applicable; however, even more important than the actual decision rules presented is the method used to arrive at decisions about test items and instruction. Hopefully this method is generalizable.

In this section we will discuss various factors that have applicability to the rules we have presented and the decision process we have proposed.

Instructional Systems and Criterion-Referenced Testing

One might define an instructional system in general as a replicable method of instruction providing feedback that can be used for revision purposes. Such systems are usually characterized by a close correspondence between test items and behavioral objectives, i.e., test items are criterion-referenced. In addition, it is usually expected that "most" of the students will get "most" of the terminal and posttest items correct.

Brennan (1970) and Popham & Husek (1969) have examined some aspects of the applicability of classical test theory to the analysis of criterion-referenced tests. Perhaps the most important implication of these analyses is that the classical normality assumptions concerning errors of measurement do not seem to be appropriate in the criterion-referenced testing situation; the errors of measurement seem to be better characterized by binomial error models (see Lord & Novick, 1968, Chapter 23). This means that many of the statistics used in classical test theory are not applicable in the criterion-referenced testing situation. For example, the biserial discrimination index is not appropriate for criterion-referenced test data, since total scores on the test are not necessarily normally distributed; a similar comment can be made about the tetrachoric discrimination index.

Another characteristic of a good instructional system is that all students who receive instruction achieve criterion performance on the posttest regardless of previous knowledge or experience (see Stolurow & Davis, 1965). Ideally, in fact, we may want all students to achieve all objectives. In such a situation all items would be non-discriminating (assuming, of course, that total test score is the criterion used for judging discriminability). This line of reasoning indicates why we have specified that non-discriminating items do not indicate a need for revision. Conversely, items that are significantly discriminating (especially negatively discriminating items) indicate a possible need for revision since the instructional system is performing worse for one group of students than for another group.

Corresponding Items

When discussing the context of the rationale that has been presented, we assumed that for each objective there exists a pretest, posttest, and terminal test item; furthermore, we assumed that the items testing a given objective are, in some sense, "corresponding," "equivalent", or "parallel".

The terms "equivalent" and "parallel" are, in the classical sense, usually applied to tests. A set of k tests are said to be "parallel" or "equivalent" if they have equal means, equal variances, and equal intercorrelations (see Gulliksen, 1950, p. 173). This does not mean, however, that there is necessarily any strict correspondence among items in the k tests. Thus, in the rationale that we have proposed, and in criterion-referenced testing in general, the classical concept of parallel tests is clearly not sufficient, since we are very concerned about the performance of students on individual items, not just entire tests. Let us, therefore, reserve the terms "parallel" and "equivalent" for entire tests, and examine the analogous issue of "corresponding" items.

We can define "corresponding" items, in general, as items that measure the same thing. Clearly, then, one requirement of corresponding items is that, in the judgment of specialists the items measure the same behavioral objective. Furthermore, just as we have a statistical criterion for parallel tests, it seems reasonable to have a similar statistical criterion for corresponding items. Thus, another reasonable requirement for corresponding items would seem to be that they have equal means, equal variances, and equal intercorrelations. Since we are assuming that items are scored dichotomously, the mean of item i is simply the proportion of correct responses (p_i) and the correlation between any two items is the phi correlation (r_ϕ).

Now, suppose we give a set of k tests to N students in order to determine whether or not the tests are parallel; i.e., whether or not the set of k means, k variances, and k(k - 1)/2 intercorrelations are equal except for sampling differences. Wilks (1946) provides a statistical test to answer this question.

Unfortunately, however, Wilks' test is not applicable for judging the equality of a set of means, variances, and intercorrelations for k dichotomously scored criterion-referenced items. Wilks' test assumes a normal multivariate population distribution, and, as we have stated previously, the assumption of normality is probably inappropriate in the criterion-referenced testing situation.

As far as we know, there is no currently available method for simultaneously testing the equality of means, variances, and correlations among dichotomously scored items that are not necessarily normally distributed. We can, however, approach a solution to the problem by applying what is usually called Cochran's Q Test (see Siegel, 1956, pp. 161-166), which is a test for the equality of means, or proportions (p_i), among dichotomized variables (in this case, test items).

Since the variance of a dichotomous variable scored zero or one is completely determined by the mean (or proportion of successes), it is clear that if the means of k items are equal, then the variances will also be equal. However, even if the means and variances of k items are equal (except for sampling differences), this does not necessarily mean that the intercorrelations are equal. The authors have no knowledge of any currently available method to test the equivalence of intercorrelations (phi-coefficients) among dichotomously scored items which may not be distributed normally in the population.

Besides the problem of non-normally distributed variables there is another problem in testing the equivalency of intercorrelations (phi-coefficients) that may not be immediately evident. Suppose we have three items (i.e., $k + 3$). In order to test whether or not the intercorrelations among the items are the same, we must take into account three different phi-coefficients: (a) r_ϕ between item one and item two, (b) r_ϕ between item one and item three, and (c) r_ϕ between item two and item three. Now, it is clear that (a) and (b) are correlated because both phi-coefficients are based on the same data for item one; (a) and (c) are also correlated since they are based on the data for item two; and finally, (b) and (c) are correlated because they are based on the same data for item three. Since the three r_ϕ's are clearly correlated, we cannot apply any of the well-known chi-square tests that are currently available for use with contingency tables. In the absence of a test of significance for examining the equivalence of intercorrelations (phi-coefficients) among k items, the evaluator will probably have to use his best judgment about whether or not the phi-coefficients are "approximately" equal.

In summary, we have defined corresponding items as items that (a) measure the same behavioral objective, (b) have the same means, (c) have the same variances, and (d) have the same intercorrelations. We have recommended Cochran's Q Test as a method for testing (b) and (c), but we are unable to specify a method for testing (d). In practice, however, the lack of a statistical test for (d) may not be too serious a limitation. Certainly, if conditions (a), (b), and (c) are fulfilled and the intercorrelations among the items are approximately equal, it is reasonable to assume that the items are "corresponding".

Comments on Data for Decision Making

For purposes of simplicity, the decision rules we have specified are based upon data from one pretest item, one terminal item, and one posttest item for each objective. There may, of course, be more than one pretest, terminal, and/or posttest item for any given objective. Such additional data can be taken into account in various ways. For example, one might merely combine the data from all the pretest (posttest or terminal) items relevant to a given objective in order to calculate the appropriate error rate. Alternatively, assuming, for example, that three posttest items test the same terminal objective, one might specify that if a student answers two of the three items correctly, then he has achieved the objective. Other alternatives are also possible; however, a multiplicity of pretest, posttest, or terminal items relevant to a given objective can complicate the interpretation of which item, if any, requires revision.

We have also assumed that every student answers every item. There are several formulas available (see Guilford, 1954, pp. 418-424) that can be used to calculate error rates with missing data. Such formulas can be used instead of Equation 2. A large amount of missing data can, however, present serious problems, especially if the sample size is small.

There are many discrimination indices available in the literature (see Guilford, 1954, pp. 424-440) than could be used to calculate BDI and PDI. In our opinion, however, the phi-coefficient and the B index (see Brennan 1970, 1972 in press) are the best indices to use with criterion-referenced tests, since they make only weak distributional assumptions, and they allow the evaluator to specify virtually any cut-off between upper and lower groups. In addition, the index B has a very useful interpretation in terms of the number of discriminations made by an item.

One further comment seems appropriate. Stolurow and Frincke (1966) have noted that there is a danger of rejecting good items (or good instruction) when the sample size is relatively small, say N = 15 or 20. In their study, Stolurow and Frincke were concerned about error rates only. Since, in this paper we examine both error rates and discrimination indices, it is certainly desirable that the sample size be sufficiently large. We believe that an N of about 25 or 30 should be adequate for most purposes. The technique we have proposed can be used with smaller sample sizes; however, the certainty with which decisions can be made is thereby reduced.

REFERENCES

Brennan, R.L. Some statistical problems in the evaluation of self-instructional programs. (Doctoral dissertation, Harvard University) Ann Arbor, Mich.: University Microfilms, 1970. No. 70-23080.

Brennan, R.L. A generalized upper-lower item discrimination index. Educational and Psychological Measurement, 2, 1972, in press.

Cronbach, L.J. & Furby, L. How should we measure "change"--or should we? Psychological Bulletin, 1970, 74(1), 68-80.

Edwards, A.L. Statistical methods. New York: Holt, Rinehart, and Winston, 1967.

Guilford, J.P. Psychometric methods. New York: McGraw Hill, 1954.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

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

Lumsdaine, A.A. Assessing the effectiveness of instructional programs. In R. Glaser (Ed.), Teaching machines and programmed learning, II--data and directions. Washington: National Education Association, 1965.

Popham, W.J. & Husek, T.R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6(1), 1-9.

Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw Hill, 1956.

Snedecor, G.W. & Cochran, W.G. Statistical methods. Ames, Iowa: Iowa State University Press, 1967.

Stolurow, L.M. & Davis, D. Teaching machines and computer-based systems. In R. Glaser (Ed.), Teaching machines and programed learning, II--data and directions. Washington: National Education Association, 1965.

Stolurow, L.M. & Frase, L.T. The logic basis and technological implications of a decision process in the development of instructional materials. Technical Recommendation No. 8, July, 1968, Harvard Computer-Aided Instruction Laboratory, Cambridge, Mass., United States Naval Academy Contract No. N00161-7339-4781.

Stolurow, L.M. & Frincke, F. A study of sample size in making decisions about instructional materials. Educational and Psychological Measurement, 1966, 26, 643-659.

Rothkopf, E.Z. Some observations on predicting instructional effectiveness by simple inspection. Journal of Programmed Instruction, 1963, 2(2), 19-20.

Wilks, S.S. Sample criteria for testing the equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. Annals of Mathematical Statistics, 1946, 17, 257-281.

FOOTNOTES

[1]Much of the research reported herein was performed pursuant to contracts with the United States Naval Academy, Contract No. N00161-70-C-0119, and the Office of Naval Research, Contract No. N00014-67-A-0298-0032.

[2] A posttest item, as we are using the term, is, in part, a measure of retention. Clearly, the evaluator must temper his decisions about revision with knowledge about the length of time intervening between instruction and testing as well as the criticality of forgetting.

[3] Items that have a virtual infinitude of possible answers have TER = 1.00; however, the evaluator should be careful not to assume that every free-response or open ended test item has TER = 1.00. Very often such items are so worded that only two or three answers are really possible, in which case TER = 0.50 or TER = 0.67.

[4] Terminal Error Rate would be a more descriptive phrase than Instructional Error Rate; however, we have chosen the latter to avoid the ambiguity involved in having TER stand for both Terminal Error Rate and Theoretical Error Rate.

[5] These decisions should not, however, be interpreted too strictly; the evaluator will still have to use some degree of subjective judgment. For example, when we say, in subsequent discussions, that an item should be revised (R), we mean that our best guess on the basis of

the data is that the item should be revised. This does not mean, however, that the item should be revised without a logical basis for revision. Also, when we say that an item (or instruction) is questionable (?), we mean that the data are not sufficient to make a definite judgment about whether or not the item (or instruction) should be revised.

[6] It is also possible that the item has neither of these faults and the objective, while being easy for most of the students, is considered to be an integral part of the total set of objectives. In this case, the item would not be revised. A similar statement can be made for Rule 16 which will be discussed later.

[7] It is also possible that the terminal item and posttest item are not measuring the same content at the same level of difficulty, even though this is an assumption underlying all the decision rules presented here.

[8] The term - 1/2N in Equation 7 is a correction for discontinuity and, as such, can be dropped if the sample size is large. Note that when TER = 1.00 Z is undefined; in this case any value of DER 0 can be considered significant.

[9] One could make a case for using error rate on the posttest item (PER) rather than error rate on the terminal item; then, however, PER - BER would involve a confounding of gain with retention, as we are using the terms in this paper.

[10] We are grateful to Mr. Eugene Millstein for developing the instructional program and collecting the data pursuant to a contract

with the Office of Naval Research, ONR Contract No. N00014-67-A-0298-0003. Our analysis of the data should not, however, be interpreted as an evaluation of the program.

[11] Recall that when derived error rates are available Rules 16-19 replace Rules 1-3 and 12-15, since the former rules are more exact statements of the latter rules. More specifically, Rule 18, in effect, replaces Rule 14. For objective number 11, application of Rule 18 indicates no need for revision, which overrides the decision made on the basis of Rule 14.

The reader will note that, in Table 4, if two or more rules indicate "no revision (NR)", we have identified only that rule which we believe is most important. There seems to be no particular advantage in identifying all the possible reasons for doing nothing!

[12] This program is being used primarily as a vehicle for testing a psychological theory of sequencing instruction. As such, the program has been purposely written to discriminate among students who have experienced different instructional sequences; the program is not meant to teach micro-economics to all students in the most effective manner.

TABLE 1

Rules for Decision-Making

Rule No.	Error Rates						Decision[a]		
	TER	BER	BDI	IER	PER	PDI	Item	Instruction	Prerequisites
1	H	H					NR		
	L	L					NR		
2	L	H					NR		
3	H	L					R		
4			-				?		
5			+						E
			-						E
6					L	0	NR	NR	
7					L	+	?	?	
					L	-	?	?	
8					H	-	R	R	
9					H	+	?	R	
					H	0	?	R	
10						+			E
						-			E
11			-			-	R		

TABLE 1 (cont'd)

Rules for Decision-Making

Rule No.	Error Rates: TER	BER	BDI	IER	PER	PDI	Decision[a]: Item	Instruction	Prerequisites
12				L	L			NR	
13				L	H			R	
14				H	L			?	
15				H	H			R	
16	DER*[b]						R		
	DER(NS)[c]						NR		
17					$RER > c_1$			R	
					$0 \leq RER \leq c_1$			NR	
18					$RER < -c_2$			?	
					$-c_2 \leq RER \leq 0$			NR	
19		$PMPG < c_3$						R	
		$PMPG \geq c_3$						NR	

[a] "NR" means no revision required.

"R" means revision is required.

"?" means the data are not sufficient to make a sound judgment about whether or not revision is required.

"E" means the prerequisites for the objective should be examined.

[b]DER is significantly greater than zero at the .05 level for a one-tailed test of significance.

[c]DER is not significantly greater than zero at the .05 level for a one-tailed test of significance.

TABLE 2

Error Rates and Discrimination Indices for a CAI Program in Micro-Economics

Objective	Raw Error Rates and Discrimination Indices						Derived Error Rates		
Number	TER	BER	BDI	IER	PER	PDI	DER	RER	PMPG
1	1.000	.750	.365	.250	.071	.205	.250*	-.179	.667
2	.875	.964	.304	.143	.643	.880**	-.089	.500	.852
3	.750	.786	.055	.107	.106	.205	-.036	-.001	.864
4	.750	.714	.650**	.214	.036	.141	.036	-.178	.700
5	.500	.604	.786**	.321	.000	.000	-.104	-.321	.469
6	.667	.714	.475*	.107	.179	.015	-.047	.072	.850
7	.750	.893	.548*	.321	.393	.510	-.143	.072	.641
8	1.000	1.000	.000	.286	.607	.535	.000	.321	.714
9	.500	.857	-.032	.071	.179	.309	-.357	.108	.917
10	.500	.500	.000	.000	.214	.309	.000	.214	1.000
11	1.000	.964	.304	.321	.143	.015	.036*	-.178	.667
12	1.000	.929	.132	.286	.429	.459	.071*	.143	.692
13	.500	.821	.737**	.214	.214	.357	-.321	.000	.739
14	1.000	.857	.420	.607	.464	.630*	.143*	-.143	.292
15	.500	.679	.580**	.143	.214	.357	-.179	.071	.789
16	1.000	1.000	.000	.143	.500	.535	.000	.357	.857
17	1.000	1.000	.000	.393	.964	.394	.000	.571	.607
18	.875	.964	.304	.679	.786	.397	-.089	.107	.296

* $p < .05$ ** $p < .01$

TABLE 3

Categorical Error Rates and

Discrimination Indices for a CAI Program in Micro-Economics

Objective Number	Raw Error Rates and Discrimination Indices[a]						Derived Error Rates		
	TER	BER	BDI	IER	PER	PDI	DER[b]	RER[c]	PMPG[d]
1	H	H	0	L	L	0	*	-	-
2	H	H	0	L	H	+	-	GT	-
3	H	H	0	L	L	0	-	-	-
4	H	H	+	L	L	0	-	-	-
5	H	H	+	H	L	0	-	LT	LT
6	H	H	+	L	L	0	-	-	-
7	H	H	+	H	H	0	-	-	-
8	H	H	0	L	H	0	-	GT	-
9	H	H	0	L	L	0	-	-	-
10	H	H	0	L	L	0	-	GT	-
11	H	H	0	H	L	0	*	-	-
12	H	H	0	L	H	0	*	-	-
13	H	H	+	L	L	0	-	-	-
14	H	H	0	H	H	+	*	-	LT
15	H	H	+	L	L	0	-	-	-
16	H	H	0	L	H	0	-	GT	-
17	H	H	0	H	H	0	-	GT	-
18	H	H	0	H	H	0	-	-	LT

[a]The cut-off value for TER, BER, IER, and PER is 0.30.

[b]"-" indicates that DER is not significantly greater than zero at the .05 level for a one-tailed test of significance.

[c]"GT" indicates that RER is "greater then" 0.20.

"LT" indicates that RER is "less than" -0.30.

"-" indicates that $-0.30 \leq RER \leq 0.20$.

[d]"LT" indicates that PMPG is "less than" 0.60.

"-" indicates that PMPG is greater than or equal to 0.60.

* $p < .05$

TABLE 4

Revision Decisions by Objective for a CAI Program in Micro-Economics

Objective No.	Decisions Using All Data[a]			Decisions Using Raw Error Rates and Discrimination Indices[a]			Decisions Using Only Pretest and Posttest Data[a,b]		
	Item	Instruction	Prerequisites	Item	Instruction	Prerequisites	Item	Instruction	Prerequisites
1	R(16)	NR(6)		NR(6)	NR(12)		NR(6)	NR(6)	
2	?(9)	R(9,17)	E(10)	?(9)	R(9,13)	E(10)	?(9)	R(9)	E(10)
3	NR(6)	NR(6)		NR(6)	NR(12)		NR(6)	NR(6)	
4	NR(6)	NR(6)		NR(6)	NR(12)		NR(6)	NR(6)	
5	NR(6)	R(19)		NR(6)	?(14)		NR(6)	NR(6)	
6	NR(6)	NR(6)		NR(6)	NR(12)		NR(6)	NR(6)	
7	?(9)	R(9)		?(9)	R(9,15)		?(9)	R(9)	
8	?(9)	R(9,17)		?(9)	R(9,13)		?(9)	R(9)	
9	NR(6)	NR(6)		NR(6)	NR(12)		NR(6)	NR(6)	
10	NR(6)	R(17)		NR(6)	NR(12)		NR(6)	NR(6)	
11	R(16)	NR(6)		NR(6)	?(14)		NR(6)	NR(6)	

TABLE 4 (cont'd)

Revision Decisions by Objective for a CAI Program in Micro-Economics

Objective No.	Decisions Using All Data[a]			Decisions Using Raw Error Rates and Discrimination Indices[a]			Decisions Using Only Pretest and Posttest Data[a,b]		
	Item	Instruction	Prerequisites	Item	Instruction	Prerequisites	Item	Instruction	Prerequisites
12	R(15)	R(9)		?(9)	R(9,13)		?(9)	R(9)	
13	NR(6)	NR(6)		NR(6)	NR(12)		NR(6)	NR(6)	
14	R(16)	R(9,19)	E(10)	?(9)	R(9,15)	E(10)	?(9)	R(9)	E(10)
15	NR(6)	NR(6)		NR(6)	NR(12)		NR(6)	NR(6)	
16	?(9)	R(9,17)		?(9)	R(9,13)		?(9)	R(9)	
17	?(9)	R(9,17)		?(9)	R(9,13)		?(9)	R(9)	
18	?(9)	R(9,19)		?(9)	R(9,13)		?(9)	R(9)	

[a]"NR" means no revision is required.

"R" means revision is required.

"?" means the data are not sufficient to make a sound judgment about whether or not revision is required.

"E" means the prerequisites for the objective should be examined.

Numbers in parentheses are rule numbers.