DOCUMENT RESUME

ED 126 103 TM 005 342

AUTHOR Reichman, Susan L.; Oosterhof, Albert C.
TITLE Strategy Guidelines for the Construction of Mastery Tests.
PUB DATE [Apr 76]
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Decision Making; *Guidelines; High School Students; Instructional Design; *Models; Pass Fail Grading; Probability; Statistical Analysis; *Student Grouping; Student Testing; *Test Construction; Time
IDENTIFIERS *Mastery Tests

ABSTRACT

Various procedures and guidelines have been suggested for the development and construction of criterion-referenced tests. The present paper proposes a comprehensive model which allows the user to identify and relate specific components which affect the optimal construction and implementation strategies of criterion-referenced tests. Furthermore, it establishes parameter values which will allow the classification of individuals into mastery or nonmastery states with prespecified levels of confidence. Although the discussed model incorporates binomial expansions, it uses parameters of selected items for establishing baselines probabilities instead of true scores derived from an assumed population of items. (Author)

Strategy Guidelines for the Construction of Mastery Tests[1]

Susan L. Reichman and Albert C. Oosterhof
Florida State University

Various procedures have been suggested for the development and construction of criterion-referenced tests. The present investigation proposes a model which allows the user to identify and relate specific factors which affect the optimal construction and implementation strategies of criterion-referenced tests. This model incorporates empirically derived data to establish situational parameters for the proposed model, and then uses this data to illustrate the model's adaptability to an applied situation.

Application of the proposed model has implications both to the areas of instructional design and mastery testing strategies. Major implications to the instructional designer include a means whereby time can be meaningfully appropriated within a course for instruction and student assessment. The model provides the instructional designer with a procedure for determining a feasible number of pass/no-pass decision points to include within a course of instruction on the basis of values established for the individual components within the model. Finally, components which have previously been considered independently, without attention to interrelationships, are incorporated in such a manner within the model that when one component is altered the user becomes aware of the resulting implications to changes in the remaining of the model.

[1] A paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April, 1976.

The Proposed Model

Although the discussed model incorporates binomial expansions, it uses parameters of selected items for establishing baseline probabilities instead of true scores derived from an assumed population of items. Furthermore, whereas parameters associated with a heterogeneous group of subjects were incorporated into the current investigation, procedures are described which allow application of the proposed model to different groups of students who vary in the degree to which they cluster around a criterion score.

The following components are included in the model:

1. Amount of student time or average response latency associated with a specific item format [t_{κ}].

2. Total amount of student time to be allocated to testing [T].

3. The number of items selected for determining pass/no-pass decisions on each decision point [k].

4. The number of mastery-status decisions made within a specific interval of instruction [n_{ω}].

5. Probability for individuals performing at a specified domain score of being placed in the correct mastery state, given the minimum passing percentage score and the average difficulty of items selected for making the pass/no-pass decision on a particular decision point

   $$[p(A)\equiv p(\phi_o,\overline{p}_{(\kappa|\pi_j)})].$$

The relationship between the number of items assigned to a decision point, the total amount of time allocated to testing, the average response latency per item, and the number of mastery-status decisions made within a specified interval of instruction is defined as

$$k= \frac{T}{n_{\omega}t_{\kappa}}$$

The probability of classifying an individual into the correct mastery state is determined by selecting the appropriate terms from the binomial

$$[A+(1-A)]^k$$

Amount of student time associated with a specific item format

The average response latency time required of the student to respond to various types of item formats has an inverse relationship to the number of items that can be presented to the student within an allocated time. Across the range of various format options (such as simple factual multiple choice, true/false, complex multiple choice as requiring problem solving, or complex mathematics and reading selections) estimates of response latency time range from 30 to 300 seconds per item. Within a specified amount of time a student can respond to finite number of items. One should consider this practical limitation when designing criterion-referenced tests.

If the instructor or instructional designer fails to take into consideration response latency of the particular type of item selected various problems could arise. Take for example a situation where fifteen minutes were allocated to testing a decision point and twenty items were required to obtain the desired probabilities of misclassification; the problem -- items selected were of the complex multiple choice type, requiring more than 60 seconds each. Had the designer taken into account the type of item and its respective response latency time various decisions could have been altered: (1) more time could have been allocated for testing, keeping the probabilities and number of items the same, (2) the amount of time and number of items could have been held constant, reducing the probabilities of misclassification, or (3) the amount of time with a shorter response latency time used. By considering the response latency time of selected items along with other components in the model, the most useful combination of components can be selected.

Total amount of student time allocated to testing

The amount of student time the instructor is willing to allocate to testing directly affects the number and type of decision points that can realistically be incorporated within any given course of instruction. Various authors such as Thorndike and Hagen (1969) and Novick and Lewis (1974) suggest that the percent of time that the instructor is willing to allocate to testing is a practical factor which limits test length. Given an upper limit to the total amount of time available for instruction and assessment, a finite number of tests consisting of a specified number and type of items can be given to a student. As an increased number of decision points are incorporated into a set course of instruction, a student must be certified as having attained mastery over a test for each respective decision point. As more tests are incorporated into the course, adjustments must be made with regard to other components in the model; altering the total amount of student time allocated for testing has an observable affect upon other components.

The number of items selected for determining mastery status

The number of items used for determining a student's mastery status within an individual domain can be calculated on the basis of values for the total amount of student time to be allocated to testing, the number of mastery status decisions made within a specified interval of instruction, and the average response latency associated with the specified item format(s). ~~Holding any two of the four variables constant,~~ the instructional designer can alter the third and compute the required value for the fourth component. In this manner values for these four variables can be manipulated and altered until a satisfactory combination is reached.

Number of mastery-status decisions within a specified interval

The number of decision points combined within a course of instruction is inversely related to the number of items that can be theoretically or practically associated with each decision point. Holding the type of item constant~~, as~~ the number of items relating to each decision point is increased more time must be set aside for assessment. Similarly, if the time allocated for testing is held constant for a given course and the number of items for each decision point is increased, the number of feasible decision points decreases.

The total number of decision points which it is feasible to include within a course of instruction is further affected by the actual type of decision point. Hambleton and Novick (1973) indicate that when total testing time is fixed and there is interest in measuring many competencies, the problem arises as to whether one should obtain very precise information about a small number of competencies or less precise information about many more competencies. Cronbach and Gleser (1965) identified this relationship as the bandwidth-fidelity dilemma; bandwidth refering to the variety of information obtained from testing and fidelity as the thoroughness of testing to obtain more complete information. Depending upon where on the bandwidth-fidelity continuum the instructional designer elects to test, varying numbers of items will be required; thus affecting the amount of time to be allocated to testing and the total type and number of decision points that it is practical to assess.

Probability of being placed in the correct mastery state

The probability of a student being placed in the correct mastery state is dependent upon the minimum passing percentage score and the average difficulty of selected items for individual's performing at a specified true domain score. The desired probability of being placed in the correct mastery state is selected on the basis of the implications associated with making a false-positive or false-negative action with respective decision points.

Altering the actual minimum passing level, or criterion level, can be used as a means to make student classifications more or less definitive (Gagne & Briggs, 1974). However, this approach also has its trade-offs. As the criterion level is moved upward toward 100% correct a greater number of mastery people may be classified into the non-mastery group; false-negative actions are more prevalent. As the criterion level is moved downward an increased number of non-mastery people will be classified into the mastery group, resulting in more false-positive actions.

The average difficulty of selected items for individuals performing at a specified domain score is the best estimate of the probability that one given item from a domain will classify an individual into an inappropriate mastery or non-mastery state. The observed difficulty for a given item is dependent upon the observed characteristics of the actual item selected and the domain score under consideration. As various items are sampled and/or different domain scores considered this probability changes. If each item was chosen completely at random from a domain of all potential items and if a given individual had an equal probability of correctly responding to each item in the domain then domain scores derived from an assumed population of items (as in the binomial model) could be used as the probability that one item would correctly classify an individual into mastery or non-mastery catagories. Since these two assumptions are seldom met in the classroom situation it would be more preferable to use the actual item parameters for establishing these baseline probabilities.

Empirical Determination of Probabilities

Collecting empirical data upon which to base binomial probabilities of correctly classifying students into mastery or non-mastery catagories is necessary due directly to assumptions underlying the binomial model. More specifically, these assumptions are that each test item is chosen randomly from a domain of all potential items, and that a given individual has an equal probability of correctly responding to each item that might be selected from the domain.

In most cases, a small number of items from the potential domain are generated to meet a specific purpose in mind. Items are not randomly selected from a large item domain. Furthermore, selected items will tend to vary in difficulty and other characteristics. This is true of even the best defined domains (e.g. multiplication of all single digit numbers includes such items as 2 x 2, 7 x 8, and 5 x 6).

Because these two implicit assumptions are not easily met, procedures based upon an assumed population of items is perhaps not the most appropriate approach to take for determining mastery test length Therefore, it was felt necessary to empirically determine, in a real situation, baseline probabilities derived from parameters of selected items rather than using probabilities derived from an assumed population of items.

Procedure

A computer program was written to carry out the data analysis procedures described below. For each subject's test score each test item was scored dichotomously and a total domain score calculated. The first criterion level was then set and all subjects classified on the basis of that domain and criterion level into mastery or non-mastery catagories. Criterion levels incorporated into the program included 100%, 90%, 80%, 70%, 60%, and 50%.

To calculate the probability, on a single item, of a subject being placed in the correct mastery state, given the minimum passing score, the average difficulty of included items was computed for subjects performing at each possible domain score. This analysis provided the baseline data for binomial expansions to determine probabilities of subject misclassification as test lenghts increased.

At this point test item subsets within the first domain were selected and analyzed to determine the effects upon probabilities of misclassifications as actual test size was reduced. Items included in each subset were not randomly selected, but were selected using a deliberate plan. This was done to more nearly represent the deliberate generation or selection of items that occurs when instructors develop tests with a specific purpose in mind. The size of the item subsets was reduced by intervals of eight until the final subset for the domain equaled eight.

For each test subset size the probability of subject misclassification, given one item, was calculated. Each test subset size was replicated five times with a mean and standard deviation calculated from the resulting five probabilities. The average probability of a misclassification was used as the baseline entry for the binomial expansions to determine probabilities of subject misclassification as test lengths are increased.

Once all test size subsets of the first domain had been analyzed under all six criterion levels a new domain was created by randomly eliminating eight test items. A total of seven domains, ranging in size from fifty-six to eight, were investigated. Within each domain test subsets were again selected and probabilities of subject misclassification determined under each of the six criterion levels.

Subjects

Probabilities of subject misclassification were empirically determined based upon data collected throughout the administration of a 56 item test to 1281 subjects; 49% females and 51% males. These subjects were from 57 volunteer classrooms in the following seven states: Arkansas - 16%, California - 12%, Kansas - 33%, Maine - 12%, New York - 16%, Utah - 7%, ~~and Wisconsin~~ - 4%. Of these 57 classrooms, 29% of the instructors identified their classrooms as rural, 36% as urban, and 35%

as suburban. All subjects were participants in fieldtesting materials developed by the Individualized Science Instructional System (ISIS); a project funded by the National Science Foundation, engaged in the development of discrete instructional modules in a variety of science topics. 92% of the subjects were taking ISIS in their science class for the first time, while 7% had ISIS last year.

The grade levels of the participants ranged from nine to twelve; 18% ninth graders, 55% tenth graders, 16% eleventh graders, and 11% twelveth graders. Ages ranged from 1% twelve to thirteen years, 14% fourteen years, 43% fifteen years, 28% sixteen years, and 15% seventeen or older. 96% of the subjects indicated that they plan to go on to college.

In indicating what the classroom teacher perceived the overall socio-economic level of their class to be in comparison to the nation, 27% of the teachers identified their classes as average, 50% below average, adn 22% above average. In rating the socio-economic level of their classes in comparison to the rest of their school, 74% identified their class as average, 22% below average, and 4% above average.

Development of the test instrument

The student test materials consisted of 56 four response multiple choice items covering the following defined domain of information: the common and scientific names of 14 major bones in the body. An item form approach was used to systematically generate each of the items, with distractors and item test position within four main test sections randomly assigned. The domain was selected from content of the ISIS minicourse Keeping Fit. This domain was selected because it represents a well-defined and finite area of information which could be completely sampled. Within this domain, items, even though generated using an item form approach, would be content include items of varying difficulty levels. Further, the domain was selected because available subjects would have various levels of attainment within the domain, ensuring a wide range of scores on the test. In regard to exposure to the contents of this minicourse, 32% of the subjects indicated that they had done the minicourse this year, 2% had done an earlier version including the domain last year, and 65% had never done the minicourse.

Two major approaches were taken to ensure the validity of the 56 test items. First, an item form approach was used to generate all items. The use of item forms has been identified by various authors as Baker (1974) and Hambleton and Novick (1973) as a systematic approach to establishing content validity. After the actual generation of the items content experts were used to determine whether or not the items for the domain did in fact represent that domain. These content experts were evaluators and writers from the ISIS project staff.

Test Characteristics

Subjects taking the test represented a heterogeneous group with respect to their exposure to the contents of the tested domain, and therefore represented a potential of sampling fairly well the continuum from total non-mastery to complete mastery. The fact that each item constructed for use in the test was carefully matched to intended domain, the consequently observed range of scores gives support to the heterogeneous nature of the sampled subjects.

TABLE 1

Test Parameters

	Item Difficulty	Point Biserial Item Discrimination
No. of Subjects = 1281		
No. of Items = 56	Q_1 = .4280	Q_1 = .5050
Mean = 30.713		
Standard Dev. = 11.851	Median = .5375	Median = .5650
Reliab. (KR-20) = .9270	Q_3 = .6720	Q_3 = .6585

Table 1 summarizes pertinant data concerning test characteristics associated with the total set of 56 items. The mean score of 30.713 corresponds to answering approximately 55% of the items correctly. The internal consistency of the test was found to be quite high as indicated by the reliability coefficient and the item discrimination values. The difficulty levels of a majority of items were contained within a rather narrow range, however a minority of items did range from very easy to very difficult.

The statistics used to describe the test often are not appropriate for domain referenced tests. However, if the range of examinee abilities is large, variability of test scores should be expected, and reference procedures for evaluating the statistical qualities of the test considered acceptable. It is expected that within the constraints of a more homogeneous group of students, e.g. a single classroom, the apparent qualities of the test would be quite different. This last point is discussed later in this paper.

Discussion of Results

Within a well-defined domain, as used in this investigation, as the actual test size was reduced the average probability of misclassifying a student remained fairly consistent, while the standard deviation increased slightly. For example, when determining scores using all 56 items, and calculating average item difficulties using sets of 48 items and setting the criterion level at 100%, an individual with a domain score of 14/56 had an average probability of misclassification of .25 with a standard deviation of .01. In comparison, with a domain size of 56, a test size of 8, and a criterion level set at 100%, an individual with a domain score of 14/56 had an average probability of misclassification of .26 with a standard deviation of .05. Using another comparison example, with a domain of 24 items, a test size of 16 items, and setting the criterion level at 70%, an individual with a score of 14/24 had an average probability of misclassification of .58 with a standard deviation of .01. In comparison, with a domain size of 24 items, a test size of 8 items, and setting the criterion level at 70%, an individual with a score of 14/221 had an average probability of misclassification of .59 with a standard deviation of .02. Using a well-defined domain, sampling error did not appear to have much of an effect upon the average probability of misclassification as the actual test size (i.e. the number of items used to determine the average item difficulties) was reduced.

As the actual domain size was reduced the average probability of misclassifying individuals performing at specific domain levels also remained quite consistent. Table 2 illustrates this by demonstrating that as the number of items used to determine domain scores was reduced from 56 to 16, there were only minor changes in the average probabilities (and associated standard deviations) of misclassifying an individual at various examiner performance levels within the domain. This would suggest that for the present content domain, 56 items provided a fairly stable estimate of baseline probabilities for classifying

Table 2

Probabilities of Student Misclassifications
Various Domain Sizes
Using Criterion Levels of 100% and 50%

	% correct on domain	average probability	standard deviation
Domain Size = 56	75	.75	.05
Test Size = 8	50	.52	.04
Criterion Level = 100%	25	.26	.05
Domain Size = 24	75	.74	.02
Test Size = 8	50	.51	.02
Criterion Level = 100%	25	.26	.01
Domain Size = 16	75	.75	.05
Test Size = 8	50	.50	.06
Criterion Level = 100%	25	.25	.05
Domain Size = 56	75	.25	.05
Test Size = 8	50	.48	.04
Criterion Level = 50%	25	.26	.05
Domain Size = 24	75	.26	.02
Test Size = 8	50	.49	.02
Criterion Level = 50%	25	.26	.01
Domain Size = 16	75	.25	.05
Test Size = 8	50	.50	.06
Criterion Level = 50%	25	.25	.05

individuals into mastery/non-mastery categories at various domain performance levels.

Holding the domain and test size constant, as the criterion level was decreased to .50, the probability of misclassifying non-mastery individuals into the mastery category increased. Similarly, under the same conditions, as the criterion level was increased the probability of misclassifying mastery individuals into the non-mastery category decreased. For example, with a domain score of 56, a test size of.24, and the criterion level set at 90%, there was a .89 probability that an individual scoring just below the criterion level had been misclassified and a .09 probability that an individual scoring above the criterion level had also been misclassified. Under the same domain and test conditions, but with the criterion level set at 50%, there was a .50 probability that an individual scoring just below the criterion level had been misclassified and a .48 probability that an individual scoring just above the criterion level had also been misclassified.

For the criterion levels incorporated in the present study (= .5 to 1.0), individuals with domain scores above the criterion level have the lowest probability of misclassifications; the farther to the right a score is from the criterion level the lower this probability. Individuals scoring just below the criterion level have the greatest probability of being misclassified; this probability again reducing as scores move downward away from the criterion level. These relationships were most predominant as the criterion level deviated away from and above 50%. The binomial model would suggest that the relationship of probabilities above and below the criterion level would be reversed for criterion levels below 50%. For instance, the probability of misclassifying individuals whose domain scores are above the criterion level would be higher than for those which are below the criterion level.

Application of the Model

The instructor or instructional designer, in applying ~~this model~~, should establish values for each of the variables with the model to best fit the conditions of the assessment system to which the model is to be applied. In implementing the model, a user must take into consideration the probable range and distribution pattern of student domain scores in order to incorporate the appropriate probabilities of student misclassification. Presented here is an example of how a designer might use this model to form decisions concerning the optimal criterion-referenced testing strategies. The illustration is limited to using the probabilities which were empirically derived from the present investigation.

An instructor, in allocating time within the total instructional effort, decided to allow no more than eight minutes for assessing initial student mastery over each decision point. It was determined that, for the particular item format to be incorporated, allowing one minute per item would provide sufficient time for students to complete the respective tests. From previous experience, it was estimated that on an initial test, students' scores would be rather symetrically distributed between upper and lower limits of answering 95% to 50% of the test items correctly; more students would be expected to perform at the center of this distribution than at the extremes. It was decided to set the criterion score level at .80.

Using the relationship defined on page 2 between the number of items assigned to a decision point and other variables included in the model, it is determined that the required response latency allows eight items to be used to assess each decision point. Within the estimated range of domain scores, given eight items assigned to each decision point, the probabilities of classifying students into the correct mastery states ranges from .03 to .53. Incorporating the subset of misclassification probabilities, additional weights are assigned to probabilities associated with the center of the domain score distribution. Incorporating the subset of probabilities corresponding to the use of eight items administered to individuals with domain scores ranging from 50% to 95%, the probabilities of misclassifications listed in Table 3 would be appropriate. These probabilities are extracted from the larger set of probabilities derived empirically in conjunction with the present investigation. (Reproduction limitations prohibit reproduction of the complete tables, however, specific sections can be provided individuals upon request.)

Weighting the probabilities to parallel the anticipated distribution of domain scores, the average probability of misclassifying an examinee would be .26 for the conditions described. In other words, approximately one out of four students would be classified into the incorrect mastery state.

If this amount of misclassification was considered unacceptable, the instructor or instructional designer could alter values given to the other variables incorporated into the model. The number of test items used to assess each decision point could be increased. If the model indicated that as a consequence, the amount of time required for testing was excessive, the average breadth of content assigned to each decision point could be reduced. One could also modify the criterion level, or alter the effectiveness of instruction preceding testing in order to change the anticipated distribution of domain scores. The most important contribution to be made by the model is that it provides a means of interrelating the consequences associated with a proposed

Table 3

Probabilities of Misclassification Of Specific Domain Scores Ranging From 50% to 95%

Score[1]	Relative Weight[2]	Probability[3]
26	1	.03
27	1	.04
28	1	.05
29	1	.05
30	1	.08
31	2	.08
32	2	.10
33	2	.11
34	2	.14
25	3	.18
36	3	.18
37	3	.24
38	4	.25
39	4	.29
40	5	.29
41	5	.25
42	4	.38
43	4	.41
44	3	.53
45	3	.41
46	3	.42
47	2	.34
48	2	.25
49	2	.23
50	2	.18
51	1	.16
52	1	.12
53	1	.07

[1] Domain scores are the number of items out of 56 that would be correctly answered.

[2] Relative weight simply reflects a symetrical distribution which is more concentrated at the center than at the extremes.

[3] Probabilities listed are those determined from a domain of 56 items using test sizes of 8 items.

assessment strategy. Then, on the basis of importance of the decision point and potential time allocated, the instructor can consider various combinations of test sizes, criterion levels and actual probabilities of misclassifications. Once various combinations have been looked at, the instructor can analyze present needs and select the best combination of values to be assigned the various components in the model.

Limitations of the Model

Three weaknesses in the present model have been identified. First, there is the matter of allocating an appropriate amount of time within a mastery-testing strategy to those students who fail to surpass the criterion performance on the initial test attempt. In the situation where initial testing occurs in the classroom and retesting is administered other than during class time, (such as in a testing center accessible at the student's discretion and need) time allocation is not a problem. However, consideration must be given to time allocations for retesting when these re-evaluations must occur during class time. At present this allocation of time is left up to the user of the model, as it is a judgement which must be estimated on the basis of previous knowledge concerning characteristics and individual differences of the students involved.

A second limitation of the present model is the establishment of parameters associated with various item forms and formats. Data for the present investigation was limited to multiple choice items written for a well-defined domain of information in a science area and administered to high school science students. As the various types of items used and the domains are changed, the baseline parameters would need to be reestablished.

A third limitation of the present model is the lack of convenient procedures for determining domain score distributions and their associated probabilities of misclassification for various domains and item test formats. Similarly, accurate and detailed information about student performance under these various conditions has not been widely collected for use in determining these domain score distributions and associated probabilities. This type of information is, however, easily obtainable and computer programs or derived tables could be made accessible to the instructional designer.

Implications of the Model

This model is unique in its treatment of three major areas. First, minimal work has been done in the area of determining just how many test items are needed to classify a person into mastery/non-mastery categories with respect to a previously stated area with a given level of confidence. Most significant works in the area eventually rest upon the binomial model in deriving probabilities of student misclassifications. Since the average probability with which a student will correctly answer each item is a function of the specific items utilized or sampled, it would be better to apply a binomial expansion to the probabilities associated with the items actually selected rather than to estimates of population parameters based on supposed domains of equivalent items.

A second unique characteristic of the model is the adaptability to specific types of students. Specifically, as information is collected on how similar groups perform on a specific type of test more efficient decisions can be made in regard to testing the next group. As classes differ in ability and actual performance, the test length and criterion level can be adjusted to accommodate desired probabilities of misclassifications.

The third unique feature of the model is that by using this procedure the instructional developer or individual instructor is provided a means by which the interrelationships of various test-related factors become apparent. Components which previously have been considered independently, without attention to interrelationships, are incorporated in such a manner within the model that when one component is altered the user becomes aware of the resulting implications to changes in the remaining elements of the model. The model provides the user with a method for selecting values for some components in the model and determining the resulting values for the remaining elements. If not satisfied with the first results, the user can repeatedly go through the model, changing values until a satisfactory combination is obtained. In this manner, the user is able to concurrently consider any individual or group of factors that will affect or be affected by any other decisions made during the developmental period of instruction and related criterion-referenced tests used for student assessment.

Directions for Future Research

In regard to establishing probabilities of student classifications, two major areas require further investigation: (1) the effect of various types of item formats, and (2) the effect of other content areas or domains. The effect of using different types of item formats, other than simple multiple choice, for the establishment of baseline probabilities needs to be determined. Similarly, the establishment of these probabilities as different content areas and less well-defined domains are used warrants further investigation. As to student time required to respond, most information about response latency time for various item formats is in the area of estimates. Specific data in regard to well-described types of items and students should be collected to aid the instructional designer or instructor in allocating testing time. Finally, a means by which the model could be easily manipulated by the instructional designer or instructor requires refinement.

As to the actual model, the authors welcome and encourage people to take exception to it. The components selected for inclusion in the model and the manner in which they were integrated represent only one of various possible options. A major advantage to this model is that it provides a way of integrating components to encourage and facilitate the instructional designer or instructor to concurrently consider effects that any decision about one component has upon the remaining elements.

The consequence of not concurrently considering components within the model has resulted in decisions being made about various components with no concern for the effect these decisions have on other elements. Of particular significance is that the instructor and designer quite freely and arbitrarily select the number of items necessary to assess student competence over an area without questioning the probability of an individual being placed into the correct or incorrect mastery state. Further, the instructor or designer is left at the point where guesses are being made in such critical areas as to how many items should be included on a criterion-referenced test, what passing level should be used, and how correct student classifications really are. The development and application of the presently proposed model, or some other comprehensive model which allows the user to identify and relate specific components which affect the optimal construction and implementation strategies of criterion-referenced tests is essential.

Summary

In this investigation a comprehensive model was proposed which would facilitate the instructional designer or individual instructor in concurrently considering various components which affect the construction of criterion-referenced tests used to make pass/no-pass decisions in regard to specified decision points. Components within the model include the average response latency time associated with the specific item format, the total amount of student time to be allocated to testing, the number of items selected for determining pass/no-pass decisions on each decision point, the number of mastery-status decisions made within the course, and the probability for individuals performing at a specified true domain score of being placed in the correct mastery state.

In order to demonstrate the empirical determination of probabilities of correctly classifying individuals into mastery/non-mastery categories, a 56 item test covering a well-defined domain of science information was administered to 1281 high school students. Baseline probabilities were determined under various domain and test sizes, using six different criterion levels in each case. Binomial expansions were then used to determine probabilities as test lengths were increased.

The data collected suggested that when a well-defined domain is established, as the actual domain size was reduced the average probability of misclassifying individuals at specific domain levels remained fairly consistant. Further, as the actual test size was reduced the baseline probability of misclassifying an individual at a given domain level also remained fairly consistant, while the standard deviation increased slightly. Increasing the criterion level resulted in an increase in the probability of misclassifying individuals with domain scores above the criterion level, while decreasing the criterion level resulted in an increase in the probability of misclassifying individuals with domain scores below that criterion level.

Application of the model demonstrated how the designer or instructor could manipulate components within the model in order to select the most efficient combination of factors to meet present needs. It was also demonstrated how this model could be used to make testing decisions for specific types of students based upon estimates of student performance with the content domain. The need for further research in the area of other domains and various types of item formats was pointed out. Most importantly, the need for further work in developing comprehensive models which provide the designer or instructor with easy methods for concurrently considering various test related components has been identified.

References

Baker, Eva. Beyond objectives: domain-referenced tests for evaluation and instructional improvement. Educational Technology, 1974 14(6), 10-16.

Cronbach; Lee J. & Gleser, Goldine C. Psychological tests and personnel decisions, (2nd ed.). Urbana: University of Illinois Press, 1965.

Gagne, Robert M., & Briggs, Leslie J. Principles of instructional design. New York: Holt, Rinehart, & Winston, 1974.

Hambleton, Ronald K. & Novick, Melvin R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10(5), 159-170.

Novick, Melvin R., & Lewis, Charles. Prescribing test lehgth for criterion-referenced measurement. Iowa City: The American College Testing Program, Research and Development Division, 1974. (ACT Technical Bulletin No. 18).

Thorndike, Robert L. & Hagen, Elizabeth. Measurement and Evaluation in psychology and education. (3rd ed.) New York: John Wiley & Sons, 1969.