ED 104 957 TM 004 417

AUTHOR Haladyna, Thomas
TITLE An Analysis of Two Procedures for Decisionmaking When Using Domain-Referenced Tests.
PUB DATE Apr 75
NOTE 22p.; Paper presented at the Annual Convention of the National Council on Measurement in Education (Washington, D. C., April 1975).

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Academic Achievement; Academic Standards; Criterion Referenced Tests; *Decision Making Skills; *Individualized Instruction; Item Sampling; Measurement Techniques; Norm Referenced Tests; Pass Fail Grading; Predictive Validity; Scoring Formulas; *Statistical Analysis; Student Evaluation; Test Construction; *Testing; *Test Interpretation; Test Reliability; True Scores
IDENTIFIERS *Domain Referenced Tests; Millman (J)

ABSTRACT

A central problem for the user of domain-referenced tests in instruction is deciding who has passed and who has failed. Two procedures were presented and discussed. The first, employing classical test theory, was found to be more useful for larger domains and where the passing standard is 70 percent or less. The sampling procedure suggested by Millman (1974) was found to be more applicable when the test size approximates the size of the domain. Neither procedure appears useful when the passing standard is high. In light of the large numbers of examinees classified as uncertain when real test data is used, it was concluded that neither procedure offers much to decisionmaking in systematic individualized instruction. (Author)

An Analysis of Two Procedures for Decisionmaking When Using Domain-Referenced Tests

Thomas haladyna
Teaching Research
A Divisi:on of the Oregon
State System of Higher Education

A Paper Presented at the Annual Meeting of the
National Countil for Measurement in Education, April, 1975.

ABSTRACT

A central problem for the user of domain-referenced tests in instruction is deciding who has passed and who has failed. Two procedures were presented and discussed. The first, employing classical test theory, was found to be more useful for larger domains and where the passing standard is 70% or less. The sampling procedure suggested by Millman (1974) was found to be more applicable when the test size approximates the size of the domain. Neither procedure appears useful when the passing standard is high. In light of the large numbers of examinees classified as uncertain when real test data is used, it was concluded that neither procedure offers much to decisionmaking in systematic individualized instruction.

One problem in systematic individualized instruction is determining which students have passed or failed a test representing an achievement domain. It has been widely advocated that a passing standard (PS) be set for institutional situations and that tests be constructed which carefully correspond to instructional intent (Millman, 1974b; Hambleton and Novick, 1973). Those examinees whose score fall at or near the PS are in jeopardy of being misclassified due to errors of measurement. One solution to this problem is to set a confidence interval around the PS and to make decisions based on how each examinee scores with respect to that confidence interval. Those who score above the confidence interval pass, those who score below fail, and those who score within the confidence interval are given remedial instruction or further testing until their status is ascertained.

The statistical estimation of confidence intervals can be done in a number of ways, depending upon the theoretical orientation. This particular study is limited to two very contrasting approaches; the first is a procedure from classical test theory, and the second is an item sampling technique recently suggested by Millman (1974a).

To begin this analysis, it is necessary to provide a useful definition of a domain-referenced test (DRT) and then briefly describe the instructional context for which the decisionmaking models are advocated. Then the two procedures are examined in light of some fundamental operations in scientific inquiry.

Defining the Construct. Hively (1974 p. 8) has described a domain as "any specified set of items." Millman (1974a) defines a DRT as a random

,ample of items from a domain. However, these brief definitions deserve more attention. It is clear from Millman's extensive treatment of DRT theory (1974b) that the use of instructional objectives was not intended to be the device for the careful specification of a set of items for a domain. What is advocated are item writing rules in the spirit of Hively, Patterson, and Page (1968) and Bormuth (1970). However, the technology for such item generation is neonatal. In the present context, a domain will be considered any set of items that is conceptually related to an instructional unit or intent. The unitary nature of the set of items is a defining trait of the domain, and any random sample of items is a DRT. The domain may be infinite in size, or it may consist of only several items. The latter instance is viewed as extremely unlikely in modern systematic instruction where items and item pools are numerous.

The Instructional Context. The current press for individualized instruction has led to many types of systematic instruction. Some of these are Bloom's mastery learning (1968), Individually Prescribed Instruction (IPI), and Program for Learning in Accordance with Needs (PLAN). The latter two were recently reviewed by Hambleton (1975). Regardless of the specific instructional system employed, most systems allow time-to-learn to vary with the individuals; and most advocate the use of frequent testing, usually prior to and following instruction. Thus a PS is required, and students must ultimately be assigned to a pass or fail category. Despite the fact that this analysis is focused on the decisionmaking issue, the problem of where to set the PS is inextricably connected with the former.

Some Fundamental Operations in Scientific Inquiry. According to Kuhn (1962), science often advances with scientific revolutions. The current trend away from classical test theory and toward new approaches to classroom achievement testing may qualify as a departure from "normal science". The test of the two approaches from a theory viewpoint is couched in logical and statistical criteria. One is initially concerned with the inferences drawn from each approach and the generality of an approach to the wide array of achievement testing situations common to systematic individualized instruction. Later, the theory is tested to see if data fits the model. The interphase between data and theory is a necessary condition in theory verification (Kaplan, 1964). In the present context, two rivaling hypothesis are examined both theoretically and empirically. Classical theory has been rejected by many "new theory" advocates, thus the discussion of classical theory focuses on these criticisms and the nature and scope of .lassical theory. While the item sampling approach is presented as a direct solution to DRT construction and use.

Two Approaches to Decisionmaking

Regardless of the approach considered, the crux of the problem in decisionmaking in this instructional setting is that of knowing about true scores. In classical theory, a true score is the expected observed score. It is estimated from the product of the reliability estimate and the standardized observed score, the result is a regressed true score estimate. In sampling theory, the observed score is considered to be an unbiased estimator of the domain score (analogous to the true score). In other approaches (e.g. Rasch models, baysian approach, and Cronback's theory of generalizability), the true scores are conceptualized differently. With each approach, the standard error (SE) may vary. Thus it is held that the procedure that yields the smaller SE for a wide variety of test situations is probably most effective for decisionmaking. Analogously, when classical reliability is estimated two different ways (KR-20 and KR-21), the latter is an underestimation which leads to overestimated SE's. In the same respect, various approaches lead to estimates of error which may be too large to be useful in instructional decisionmaking. If an approach leads to assigning most students to a failing status or uncertainty, one has to question the usefulness of the model, not on logical or statistical criteria, bu~ on empirical.

Ideally, the rationale for setting a PS should be one of predictive validity. Those scoring above the PS are very likely to be successful in another unit of instruction or on a job. Those not passing have little likehood of future success in the next instructional unit or job. In this ideal situation, the distribution of test scores is bimodal, with noninstructed students scoring at the floor or the scale and instructed students scoring at the ceiling. With the PS set to minimize the errors of misclassification, an approach to setting confidence intervals bears importantly on decisionmaking. This might lead to a suspicion that DR tests might have the optimal PS at the midpoint of the achievement scale.

The Classical Approach. With dichotomously scorable items, the SE is computed using a KR-20 estimate of reliability. In classical theory, the SE does not apply to error surrounding the observed scores. Instead, it refers to the distribution of observed scores around a true score. Taking the PS as a point on the scale where a true score must exist in order to justify a passing status, those persons whose true scores fall at the PS will have observed scores plus or minus two SE's around the PS about 95% of the time. To minimize errors of misclassification, we either continue instruction or provide specific remedial instruction as determined from subscale scores. In other words, we attempt to change the student's status positively so an accurate assignment of "pass" can be made.

Critics of classical theory for such testing have maintained that classical test theory is a norm-referenced (NR) approach to measurement and does not yield the type of information required in a criterion-referenced (CR) situation. According to this argument (e.g. Popham and Husek, 1969; Carver, 1974), a NR test theory yields information about the relative differences among examinees, whereas CR tests yield information about the percentage of tasks (test items) that a student can do (answer correctly) from a well-defined universe of tasks. Donlon (1974) and Millman (1974b) have discussed the semantic difficulties of CR, NR, DR. The problem with definitions of the concepts has made the study of DRT's more difficult.

It has been popularly held that classical test theory leads to NR test interpretations, while item sampling theory leads to DR test interpretations. However, there is evidence to dispute these beliefs. Nunnally (1967) has presented classical theory as a "domain sampling" model. Any test is "a random sample of items from a hypothetical domain of items" (Nunnally, 1967, p. 175). Lord and Novick (1968) have also defined classical theory as a random sampling procedure from a well-defined set of test items. The randomness in sampling items from the domain is quite explicit in theory, although admittedly seldom practiced. Thus it is the use of classical theory rather than the theory itself that appears faulty.

Classical test scales yield two fundamental types of information, absolute and relative. The absolute information is seen as DR, and the relative information is NR. Donlon (1974) among others, has clarified this relationship and extended our understanding of the various test uses to a number of applications. As Ebel (1974) has stated, a test is a test. What we choose to do with the results determines the designation CR, DR, NR.

If a test is constructed in a manner specified in classical theory, it is held that NR or DR interpretations are possible. Based on this argument, classical theory is advocated as a useful approach to decisionmaking along with a constellation of other approaches, many of which were cited briefly earlier in this paper.

The statistical aspects of decisionmaking have been criticized by Popham and Husek (1969) and later by others (e.g. Carver, 1974). The issue here is one of variance. Scores following instruction are said to be restricted to the degree that classical estimates of reliability are useless. Millman and Popham (1974) have also argued that variance is actually an irrelevant concept in CR or DR testing, since the measurement requirements involve only a person's status with respect to a well-defined domain of items. The suspicion that variance is reduced following instruction has not been empirically verified. In fact, when instruction is not as effective as one might hope, the opposite appears true; variance is quite substantial. Woodson (1974a, 1974b) has argued persuasively that the suspected lack of variance may be due to a restricted and inappropriate sampling of examinees. Since the test is calibrated to discriminate between instructed and noninstructed persons, items should be calibrated on the entire range of abilities. This was empirically substantiated with CR tests in one study by Haladyna (1974).

However, the attention given to variance and reliability may be misdirected. As previously noted, test variance has much to do with the estimation of reliability but nothing to do with computing the SE. Reliability is only a device to gain information so a SE can be computed. Since the SE are the important statistics in decisionmaking, and SE are constant regardless of the sample tested; it would seem appropriate to compute a SE from <u>any</u> sample of examinees.

To summarize, the essence for the logical-statistical rationale for useing classical theory in decisionmaking in the DR context is that any achievement test is viewed simply as a means for measurement. What occurs following that measurement is viewed as "NR" "DR" or "CR". Strictly speaking, the definition offered by advocates of DRT theory is semantically identical to that classical theory. In this respect the use of classical theory for the DR test use is entirely consistent.

Item Sampling. The sampling approach for DRT's presented by Millman (1974) is an application of item sampling theory as described in Chapter 11 of Lord and Novick, (1968). The procedure calls for random samples of items for any test to be drawn from a well-defined domain of items. No restrictions are put on the domain, unlike classical theory where homogenity is a useful concept. Some of the assumptions of the item sampling model are: (a) a domain is definable in terms of items which need not be conceptually or empirically homogeneous as a domain is in classical theory, (b) any examinee's score is an unbiased estimator of his domain score, (c) the score and the interpretation of the score are independent of any other examinee's score or of the qualities of the test (i.e. test variance, reliability, and item discrimination). In fact, Millman (1974b) has maintained that the tampering of items in a domain may limit or change the quality of the domain. Item analysis is thereby restricted to locating and discarding or revising defective items. In classical theory, one seeks items that measure a domain through item analysis or similar procedures.

An uncertainty band is constructed which is conceptually analogous to the confidence interval in classical theory. Two UB's, like plus or minus two SE's, forms a 95% confidence interval. It is interesting to note that a classical SE is independent of the PS, while the item sampling UB is dependent on the PS. One reason for the latter is that there is no accounting of the source of measurement error due to ambiguous or non-discriminating items. Therefore, one must conclude that the PS is set for a DRT to adjust for items which are variably discriminating. Further, the PS is adjusted upwards or downwards to compensate for the decreased or increased amount of measurement error arising from variable item discrimination indexes. Without this assumption, the item sampling procedure would not account for a source of error which is built into the classical model.

Millman (1974a) has stated that those students falling in the uncertainty band should be given more test items until their status is determined. The administration of more items decreases the size of the UB so that more pass or fail assignments can be made. If a student has scored at or extremely close to the PS, the number of items needed could be inordinate. Rather than take longer tests, it might be advisable to offer remedial instruction based on subscale information from the test. However, in the classical approach, subscale information has been found to be quite unreliable (Haladyna 1974).

Empirical Aspect of the Analysis

This part of the analysis begins with an application of item sampling to hypothetical situations. The tables constructed (see Table 1) are by no means exhaustive but are meant to be illustrative of a wide variety of instructional situations that are encountered in a great many individualized instruction systems. The second part is an application of both procedures to sets of achievement data which meet the requirements of a DRT. It is in the second phase that the criterion of effectiveness becomes crucial.

In Table 1, UB's are presented for a variety of situations where the domain size is unspecified, 1000, 500, 100, 50 and 25; where the test varies from 5, 10, 20, 30, 50, 75, 100; where the PS varies from 50, 70, 90, 99. The UB is a percentage scale. Large UB's indicate the potential for ineffective decisionmaking whereby too many students are assigned uncertain status and where the confidence zone is too large with respect to the scale.

Table 1 reveals that whenever the PS is high or the test size is extremely small, no one can be assigned a passing status. Therefore, the sampling procedure cannot be applied to these situations. The cutoff for this appears to be in the high 80%'s for most situations. That is, if a PS is higher than around 85%, passing status can seldom be made using the sampling plan.

In situations where test size (n) is small, the UB is also quite large. This result is consistent with classical theory where Haladyna (1974) reported low subscale reliabilities for CR tests. The rest of Table 1 serves to illustrate that when n approaches N and the PS is high, the UB's are very small. When the PS is low, between 50 and 70%, UB's are larger. For example, when a 50% confidence interval is justified and a 30 item test is used to measure a large domain, the uncertainty region includes the range of scores

TABLE 1

Uncertainty Bands as a Function of Domain Size (N), Test Size (n), and the Passing Standard (PS)

	Passing Standard			
(N =∞) n	50	70	90	99
5	22	20	13	4
10	16	15	9	3
20	11	10	7	2
30	9	8	5	2
50	7	6	4	1
75	6	5	3	1
100	5	5	3	1

	Passing Standard			
(N = 1000) n	50	70	90	99
5	22	20	13	4
10	16	14	9	3
20	11	10	7	2
30	9	8	5	2
50	7	7	4	1
75	6	5	3	1
100	5	4	3	1

(N = 500) n	50	70	90	99
5	22	20	13	4
10	15	14	9	3
20	11	10	7	2
30	9	8	5	2
45	7	7	4	1
75	5	5	3	1
100	4	4	3	1

(N = 100) n	50	70	90	99
5	22	20	13	4
10	15	14	9	3
20	16	9	6	2
30	8	7	5	2
45	6	5	3	1
75	3	3	3	1/2
99	1/2	1/2	1/3	0

(N = 50) n	50	70	90	99
5	21	20	13	4
10	14	13	9	3
20	9	8	5	2
30	66	5	3	1

(N = 25) n	50	70	90	99
5	20	19	13	4
10	12	11	8	2
15	8	[illegible]	5	2
20	5	4	2	0

f› m 32% to 68% inclusive. To what degree this interval includes examinees is one of empirical determination. If a bimodal distribution exists representing the instructed and non-instructed groups, then such a model would lend to more effective decisionmaking.

When N is large (over 1000), which is not unusual in terms of present day item and objective banks, the UB's appear considerably larger relative to UB's for similar test lengths from smaller domains. Thus the UB appears most suited for domains of small size (N).

To summarize these findings: (a) the sampling approach does not yield useful decisionmaking capabilities when the PS is high (over 85% approximately); (b) UB's are very large when the PS is low, between 50% to 70%, suggesting that if instruction is less than superior or measurement error is large, far too many students may be found in the uncertainty band; (c) when the domain size is small, and the test size is relatively large, UB's are more effective in decisionmaking than in other situations. The empirical question that arises is what proportion of examinees are given passing, uncertain, and failing assignments when real data is used? We turn to the second phase of this empirical aspect of the analysis.

The data employed here are quite varied and non-representative for all possible DRT's. Nonetheless, the tests are DR, and some inferences may be validly drawn, however limited in generality they are.

The first set of data was taken from an undergraduate measurement course where the tests were CR, the PS was 70%, and the instructional system mastery. Although the tests were objective-referenced, items were pooled into conceptually homogeneous domains, and items were randomly sampled into test forms. Thus the tests were CR and DR by virtue of the defining characteristics of each. Tests were administered before and after instruction and ranged in size from 41 to 45 items. Test characteristics were reported by Haladyna (1974). The subscale information is omitted due to the fact that the SE's were extremely large relative to the variance of the subscales. Since the sampling approach also leads to large UB's with short scales, these data were omitted from further consideration.

The second set of data conforms more loosely to the DRT definition previously given. The tests were of high quality, and the items were objective-referenced. The test forms were drawn from a pool of items representing the domain of dental anatomy, and the items were keyed to a five volume dental anatomy text. The test was administered in a number of dental schools as an achievement test, although the use of a PS is difficult to determine from school to school. Despite these limitations, the tests minimally meet the requirements for a DRT.

In Table 2, the SE's and UB's for the first set of data are presented for one form for each of three instructional units. Also presented is the percentage of students falling in the categories of pass, uncertain, and fail for each approach. The PS actually used was 70%. If this standard had been applied to the students in this classroom testing situation, about the same number of students would be classified in each category regardless of the procedure. As the PS was lowered, the classical procedure proved to be more effective. As the PS was raised, the sampling approach was more effective. Both approaches resulted in far too many students being categorized as uncertain. Since it is assumed that any system that leads to uncertainty about a great number of examinees is less than useful, both approaches must be rejected. On the other han', the fault may lie with the PS. If for purposes of validity or increasing motivation or decreasing anxiety, it is likely that the more appropriate PS should be 50%, the

TABLE 2

Classical Confidence Interval, Uncertainty Band,

Percentages of Passes, Fails, and Uncertains for Three DRT's

Unit One	Classical Confidence Interval	Percentage Falling In			Uncertainty Band	Percentage Falling In		
		Fails	Uncertain	Pass		Fails	Uncertain	Pass
PS								
60	24.88	3	41	58	27.88	3	47	50
65	24.88	6	52	42	27.16	6	47	37
70	24.88	11	67	22	26.16	11	67	22
75	24.88	19	76	5	24.72	19	76	5

Unit Two	Classical Confidence Interval	Percentage Falling In			Uncertainty Band	Percentage Falling In		
		Fails	Uncertain	Pass		Fails	Uncertain	Pass
PS								
60	25.36	2	42	56	28.00	0	55	45
65	25.36	3	68	29	27.28	3	68	29
70	25.36	5	83	12	26.36	5	83	12
75	25.36	11	80	9	24.76	11	80	9

Unit Three	Classical Confidence Interval	Percentage Falling In			Uncertainty Band	Percentage Falling In		
		Fails	Uncertain	Pass		Fails	Uncertain	Pass
PS								
60	24.16	2	63	35	25.32	2	63	35
65	24.16	4	8	16	24.64	4	80	16
70	24.16	7	91	2	23.68	13	85	2
75	24.16	16	84	0	23.68	22	78	0

classical proceaure would lead to a smaller confidence band and thus prove more effective.

Looking at the second set of data, shown in Table 3, two 100 items forms of the test were used. Here, there is less tangible evidence for setting a PS, however, ideally we would use some prior information for the establishment of a PS. Thus we can only speculate that if a PS shown in Table 3 was 70% for form A, 64% of all students would be assigned to a doubtful or failing status, while 77% would have a similar fate using the sampling procedure. Again, the two procedures result in the preponderance of examinees falling in the doubtful or failing ranges. Only when the PS is quite low (65%) does either procedure lead to seemingly efficient decisionmaking. That raises the question: For what reason does one set a PS? Is it to make decisionmaking more efficient, to ensure the valid assignment of examinees to the next step in instruction, to lower anxiety, to motivate? Hopefully future endeavors in decisionmaking will consider some of these variables in a systematic way. To be sure, the data presented in Table 2 and 3 reveal that both classical and item sampling approaches have many serious limitations.

Conclusions

In many respects this analysis has revealed that there is much to be done in the theory of measurement with respect to decisionmaking in individualized systematic instruction. There is little support for either classical procedures or for the item sampling approach as an aid to decisionmaking. Neither appears to meet the criterion of effectiveness, and a relative comparison of the merits of these two would only lead to irrelevant information. In other words, neither approach appears to contribute importantly to decisionmaking in this instructional context.

TABLE 3

Classical Confidence Interval, Uncertainty Band,

Percentage of Passes, Fails, and Uncertains for

Two Parallel Forms DRT's

Form A	Classical Confidence Interval	Percentage of Those Falling in Categories Of			Uncertainty Band	Percentage of Those Falling in Categories of		
		Fails	Uncertain	Pass		Pass	Uncertain	Fails
PS								
60	11.68	2	23	75	15.20	11	31	68
65	11.68	6	35	69	14.80	6	47	57
70	11.68	17	47	36	14.20	13	64	23
75	11.68	45	35	20	13.40	24	53	23
80	11.68	40	56	6	12.40	55	40	5

Form B	Classical Confidence Interval	Percentage of Those Falling in Categories of			Uncertainty Band	Percentage of Those Falling in Categories of		
		Fails	Uncertain	Pass		Pass	Uncertain	Fails
PS								
60	15.44	1	6	93	15.20	1	6	93
65	15.44	1	35	64	14.80	1	35	64
70	15.44	6	48	46	14.20	6	48	46
75	15.44	19	59	22	13.40	19	52	29
80	15.44	31	59	10	12.40	39	49	12

This analysis does raise several crucial issues in test development and decisionmaking. In many respects, the discussion of DRT's has been reduced to instances where domains are loosely defined. It is clear from the writing of Hively et.al. (1973) and Bormuth (1971) that much more was intended in DR testing. The issue that arises in classical theory and the DR approach is how to define a domain. In the classical approach, one looks for items that measure well the hypothetical domain. This procedure is much like the development and validation of a construct: (a) define the construct abstractly, (b) hypothesize measures of that construct, (c) test to see if observables (items) measure that construct. High interitem correlations are essential in establishing the content (factorial) validity as well as construct validity of the items and tests. Items that don't belong to the domain have low discrimination indexes and are discarded or rejected. This is similar to the case in Rasch scaling, where items either fit or don't fit the latent trait. In the domain-referenced approach, the domain is rigorously defined via item forms or item writing rules and items generated in conformance. It is assumed that the rigor that goes into the item construction procedures will yield better measures. The hypothesis that any procedure leads to better measures needs to be empirically tested.

Finally, a number of procedures were very briefly described as approaches to decisionmaking. It would be useful to test the applicability of these approaches with test data. The Baysian approach offers a procedure which is a threshold loss function rather than a squared-error less function. The former is said to lead to smaller SE's in decisionmaking (Hambleton & Novick, 1973); if so and to what degree is largely indeterminate at the present. In Rasch model, SE's become small when an examinee is matched to test items, that is, he misses 50% of the items. This is contrary to the principle of randomly sampling items from a domain which is prominent in classical theory and item sampling theory. In Cronbach's theory of generalizability, test scores are used to estimate universe scores and are regressed depending upon the group from which the examinee came. Again, the question arises, what is the relative degree of error of classification? The problem remains to be studied.

Finally, the problem of where to set the PS is crucial to the decision-making process as revealed by much of the data presented in this analysis. In many respects, if discrimination between instructed and non-instructed students is desired, setting the PS at the midpoint of the achievement scale appears to be most justifiable. The bimodal distribution has the fewest examinees at the middle of the scale. In this situation, it is clear that the classical approach works more effectively.

REFERENCES

Bloom, B. S. Learning for mastery. *Evaluation Comment*, *1*, 1-12.

Bormuth, J. R. On the theory of achievement test items. Chicago: U. of Chicago Press, 1970.

Carver, R. P. Two dimensions of tests, psychometric and edumetric. *American Psychologist*, 1974, *29*, 512-518.

Donlon, T. F. Some needs for clearer terminology in criterion-referenced testing. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974.

Ebel, R. L. Evaluation and educational objectives, *Journal of Educational Measurement*, 1973, *10*, 273-279.

Haladyna, T. M. An investigation of full and subscale reliabilities of criterion-referenced tests. A paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974.

Hambleton, R. K. A review of testing and decisionmaking procedures for selected individualized instructional programs. *Review of Educational Research*, 1975.

Hambleton, R. K. and Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, *10*, 159-170.

Hively, W. Introduction to domain-referenced testing. *Educational Technology*, *14*, 5-9.

Hively, W., Patterson, H. L., & Page, S. H. Generaliz bility of performance by job corps trainees on a universe-defined system of achievement tests in elementary mathematical calculation. Paper presented at the annual meeting of the American Educational Research Association, February, 1968.

Kapl A. *The conduct of inquiry*. San Francisco: Chandler, 1964.

Kuhn, T. S. *The structure of scientific revolution*. Chicago: U. of Chicago Press, 1962.

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

Millman, J. Determining test length: Passing scores and test lengths for domain-referenced tests. *Review of Educational Research*, 1973, *43*, 205-216.

Millman, J. Sampling plans for domain-referenced tests. *Educational Technology*, 1974(a) *14*, 17-21.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.) Evaluation in education. San Francisco: McCutchan, 1974(b).

Millman J. and Popham, W. J. The issue of item and test variance for criterion-referenced tests: a clarification, Journal of Educational Measurement, 1974, 11, 137-138.

Nunnally, J. Psychometric theory, New York: McGraw-Hill, 1967.

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

Woodson, M.I.C.E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 63-64.

Woodson, M.I.C.E. The issue of item and test variance of criterion-referenced tests: a reply. Journal of Educational Measurement, 11 139-140.