DOCUMENT RESUME

ED 261 076 TM 850 478

AUTHOR	Silva, Sharron J.
TITLE	A Comparison of Traditional Approaches and Item Response Approaches to the Problem of Item Selection for Criterion-Referenced Measurement.
PUB DATE	Apr 85
NOTE	23p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Analysis of Variance; Comparative Testing; *Criterion Referenced Tests; Elementary Education; Hypothesis Testing; *Item Analysis; *Latent Trait Theory; *Mastery Tests; Mathematical Models; Reading Tests; Test Construction; Test Items; *Test Reliability; Test Theory
IDENTIFIERS	Prescriptive Reading Inventory

ABSTRACT

Test item selection techniques based on traditional item analysis methods were compared to techniques based on item response theory. The consistency of mastery classifications in criterion referenced reading tests was examined. Pretest and posttest data were available for 945 first and second grade students and for 1796 fourth to sixth grade students who were given the PRI Reading Systems Instructional Objectives Inventory. Three traditional item analysis procedures were used: the Cox-Vargas index; the point biserial correlation applied to combined pretest-posttest data; and the phi coefficient. Two approaches derived from item response theory were also used: item information at the cut-off score as estimated by the one parameter and the three parameter models. The results indicated that the two item response models produced classification consistencies, indicated by coefficient kappa, that were superior to the three traditional procedures. Furthermore, the three parameter model appeared to be superior to the one parameter model. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED261076

A Comparison of Traditional Approaches and Item Response Approaches to the Problem of Item Selection for Criterion-Referenced Measurement

SHARRON J. SILVA, PH.D.

Evaluation Associate

American Red Cross

April 1985

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
S. J. Silva

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
- Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 850 478

This study examined the effect of applying traditional item selection techniques and item selection techniques based on item reponse theory to criterion-referenced instruments by comparing the consistency of the resulting mastery classifications. The findings suggest that item response models are better at preserving mastery classifications (as measured by Coefficient Kappa) than the three traditional approaches included in the study.

Introduction

Two global objections have been raised as barriers to the application of quantitative item selection procedures to criterion-referenced measures:

1. Such procedures typically are affected by restrictions in the variance of the test or the items. For criterion-referenced measures, sufficient item and score variability are not always reasonable expectations (Popham & Husek, 1969; Crehan, 1974; Hambleton, Swaminathan, Algina & Coulson, 1978); and

2. To ensure that the test score can be interpreted as intended, maintaining a carefully balanced distribution of content is generally viewed as essential (Cox, 1964; Cox & Vargas, 1966; Popham & Husek, 1969; Popham, 1975; Popham, 1978; Hambleton, Swaminathan, Algina & Coulson, 1978; Berk, 1980).

In light of these two issues, many theorists have recommended that if item analysis results and the judgment of content experts disagree, the content experts should prevail (Cox & Vargas, 1966; Ebel, 1968; Popham & Husek, 1969; Millman & Popham, 1974; Millman, 1974; Kifer & Bramble, 1974). However, even in this limited context, there is little agreement as to which item analysis procedure should be used. This study compared three traditional item analysis procedures (the Cox-Vargas index, the point biserial correlation applied to combined pretest-posttest data, and the phi coefficient) to two approaches derived from item response theory (item information at the cut-off score as estimated by the one parameter and three parameter models).

Research has offered little evidence to establish any one of these five procedures as best for criterion-referenced tests, though approaches based on item response theory have two conceptual advantages over traditional approaches to item analysis, particularly with respect to the two issues cited above.

1. Item parameters for the item response models are presumed to be invariant across examinee populations (Wright & Panchapakesan, 1969; Hambleton & Gorth, 1971; Wright, 1977; Baker, 1977; Hambleton & Cook, 1977; Lord, 1980); and
2. Once an item set has been calibrated, an unbiased estimate of examinee ability can be secured from any subset of items (Wright & Panchapakesan, 1969; Hambleton & Cook, 1977).

These conceptual advantages have been explored more extensively for norm-referenced tests than for criterion-referenced tests, since most research examining the applications of item response models to criterion-referenced instruments encountered problems with item parameter estimation that interfered with the completion of the studies.

For the three parameter model, these problems include low discrimination values (Bejar, Weiss & Kingsbury, 1977; McKinley & Reckase, 1980), high or unestimatable c values (Bejar et al., 1977; McKinley & Reckase, 1980; Douglass, 1981), and low item difficulties (McKinley & Reckase, 1980). For all of these studies, parameter estimates were based on posttest data for a college-level respondent population. Thus, it is plausible that these item parameters were poorly estimated for lack of sufficient observations throughout the ability range measured by the test. For the one parameter model, the primary concern was the tendency for this model to reject a high percentage of items as non-fitting (Kifer & Bramble, 1974; Van der Linden, 1981).

Such implausible item parameter estimates have been an obstacle to research and to the wider application of these models for test development. In this regard, the present stuay has attempted to examine these problems by addressing two important issues relating to the application of item response theory to criterion-referenced instruments: (1) Can the item estimation problems seen in previous applications of item response models to criterion-referenced tests be overcome by using a more diverse respondent population? (2) Is there evidence that the two item analysis procedures based on item response theory are superior to the three traditional item analysis approaches in selecting item that best maintain consistent mastery classifications?

Methodology

The original mastery criterion used in this study was based on a set of items determined by content specialists to measure a specific trait. Less relevant items were added to this initial set to simulate invalid items, since the original item pool was taken from a standardized test which was unlikely to contain problem items. Thus, a preliminary item pool was constructed with two component parts:

Set 1 (Criterion Set): A set of 50 or more items appropriately sampled from a relatively homogeneous domain, and

Set 2 (Less Valid Items): A smaller set of items, less appropriate than those in Set 1 for measuring the domain.

The five item selection procedures were then applied to this preliminary item pool (Set 1 + 2). Each of the five item analysis procedures was applied to the pooled Set 1 and Set 2 items under circumstances intended to represent optimal applications. If previous research or expert opinion suggested potential problems in the application of any of these procedures, pains were taken to ensure that such circumstances were avoided by constructing a subpopulation for each statistic that took such potential difficulties into account.

Item were ranked by all statistics at each of four cut-off scores determined to be appropriate to the range of ability covered by the instrument. Each of these cut-off scores reflects an extension of the scoring procedures recommended for determining mastery on the test.

For each item analysis procedure, a long and short test was constructed of the items ranked as best by each procedure. For the traditional item analysis procedures, new cut-off scores were developed for these resulting tests through the use of discriminant analysis. For the item response models, appropriate equations were used to determine cut-off scores. Finally, students were ranked as masters or non-masters on each of the resulting tests and these mastery decisions were compared to mastery decisions made based on the original criterion-referenced test. Coefficient kappa was selected as the appropriate measure of classification consistency since this statistic controls for the level of agreement that would be expected by chance (Swaminathan, Hambleton, and Algina, 1974).

An analysis of variance was conducted using these kappas as the outcome measure. This analysis of variance included one fixed-level

ERIC

treatment factor, the item analysis statistic used to generate each test. In addition, three blocking factors were included to reduce unexplained variance. These blocking factors were 1) the test used, 2) the scoring criterion (50 percent, 62.5 percent, 75 percent, or mastery by objective), and 3) the length of the test (original length or shortened). The ANOVA contained one observation per cell, thus no direct estimate of the error term was available. The third-order interaction was used as a substitute for the error term. This approach results in a conservative test of the hypothesis (Dayton, 1970).

Only one contrast was tested. This contrast reflected the hypothesis that the classification consistency resulting from tests developed using item response models was greater than the classification consistency resulting from tests developed using the traditional statistics.

Data Source

The Word Attack and Usage Subtests for the PRI Reading Systems Instructional Objectives Inventory, Levels B and D, were used as the original criterion-referenced instruments. These items are referred to in this study as Set 1 items. The less valid items added to these subtests to simulate invalid or defective items (Set 2 items) were taken from other sections of the same tests. For the Level B test, pretest and posttest responses were available for 945 students, primarily in Grades 1 and 2. For Level D, pretest and posttest responses were available for 1[illegible]96 students, primarily in Grades 4-6. Various subpopulations were constructed to represent optimal applications of each item analysis procedure.

Findings

1. Item Parameters

Table 1 shows the distribution of the item discrimination parameters for Set 1 (original subtest) and Set 2 (less valid) items in Levels B and D. The item parameters obtained for the Level B items differ considerably from those described in most previous studies. In contrast

Table 1

Distribution of
Item Discrimination Parameters
Three Parameter Model

	LEVEL B		LEVEL D	
Range	Set 1 Items	Set 2 Items	Set 1 Items	Set 2 Items
0.0<=a<0.2	0	1	1	0
0.2<=a<0.4	0	4	0	0
0.4<=a<0.6	0	8	10	1
0.6<=a<0.8	2	1	11	3
0.8<=a<1.0	9	1	15	4
1.0<=a<1.2	19	0	15	5
1.2<=a<1.4	11	0	9	1
1.4<=a<1.6	10	0	6	1
1.6<=a<1.8	8	0	0	0
1.8<=a<2.0	6	0	1	0
2.0(maximum)	15	0	0	0
TOTAL	80	15	68	15

(Level B, 945 respondents; Level D, 891 respondents) to the study by Bejar et al. (1977), where a large proportion of items were rejected because they had a values below .8, only two of the 80 items in Set 1 (Level B) had such low discrimination values; however, in Set 2 (Level B), fourteen of the fifteen items had low discrimination values. This is a plausible finding, in that one would expect the Set 2 items, which are drawn from a different subtest, to discriminate Word Attack and Usage skills poorly. In contrast, the item parameter estimates obtained for the Level D test differ from those obtained for Level B. For example, 22 of the Set 1 items (32 percent) have low discrimination values (below .8). In addition, the item discriminations for Set 1 and Set 2 items are quite comparable for Level D.

This table also points out a different type of estimation issue that emerged in the present study. For Level B, Set 1, fifteen items (19 percent) have been set to the maximum discrimination (AMAX) value of 2.0. It is likely that these high discrimination values reflect the wide ability range of the population used to generate these parameter estimates. In general, it is not desirable to have a large number of a's set to this maximum value (Wingersky, Barton & Lord, 1982); however, this LOGIST run was reviewed by Martha Stocking and Frederick Lord, the program designers, who felt that rerunning the program to raise the AMAX value would not substantially improve the estimation of these parameters (personal correspondence, April 17, 1984).

Table 2 presents the distribution of item difficulty parameters for Levels B and D, Sets 1 and 2.

Table 2

Distribution of
Item Difficulty Parameters
Three Parameter Model

	LEVEL B		LEVEL D	
Range	Set 1 Items	Set 2 Items	Set 1 Items	Set 2 Items
-3.0<=b<-2.5	0	0	1	0
-2.5<=b<-2.0	0	1	8	0
-2.0<=b<-1.5	0	0	7	1
-1.5<=b<-1.0	8	2	9	1
-1.0<=b<-0.5	11	4	8	1
-0.5<=b< 0.0	24	3	12	0
0.0<=b< 0.5	28	3	17	1
0.5<=b< 1.0	9	1	5	6
1.0 or more	0	1	1	5
TOTAL	80	15	68	15

(Level B, 945 respondents; Level D, 891 respondents)

In contrast to the a parameter estimates, for Level B, the b values show no marked difference between Set 1 and Set 2. Theoretically, the two sets of items should not differ in terms of difficulty, since both sets were developed for administration to the same students. The values for Level D in Table 2 suggest that the Set 2 items for Level D are distinguished from the Set 1 items primarily by their greater difficulty. Overall, the range of item difficulties for Level D is somewhat greater than the range for Level B. In contrast to previous research, the Set 1 items for both Level B and Level D appear well matched to the ability level of the sample of students. This may be explained by the fact that the respondent pool used to estimate these item parameters included both pretest and posttest responses.

The c parameter estimates for Levels B and D are shown in Table 3. As can be seen from the table, the c values for twelve of the fifteen Set 2 items in Level B were unestimatable and thus were assigned by the program to the mean value of the remaining estimated c values. In contrast, only nine of the 80 Set 1 items (11 percent) had c values that could not be estimated. C values were unestimatable for 39 of the Level D Set 1 items (57 percent).

Table 3

Distribution of
Pseudochance Parameters
Three Parameter Model

	LEVEL B		LEVEL D	
Range	Set 1 Items	Set 2 Items	Set 1 Items	Set 2 Items
0.00<=c<0.05	4	0	0	0
0.05<=c<0.10	5	0	0	0
0.10<=c<0.15	8	0	4	1
0.15<=c<0.20	18	1	7	1
0.20<=c<0.25	19	0	6	5
0.25<=c<0.30	14	2	5	5
0.30<=c<0.35	3	0	1	0
0.35<=c<0.40	0	0	4	0
0.40<=c<0.45	0	0	2	0
Unestimated c values	9	12	39	3
Total	80	15	68	15

(Level B, 945 respondents; Level D, 891 respondents)

Thus, the Set 1 item parameters for Level D present some of the same concerns noted in previous research, despite the fact that these parameter estimates were secured using a pooled pretest-posttest population: 32 percent of the a values were below .8 and 57 percent of the c values were unestimatable. The Set 1 item parameter estimates for Level B present no such concerns..

There are at least three possible explanations for these differences between Level B and Level D. First, the inclusion of fifteen Set 2 (less valid) items in the calibration may have had a greater effect on the item parameter estimates for Set 1 Level D by affecting the dimensionality of the item pool. In this regard, it should be noted that the Set 2 items for Level B were all selected from the same subtest. For Level D, the Set 2 items were selected from several different parts of the test. Thus, if dimensionality were the problem, one would expect it to appear in the Level B data as well.

A second possible explanation is the fact that the Level D respondent population shows more overlap between pretest and posttest than Level B. The median difference between pretest and posttest scores for Level B was 17 points; for Level D the median difference was 6 points.

Finally, it is possible that at the later grades represented by the Level D test, the skills required to respond to questions about Word Attack and Usage are not as distinct from other areas of the test. One related finding that tends to confirm this last possibility is the fact that the traditional item selection statistics were also more consistent in eliminating Set 2 items for Level B than in eliminating Set 2 items for Level D.

ERIC

It is instructive to compare these findings with the results of the one parameter analysis of these same items (Table 4). In past research relating to criterion-referenced instruments, one difficulty encountered with the one parameter model has been the large percentage of items rejected as non-fitting. In this study, this was true for the Set 2 items but less true for Set 1 items.

Table 4

Distribution of
Difficulty Parameters
One Parameter Model

	LEVEL B		LEVEL D	
Range	Set 1 Items	Set 2 Items	Set 1 Items	Set 2 Items
-3.5<=b<-3.0	0	0	1	0
-3.0<=b<-2.5	0	0	2	0
-2.5<=b<-2.0	1	0	3	0
-2.0<=b<-1.5	3	0	4	1
-1.5<=b<-1.0	9	0	7	0
-1.0<=b<-0.5	6	1	9	2
-0.5<=b< 0.0	19	0	7	0
0.0<=b< 0.5	14	0	12	0
0.5<=b< 1.0	17	0	10	1
1.0<=b< 1.5	5	0	7	5
1.5<=b< 2.0	0	0	1	2
2.0<=b< 2.5	0	0	1	0
Non-fitting Items	6	14	4	4
TOTAL	80	15	68	15

(Level B, 492 respondents; Level D, 487 respondents)

Fourteen of the fifteen Set 2 items included in the analysis for Level B were rejected as non-fitting. The basis for this classification was the total t-fit statistic produced by BICAL. (Note that Wright, Mead and Bell (1980) recommend a cut-off value of 2.0 for distinguishing such non-fitting items.) Only six of the 80 Set 1 items were identified as non-fitting. For Level D, eight of the items were identified as non-fitting. Four of these non-fitting items were from Set 2. Thus, only six percent of the Set 1 (Level D) items were identified as non-fitting.

Comparison of Consistency of Mastery Classifications

As explained previously, an analysis of variance was conducted to compare the consistency of mastery classifications from all resulting tests with the original mastery classifications arrived at using only the Set 1 items. The details of this analysis of variance, including the main effect and blocking variables, are presented in Table 5. As can be seen from the table, the contrast of interest (comparing the item response models and traditional models) is significant, as are all of the blocking variables. Indeed, the contrast in question accounts for nearly all the variance associated with the main effect.

Despite this finding, it is not clear from this study that the improved classification consistency for the item response models is a direct result of better item selection. An examination of the items elminated by each item analysis procedure does not suggest that the two item response models were better at identifying Set 2 (less valid) items for elimination. Several explanations should be considered. First, it is possible that the Set 2 items were not uniformly the worst items,

Table 5

Analysis of Variance
Comparing Classification Consistency
Resulting from Tests Based on
Item Response Versus Traditional
Item Selection Approaches

Source	SS	DF	MS	F	P
MAIN EFFECT					
Item Analysis Procedure	.033494	4	.008374	15.04	<.001
BLOCKING FACTORS					
Level	.045960	1	.045960	82.57	<.001
Mastery Criterion	.736586	3	.245529	441.10	<.001
Length	.023096	1	.023096	41.49	<.001
ERROR ESTIMATE					
Third-Order Interaction	.006680	12	.000557		
CONTRAST					
Traditional vs. IRT	.033423	1	.033423	60.01	<.001

particularly at those criteria where their relatively poor discrimination might be counteracted by their more appropriate difficulty. For Level D, the fact that none of the five item analysis procedures consistently identified most of the Set 2 items for elimination tends to support this possibility.

A second factor that should be considered in interpreting the

ERIC

significance of the contrast in Table 5 is the possibility that the procedures used to set the various cut-off scores had an effect on consistency of mastery classifications. For the item response models, there are clear formulas for setting comparable cut-off scores for various tests. No such formulas are available for the traditional item analysis procedures. While every attempt was made in this study to select the best cut-off score, it is possible that other cut-off scores might have produced more consistent mastery classifications.

Table 6 shows the means and standard deviations of kappas for the tests produced by the five item analysis procedures.

Table 6

Mean Kappas for Tests
Produced by Five Item Analysis Procedures

Item Analysis Procedure	Mean Kappa	Sigma	Number of Cases
Cox Vargas Index	.8022	.1119	16
Point Biserial Correlation	.8040	.1003	16
Phi	.8051	.0959	16
Three Parameter Model	.8452	.1674	16
One Parameter Model	.8458	.1075	16

ERIC
Full Text Provided by ERIC

The table shows that the two item response models produced mean kappas of .85 while the traditional models produced mean kappas of .80 to .81. These results appear to suggest that for this type of application there is no advantage to using the three parameter model in preference to the one parameter model. However, it should be observed that the standard deviation for the kappas for the three parameter model is somewhat larger than the standard deviations associated with the other item analysis procedures. This large standard deviation is primarily due to the kappas for the mastery-by-objective scoring criterion. This scoring criterion was the only one for which weighted scoring could not be used to determine the initial mastery classifications; thus, the original mastery designations were not arrived at within the context of the model. When the four kappas relating to the mastery-by-objective criterion are eliminated, the mean for the three parameter model is .90 and the standard deviation is .06. This suggests that the three parameter model may have advantages over the one parameter model that are not evident from the table above.

It is instructive to examine the number and percent of students classified as masters by tests developed using each of the five item analysis procedures at each mastery criterion. These data are presented in Table 7. For the three parameter model, item-weighted scoring was used to arrive at the original mastery decisions instead of the number-right scoring used for the remaining four item analysis procedures. For those criteria where this item weighting procedure was used, the three parameter model consistently classified more students as masters.

Two other features of this table merit comment. First, it is clear from Table 7 that the objective-based scoring procedure produced the least

Table 7

Number and Percent of Students
Classified as Masters
Using Tests Developed by Each Item Analysis Procedure

(Classification Consistency Subpopulations)

Item Analysis Procedure	Original Subtest	Long Test	Short Test
	LEVEL B AT 50 %		
COXV	656 (82.3%)	678 (85.1%)	675 (84.7%)
PTB	656 (82.3%)	686 (86.1%)	677 (84.9%)
PHI	656 (82.3%)	686 (86.1%)	677 (84.9%)
1P	656 (82.3%)	669 (83.9%)	664 (83.3%)
3P	702 (88.1%)	704 (88.3%)	698 (87.6%)
	LEVEL B AT 62.5 %		
COXV	576 (72.3%)	563 (70.6%)	575 (72.1%)
PTB	576 (72.3%)	568 (71.3%)	569 (71.4%)
PHI	576 (72.3%)	572 (71.8%)	577 (72.4%)
1P	576 (72.3%)	583 (73.1%)	593 (74.4%)
3P	626 (78.5%)	631 (79.2%)	627 (78.7%)
	LEVEL B AT 75 %		
COXV	466 (58.5%)	530 (66.5%)	549 (68.9%)
PTB	466 (58.5%)	539 (67.6%)	548 (68.8%)
PHI	466 (58.5%)	528 (66.2%)	529 (66.4%)
1P	466 (58.5%)	477 (59.8%)	487 (61.1%)
3P	520 (65.2%)	519 (65.1%)	522 (65.5%)
	LEVEL B OBJECTIVE		
COXV	325 (40.8%)	230 (28.9%)	221 (27.7%)
PTB	325 (40.8%)	227 (28.5%)	239 (30.0%)
PHI	325 (40.8%)	239 (30.0%)	274 (34.4%)
1P	325 (40.8%)	231 (29.0%)	254 (31.9%)
3P	325 (40.8%)	303 (38.0%)	299 (37.5%)

(Based on responses from 797 students)

ERIC

Table 7 (Continued)

Number and Percent of Students
Classified as Masters
Using Tests Developed by Each Item Analysis Procedure

(Classification Consistency Subpopulation)

Item Analysis Procedure	Original Subtest	Long Test	Short Test
	LEVEL D AT 50 %		
COXV	862 (88.1%)	794 (81.2%)	772 (78.9%)
PTB	862 (88.1%)	826 (84.5%)	790 (80.8%)
PHI	862 (88.1%)	804 (82.2%)	785 (80.3%)
1P	862 (88.1%)	883 (90.3%)	862 (88.1%)
3P	909 (92.9%)	913 (93.4%)	908 (92.8%)
	LEVEL D AT 62.5 %		
COXV	735 (75.2%)	696 (71.2%)	679 (69.4%)
PTB	735 (75.2%)	721 (73.7%)	700 (71.6%)
PHI	735 (75.2%)	683 (69.8%)	655 (67.0%)
1P	735 (75.2%)	758 (77.5%)	797 (81.5%)
3P	799 (81.7%)	807 (82.5%)	802 (82.0%)
	LEVEL D AT 75 %		
COXV	533 (54.5%)	573 (58.6%)	549 (56.1%)
PTB	533 (54.5%)	598 (61.1%)	580 (59.3%)
PHI	533 (54.5%)	558 (57.1%)	550 (56.2%)
1P	533 (54.5%)	582 (59.5%)	626 (64.0%)
3P	598 (61.1%)	619 (63.3%)	611 (62.5%)
	LEVEL D OBJECTIVE		
COXV	384 (39.3%)	385 (39.4%)	395 (40.4%)
PTB	384 (39.3%)	362 (37.0%)	362 (37.0%)
PHI	384 (39.3%)	377 (38.5%)	371 (37.9%)
1P	384 (39.3%)	410 (41.9%)	438 (44.8%)
3P	384 (39.3%)	602 (61.6%)	603 (61.7%)

(Based on responses from 978 students)

consistency in the percentage of students classified as masters across all five item analysis procedures. A second striking feature of this table is the level of consistency of the number of students rated as masters by the three parameter model. Eliminating the objective-based scoring criterion from consideration, there is no more than two percent difference in the percentage of students classified as masters by each of the three tests (original subtest, long test and short test) for the three parameter model applications.

Conclusions

These findings suggest that more plausible item parameter estimates can be obtained for item response applications to criterion-referenced tests when calibration samples of sufficient diversity are used. In addition, the ability of the two models to distinguish items drawn from the Word Attack and Usage Subtest from items drawn from other sections of the test suggest that the item characteristics measured by the one and three parameter models are relevant to the purposes of criterion-referenced testing.

With respect to the final comparison of mastery classifications, the two item response models did produce classification consistencies that were superior to the three traditional item analysis procedures included in this study. The mean kappa for the item response procedures was .85, compared to a mean of .80 for the three traditional approaches ($p<.001$). This result suggests than when there is sufficent variance in the respondent population to produce good item parameter estimates, the one and three

ERIC

parameter models produce better classification consistency than the three traditional item analysis approaches. It is not clear whether this is solely a consequence of better item selection or whether this superior classification consistency may also be the result of better procedures for setting comparable cut-off scores for different tests.

While the focus of the study was not on comparisons between the two item response models, there is evidence to suggest that the three parameter model may be superior to the one parameter model in producing consistent mastery classifications. Within the context of the three parameter model, the agreement in mastery classification between various related tests is impressive. The mean kappa for the three parameter tests scored using item-weighted scoring was .90. The variation in the percentage of students classified as masters by comparable tests was less than two percent.

ERIC

REFERENCES

Baker, F. (1977). Advances in item analysis. Review of Educational Research, 47, 151-178.

Bejar, I., Weiss, D. & Kingsbury, G. (1977). Calibration of an item pool for the adaptive measurement of achievement. Minneapolis, Minnesota: Minnesota University. (ERIC Document Reproduction Service No. ED 146 231)

Berk, R. (1980). Item analysis. In R. A. Berk (Ed.), Criterion-referenced measurement: the state of the art. Baltimore: John Hopkins Press.

Cox, R. C. (1964). An empirical investigation of the effect of item selection techniques on achievement test construction. (Doctoral dissertation, Michigan State University, 1964). Dissertation Abstracts, 25, 6386. (University Microfilms No. 65-1725)

Cox, R. C. & Vargas, J. (1966). A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Chicago: Annual Meeting of the National Council on Measurement in Education. (ERIC Document Reproduction Service No. ED 010 517)

Crehan, K. D. (1974). Item analysis for teacher-made mastery tests. Journal of Educational Measurement, 11, 255-262.

Dayton, C. M. (1970). The design of educational experiments. New York: McGraw-Hill Book Company.

Douglass, J. B. (1981). A comparison of item response theory models for use in a classroom examination system: promising applications of latent trait models and evidence for their validity. Los Angeles: Annual Meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 199 277)

Ebel, R. L. (1968). The value of internal consistency in classroom examinations. Journal of Educational Measurement, 5, 71-73.

Hambleton, R. K. & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14, 75-96.

Hambleton, R. K. & Gorth, W. P. (1971). Criterion-referenced testing: issues and applications. Amherst, Massachusetts: Massachusetts University, Amherst School of Education. (ERIC Document Reproduction Service No. ED 060 025)

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: a review of technical issues and developments. Review of Educational Research, 48, 1-47.

Kifer, E. & Bramble, W. (1974). The calibration of a criterion- referenced test. Chicago: Annual Meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 091 434)

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

McKinley, R. & Reckase, M. (1980). A successful application of latent trait theory to tailored achievement testing. Columbia, Missouri: Missouri University. (ERIC Document Reproduction Service No. ED 190 651)

Millman, J. (1974). Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: current applications. Berkeley, California: McCutchen.

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

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

Swaminathan, H., Hambleton, R. K. & Algina, J. (1974). Reliability of criterion-referenced tests: a decision-theoretic formulation. Journal of Educational Measurement, 11, 263-267.

Van der Linden, W. J. (1981). A latent trait look at pretest-posttest validation of criterion-referenced test items. Review of Educational Research. 1981, 51, 379-402.

Wingersky, M.S., Barton, M,A. & Lord, F.M. (1982). Logist User's Guide. Princeton, N. J.: Educational Testing Service.

Wright, B. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

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

ERIC