DOCUMENT RESUME

ED 283 866 TM 870 391

AUTHOR Hambleton, Ronald K.; And Others
TITLE Optimal Item Selection with Credentialing Examinations.
PUB DATE Apr 87
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Content Validity; Cutting Scores; Difficulty Level; Higher Education; *Item Analysis; *Latent Trait Theory; Mastery Tests; Statistical Studies; *Student Certification; *Test Construction; Test Items; *Test Length
IDENTIFIERS LOGIST Computer Program

ABSTRACT

The study compared two promising item response theory (IRT) item-selection methods, optimal and content-optimal, with two non-IRT item selection methods, random and classical, for use in fixed-length certification exams. The four methods were used to construct 20-item exams from a pool of approximately 250 items taken from a 1985 certification exam in a health field. Mastery status on the criterion test was determined for candidates by administering the full item pool. For the two IRT methods, separate test versions represented cutoff scores of 65%, 70%, and 75%. Results compared: (1) overlap in items selected; (2) exam information curves; (3) accuracy of decisions resulting from use of the exams. Optimal exams typically provided three to four times more information near the cutoff scores than exams constructed by the random method. The content-optimal method performed nearly as well as the optimal method. These results support the possibility of shortening conventionally constructed credentialing exams without losing decision accuracy. For a pre-specified exam length, optimal item selection can improve decision accuracy over non-optimal methods. (LPG)

Optimal Item Selection with Credentialing Examinations

Ronald K. Hambleton, Dean Arrasmith
University of Massachusetts at Amherst

and

I. Leon Smith
Professional Examination Service

Abstract

The purposes of the study were to compare two promising item response theory (IRT) item selection methods, optimal and content-optimal, with two non-IRT item selection methods introduced to provide baseline results, random item selection and classical item selection. The effects of the four item selection methods were compared in three ways: (1) overlap in items selected, (2) exam information curves, and (3) accuracy of decisions resulting from the use of the exams.

The four item selection methods were used to construct 20-item exams from an item pool of (approximately) 250 test items. Mastery status on the criterion test was determined for candidates by administering the full item pool. Three cut-off scores were also studied: 65%, 70%, and 75%.

The results showed that the optimal exams typically provided 3 to 4 times more information near the cut-off scores than the exams constructed with the random method. Also, the content-optimal method produced nearly as good results as the optimal method. Classical method results were, in general, better than the random method but not nearly as good as the optimal methods.

The results highlighted the potential of optimal and content-optimal item selection methods for improving the decision-making capabilities of fixed-length certification exams. One consequence of these results is the potential for shortening conventionally constructed credentialing exams without losing decision accuracy. Alternatively, with a pre-specified length for a credentialing exam, the optimal item selection methods can improve decision accuracy over other non-optimal item selection methods.

5/12/87

Optimal Item Selection with Credentialing Examinations[1,2]

Ronald K. Hambleton, Dean G. Arrasmith
University of Massachusetts at Amherst

and

T. Leon Smith
Professional Examination Service

Credentialing examinations in the United States and Canada might be described in two ways: important and lengthy. The importance of these exams is clear when it is noted that over 900 professions now use the results of credentialing exams to award certificates, diplomas, or licenses. In many of these same professions, a person cannot practice until a credentialing examination (or a recredentialing examination, in many cases) has been passed.

Another common characteristic of credentialing exams is their unusual length. Exams with 200 to 500 items are regularly found in practice. The excessive lengths of many of these exams are often defended by their developers on the grounds that high levels of content validity and reliability are needed. Also, since credentialing exams are rarely pilot-tested, exam developers argue that extra items are needed so that "bad" items identified following exam administrations can be eliminated from exam scoring without fear of shortening exam lengths to the point where the psychometric properties of exam scores would be unacceptable.

There appears to be a widespread belief among those who sponsor credentialing exams (e.g., associations, agencies, etc.) that long

[1] Laboratory of Psychometric and Evaluative Research Report No. 157. Amherst, MA: University of Massachusetts, 1987.

[2] A paper presented at the annual meeting of AERA, Washington, 1987.

exams are better than short exams. But, 200 to 500 exam items with 5 to 6 hours of exam administration time seems excessive. In addition, shorter exams could be an improvement over the longer exams if the limited exam development funds were used to improve the smaller number of necessary exam items.

Hambleton and de Gruijter (1983) and de Gruijter and Hambleton (1983) demonstrated, using computer simulated exam data, the advantages of another method for improving exams that also reduces exam length: optimal item selection. For any given exam length, the most valid exam for separating candidates into "passes" and "failures" includes items that discriminate effectively near the cutoff score on the exam score scale (Lord, 1980; Lord & Novick, 1968). Such an exam is constructed using optimal item selection (Hambleton & Swaminathan, 1985). But credentialing exam development specialists have not usually taken advantage of optimal items for an exam, perhaps because they are unfamiliar with the general approach and/or with item response theory (IRT), a test theory framework that must be understood and used in optimal item selection.

Instead, classical item statistics are often used by exam developers in item selection, but these statistics have limited usefulness in constructing exams to discriminate effectively at a cut-off score of interest. The main shortcoming is that classical item statistics (item difficulty and discrimination indices) are defined over a population of candidates. The cut-off score set to separate "passes" and "failures" is defined over a domain of content.

Unfortunately, classical item statistics and the cut-off score are not defined on the same scale and therefore the item statistics cannot be used conveniently in selecting an optimal set of items for an exam. Optimal item selection requires that item statistics and the cut-off score be defined on the same scale. Item response theory can provide the needed scale when an item response model can be found to fit the exam data (Lord, 1980; Hambleton & Swaminathan, 1985).

Optimal item selection, however, is not without problems. One problem is that when statistical criteria only are used in item selection, there is the great risk of producing exams which lack content validity. Computerized adaptive testing is often criticized for the same reason. It appears that optimal item selection algorithms will need to be modified to include content considerations to avoid what seem to be a legitimate criticism. The effects of modifying optimal item selection algorithms to accommodate content considerations are unknown.

The purposes of the present paper were to compare two promising item selection methods, optimal and content-optimal (the optimal method modified to include content considerations) with two item selection methods introduced to provide some baseline results, random item selection and classical item selection. Complete details on the methods are provided in the next section. The effects of the four item selection methods were compared in three ways: (1) overlap in items selected, (2) exam information curves, and (3) decision accuracy. In addition, several cut-off scores were studied to investigate the effects of item selection methods when the cut-off scores and associated exam passing rates varied substantially.

Method

Exam Item Pool

The basic data for the study came from a certification examination in the health field administered in 1985. A three-parameter IRT model analysis was carried out on the exam data to provide item statistics and corresponding item information functions for later use in the exam development process. The item calibrations were carried out using LOGIST (Wood, Wingersky, & Lord, 1976).

Item Selection Methods

Four item selection methods were compared:

1. Random. Exam items were selected without regard for their item statistics or content. (We note, however, that all available items had been carefully reviewed by a committee and judged acceptable for use in the exams.) Random item selection, subject usually to some content constraints, is a commonly used item selection method (Hambleton, 1982; Hambleton & Rogers, 1986).
2. Optimal. Exam items were selected which provided maximum information at the cut-off score of interest. Item content was not a factor in item selection.

3. Content-Optimal. Exam items were selected which provided maximum information at the cut-off score of interest subject to the constraint that the final version of the exam must meet the content specifications approved by the exam committee.

4. Classical. Items were selected that had (1) p-values between (about) .40 and .80 and (2) the highest classical item discrimination indices (biserial correlations). In addition, the exam needed to meet the content specifications approved by the exam committee.

For the purpose of this investigation, exams consisting of 20 items were constructed. Exam length was kept short to minimize the overlap with the criterion exam which is described in the next section.

The content specifications for the criterion exam were organized by the national committee for the specialty into a two dimensional grid. The percentage of items in each cell of the two-dimensional grid were followed as closely as possible in building 20-item content valid exams with the content-optimal and the classical item selection methods.

Criterion Test

One of the criteria for evaluating the item selection methods was the pass/fail decisions resulting from the administration of the (approximately) 250-item certification exam. Of interest was the match between pass/fail decisions based on this criterion test with pass/fail decisions based on the 20-item exams constructed using the four item selection methods. Since the items selected for the 20-item exams were from the pool of items defined by the criterion test, the overlap in exam items (albeit slight) between the short exams and the criterion test inflates the levels of agreement between decisions based on the 20-item exams and the criterion test. Fortunately, this overlap did not influence the results addressing the comparison of methods because the slight positive bias in assessing agreement was common to all four item selection methods.

Cut-off Scores

Three cut-off scores for the criterion test were considered in the study: 65%, 70%, and 75%. These cut-off scores resulted in (approximate) passing rates of 90%, 80%, and 50% in the sample of (over) 1500 candidates, respectively. The corresponding cut-off scores on the exam ability scale were -1.00, -0.50, and .125, respectively, and obtained using the test characteristic curve for the total set of test items in the criterion test (Hambleton & Swaminathan, 1985). The cut-off scores on the ability scale were the points used to build optimal and content-optimal exams.

Procedure

For each cut-off score (65%, 70%, and 75%), 20-item optimal and content optimal exams were constructed. In addition, single 20-item exams using the random and classical methods were constructed. In total, eight 20-item exams were constructed from the available pool of test items - optimal and content-optimal exams at each of three cut-off scores, plus one exam constructed using the random item selection method and one exam constructed using the classical item selection method.

For each of the 20-item exams, candidate exam item scores were obtained, exam ability scores were estimated, and pass-fail decisions were made by comparing the ability estimates to the correct cut-off score (-1.00 with the 65% cut-off score, -0.50 with the 70% cut-off score, and .125 with the 75% cut-off score).

Evaluation of the Item Selection Methods

For each cut-off score, and item selection method, five evaluative criteria were of interest:

1. Percent of non-masters (as determined by the criterion test) who failed the 20-item exam (correct decisions) and who passed the 20-item exam (incorrect decisions).

2. Percent of masters (as determined by the criterion test) who passed the 20-item exam (correct decisions) and who failed the 20-item exam (incorrect decisions).

3. Overall accuracy rate (percent of candidates who were correctly classified).

These three statistics were calculated, first, for the total pool of candidates, and second, for the subsets of candidates scoring near the cut-off score. In the second analysis, only candidates scoring within one standard error of measurement of the cut-off score (about three score points) on the criterion test were included. The second set of statistics was calculated because it is among candidates scoring near the cut-off score that optimal or content-optimal item selection methods might be expected to be the most useful. Considerable interest is centered in exam development on this group because these candidates are the ones who are most likely to be misclassified.

Two other criteria were also used to interpret the results:

4. The information functions for exams constructed with the four item selection methods.

5. The probabilities of misclassification with the various exams.

Results

IRT Goodness of Fit Studies

Tables 1 and 2 provide information concerning IRT model-exam data fit. Unless the chosen IRT model fit the exam data, the research would have had little merit. In fact, the fits of the one-, two-, and three-parameter logistic models to 75 randomly-chosen exam items from the total pool of items were all quite good, though the two- and three-parameter models fit the test data somewhat better. (The c-parameter was set equal to .20 for all items with the three-parameter model.) Only a subset of items were analyzed in this phase of the research because of the limits of the LOGIST program, and our belief that a random set of items would be quite sufficient for addressing the model-data fit question.

Insert Tables 1 and 2 about here.

About 70% of the standardized residuals calculated for each test item in 12 equal-sized intervals between -3.0 and +3.0 on the ability scale using the two- and three-parameter logistic curves) had values less than one. Less than 1% of the standardized residuals exceeded a value of three. Clearly, the two- and three-parameter models provided excellent fits to the test data.

Also, the misfit statistics (the standardized resiauals) reported in Table 2 for the two- and three-parameter models were not correlated with the content categories of the exam items. The results from the one-parameter model were very different and had this model been used in our later work, an oversampling of items from a few of the content categories would have resulted. The findings in Tables 1 and 2 lent support to (1) the credibility of the unidimensionality assumption for the full set of exam items and (2) our decision to proceed in the research with the three-parameter model.

Parameter Estimation

The actual LOGIST runs were carried out with the c-parameter in the three-parameter model set to a value of .20. Table 3 provides information pertaining to the item difficulty (b-parameter) and the item discrimination (a-parameter) estimates for full set of test items. An analysis of Table 3 revealed that many items were of very limited value in the optimal or content-optimal exams of interest, either because they were very easy (high negative b-values) or non-discriminating (low a-values). Also, the limited variability among the a-parameter estimates reduced the effectiveness of the optimal and content-optimal item selection methods. In general, the optimal item selection methods will be most useful when there is considerable variability among the test items in an item pool.

Insert Table 3 about here.

Overlap in the 20-Item Exams

Table 4 shows the overlap of items in the exams constructed at each cut-off score. Use of the optimal and content optimal item selection methods resulted in considerable overlap, which was to be expected, regardless of the cut-off score. The random method, also as expected, did not overlap to any extent with the other three methods. The classical method overlapped moderately with the optimal and content-optimal at the high cut-off score (75%) and overlapped only slightly at the lower cut-off scores (65%, 70%). This finding seems to indicate that when the chosen cut-off score is far from the center of the exam score distribution (at 65% only 14% of the candidates failed), optimal (and content-optimal) exams look very different than exams constructed using classical methods. On the other hand, when the cut-off score is near the center of the exam score distribution (at 75%, 48% of the candidates failed), classical methods function more like optimal exams. In practice, however, the cut-off score for a certification exam is seldom close to the mean exam score.

Insert Table 4 about here.

Exam Information Curves

Figure 1 provides the exam information curves for the four 20-item exams at each cut-off score. An analysis of the exam information curves shows, typically, that the information is 3 to 4 times greater from the optimal and content-optimal methods than the random method. Such improvements in exam information mean that the standard errors of ability estimates for candidates around the cut-off score with the optimal exams will be about 50% smaller than the standard errors associated with exams constructed using the random method. Therefore, substantially fewer of these candidates will be misclassified. The differences in information functions for exams constructed with the optimal methods and the classical method were less, however the differences were still of practical importance, especially at the two lower cut-off scores.

Table 5 provides some results which address the probabilities of misclassification for candidates with ability scores -2.5, -2.0, -1.5, -1.0, -.5, 0, .5, and 1.0 on each of the four exams and with each of the cut-off scores, 65%, 70%, and 75%. The probabilities were obtained by assuming a normal distribution of ability estimates for each ability level of interest and a standard deviation for the normal distribution equal to the standard error of estimation associated with the exam used to obtain the ability estimates. The standard error of ability estimation equals $1/\sqrt{\text{Info}(\theta)}$ where Info(θ) is the information provided by the exam at the ability level of interest (Hambleton & Swaminathan, 1985). The statistics reported in Table 5 confirms the substantial

Insert Table 5 about here

theoretical advantages of optimal and content-optimal item selection algorithms. For example, consider θ=-2.0 and cut-off score = 65%. The probability of misclassifying the examinee using the exams constructed with the random and classical methods is at least four times larger than the probabilities of misclassification associated with the two optimal item selection methods. In fact, at nearly every ability level and for every cut-off score, the optimal and content-optimal item selection methods produced exams that substantially outperformed the exams constructed using the other two item selection methods.

Analysis of Decision Accuracy

Tables 6 and 7 provide summaries of the decision accuracy results for the total and constrained samples of candidates. Results in the

Insert Tables 6 and 7 about here.

tables highlight the actual decision accuracy results for the various exams and cut-off scores. Though the gains in decision accuracy with optimal and content optimal item selection methods with the real data were modest in size over the other two methods (they ranged from 1% to 16%), they are of practical significance. Recall first that these improved results were obtained without any increase in exam length. Any increases, however slight, as long as they do not involve major new test development expenses or an excessive amount of time, would seem worthy of serious consideration by certification boards in view of the desirability of increasing the decision accuracy (i.e. validity) of their exams. Improved decision accuracy resulted with the non-masters groups especially, in part, because on the average these groups were closer to the cut-off scores. Second, rather sizeable increases in exam length with the random and classical methods would be required to obtain even 3% to 4% increases in decision accuracy. Using the real data and the random method, the levels of decision accuracy as a function of exam length for the three cut-off scores were calculated. Even a gain in decision accuracy of 4% would require an exam constructed with the random method which would be nearly double in length! Thus, small gains in decision accuracy corresponded to rather large changes in exam length.

Conclusions

The evidence collected from this study showed that the optimal exams typically provided 3 to 4 times more information than the random exams and resulted in practically significant improvements in decision accuracy. To address the legitimate complaint that optimal exams may lack content validity, a method that balanced content with statistical considerations was also studied. The content-optimal method, also, produced very promising results. In fact, the results from this method were almost as good as the results obtained with the optimal method and in a few cases the results were better. Classical method results were, in general, better than those results obtained with the random method but not as good as the optimal methods.

The results from this study highlight the potential of both *optimal* and *content-optimal* item selection methods for improving the decision-making accuracy (i.e., validity) of fixed-length credentialing exams. If exam lengths are fixed, *optimal* and *content-optimal* methods can lead to increased decision accuracy over non-optimal item selection methods. Alternatively, if decision-accuracy results with the non-optimal item selection methods are acceptable, the use of *optimal* item selection methods can lead to substantially shorter exams without any reduction in decision-accuracy. This finding should be especially important and interesting to credentialing exam boards who may wish to shorten their exams without affecting the levels of decision-accuracy obtained from their credentialing exams constructed with non-optimal item selection methods.

In conclusion, one final point should probably be made about the results. Though the results from applying *optimal* item selection methods in this study were positive, even more positive results are likely to be observed in other applications. This is because *optimal* item selection methods will be most effective when applied to large statistically diverse item pools. In this study, the item pool consisted of relatively homogeneous test items. That is, the exam items showed very little variability in their discriminating power.

References

de Gruijter, D.N.M., & Hambleton, R.K. (1983). Using item response models in criterion-referenced test item selection. In R.K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R.K. (1982). Advances in criterion-referenced testing technology. In C. Reynolds & T. Gutkin (Eds.), Handbook of school psychology. New York: Wiley.

Hambleton, R.K., & de Gruijter, D.N.M. (1983). Application of item response models to criterion-referenced test item selection. Journal of Educational Measurement, 20, 355-367.

Hambleton, R.K., & Rogers, H.J. (1986). Technical advances in credentialing examinations. Evaluation & the Health Professions, 9, 205-229.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Hingham, MA: Kluwer-Nijhoff.

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

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

Wood, R.L., Wingersky, M.S., & Lord, F.M. (1976). LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, NJ: Educational Testing Service.

Table 1

Summary of the Absolute-Valued Standardized Residuals for 75 Items in the Certification Examination

IRT	Percent of Absolute-Valued Standardized Residuals			
MODEL	|0 to 1|	|1 to 2|	|2 to 3|	|over 3|
1-p	6i.4	30.5	7.0	1.1
2-p	70.4	26.5	2.9	0.1
3-p	70.2	26.5	2.9	0.4

Table 2

Association Between Absolute-Valued
Standardized Residuals and Item Content
on 75 Items in the Certification Examination

Content Category[1]	Number of Items	Percent of Standardized Residuals					
		1-p Model		2-p Model		3-p Model	
		SR(<.80)	SR(≥.80)	SR(<.80)	SR(≥.80)	SR(<.80)	SR(≥.80)
		(n=26)	(n=45)	(n=43)	(n=28)	(n=45)	(n=26)
X_1	40	25.0	75.0	52.5	47.5	60.0	40.0
X_2	13	46.2	53.8	76.9	23.1	69.2	30.8
X_3	8	87.5	12.5	62.5	37.5	62.5	37.5
X_4	5	60.0	40.0	60.0	40.0	60.0	40.0
X_5	5	0.0	100.0	80.0	20.0	80.0	20.0
		$x^2 = 66.26$		$x^2 = 3.35$		$x^2 = 1.08$	
		d.f. = 4, $p < .001$		d.f. = 4, $p = .50$		d.f. = 4, $p = .90$	

[1] For test security reasons, the content categories cannot be identified.

Table 3

Summary of the Item Parameter Estimates[1]
(c = .20)

)iscrimination Parameter Estimates	Difficulty Parameter Estimates					
	< -2	-2 to -1	-1 to 0	0 to 1	1 to 2	over 2
less than .30	17.7	4.8	4.0	3.6	2.0	3.2
.30 to .60	17.3	16.5	9.6	6.4	2.4	1.2
over .60	2.8	4.8	1.2	2.0	0.4	0.0

[1] Percent of test items are reported.

Table 4

Percent of Overlap in the 20-Item Exams

Cut-off Score	Item Selection Method	Item Selection Method		
		2	3	4
65%	1. Random	5%	10%	5%
	2. Optimal	-	75%	5%
	3. Content-Optimal	-	-	15%
	4. Classical	-	-	-
70%	1. Random	10%	10%	5%
	2. Optimal	-	80%	20%
	3. Content-Optimal	-	-	25%
	4. Classical	-	-	-
75%	1. Random	10%	10%	5%
	2. Optimal	-	85%	30%
	3. Content-Optimal	-	-	35%
	4. Classical	-	-	-

Table 5

Summary of Misclassification Probabilities
for Various Cut-off Scores, Exams, and Ability Levels[1]

Cut-off Score	Method	Ability -2.5	-2.0	-1.5	-1.0	-.5	0.0	0.5	1.0
65%	Random	9.7	17.3	30.9	50.0	30.0	14.9	6.8	3.3
	Optimal	0.8	3.0	15.7	50.0	17.7	5.4	2.5	1.9
	Content-Optimal	1.6	4.2	17.0	50.0	17.4	4.7	1.6	1.0
	Classical	14.2	19.0	29.8	50.0	23.9	6.3	1.0	0.2
70%	Random	4.2	7.9	15.9	30.9	50.0	30.2	16.0	8.3
	Optimal	0.5	0.8	3.5	16.9	50.0	17.6	4.5	6.0
	Content-Optimal	0.6	1.0	4.0	17.5	50.0	18.5	5.2	4.0
	Classical	7.7	9.4	14.5	26.5	50.0	22.2	6.1	1.4
75%	Random	1.2	2.2	5.2	12.2	25.6	44.8	35.5	21.0
	Optimal	0.6	0.6	0.8	2.7	12.1	40.5	24.2	7.0
	Content-Optimal	0.6	0.6	0.7	2.7	12.5	40.8	25.1	7.7
	Classical	3.0	3.1	4.3	8.0	18.8	42.4	28.1	10.1

[1] The authors would like to thank Alison Zhou for preparing these results.

Table 6
Decision Accuracy Results
(Total Sample)

Cut-off Score	Method	Non-Masters		Masters		Overall Accuracy[1]
		Fail	Pass	Fail	Pass	
65%	Random	69.7%	30.3%	14.0%	86.0%	83.8%
	Optimal	79.6%	20.4%	9.0%	91.0%	89.4%
	Content-Optimal	81.5%	18.5%	8.1%	91.9%	90.5%
	Classical	73.0%	27.0%	9.5%	90.5%	88.1%
70%	Random	72.4%	27.6%	16.7%	83.3%	80.3%
	Optimal	80.2%	19.2%	12.5%	87.5%	85.5%
	Content-Optimal	78.8%	21.2%	12.7%	87.3%	84.9%
	Classical	77.6%	22.4%	13.1%	86.9%	84.4%
75%	Random	76.3%	23.7%	30.2%	69.8%	73.0%
	Optimal	85.9%	14.1%	23.1%	76.9%	81.2%
	Content-Optimal	85.4%	14.6%	23.1%	76.9%	80.9%
	Classical	82.2%	17.8%	22.8%	77.2%	79.6%

[1] Overall Accuracy is the percent of masters who pass and non-masters who fail in the total sample of (over) 1500 examinees for the 20-item exams.

Table 7
Decision Accuracy Results
(Constrained Sample)

Cut-off			Non-Masters			Masters		Overall
Score	Method	N	Fail	Pass	N	Fail	Pass	Accuracy[1]
65%	Random	79	54.6%	45.4%	268	40.0%	60.0%	58.4%
	Optimal	79	68.5%	31.5%	268	35.8%	64.2%	65.6%
	Content-Optimal	79	70.4%	29.6%	268	33.7%	66.2%	67.5%
	Classical	79	59.3%	40.7%	268	34.2%	65.8%	63.8%
70%	Random	178	62.2%	37.8%	437	38.6%	61.4%	65.3%
	Optimal	178	69.3%	30.7%	437	30.2%	69.8%	69.0%
	Content-Optimal	178	67.1%	32.9%	437	30.5%	69.5%	67.9%
	Classical	178	61.6%	38.4%	437	31.9%	68.1%	66.7%
75%	Random	307	60.2%	39.8%	507	40.4%	59.6%	59.8%
	Optimal	307	73.7%	26.3%	507	35.7%	64.3%	68.2%
	Content-Optimal	307	73.1%	26.9%	507	24.6%	65.4%	68.6%
	Classical	307	67.0%	33.0%	507	36.3%	63.7%	65.2%

[1] Overall Accuracy is the percent of masters who pass and non-masters who fail in the constrained samples for the 20-item exams.

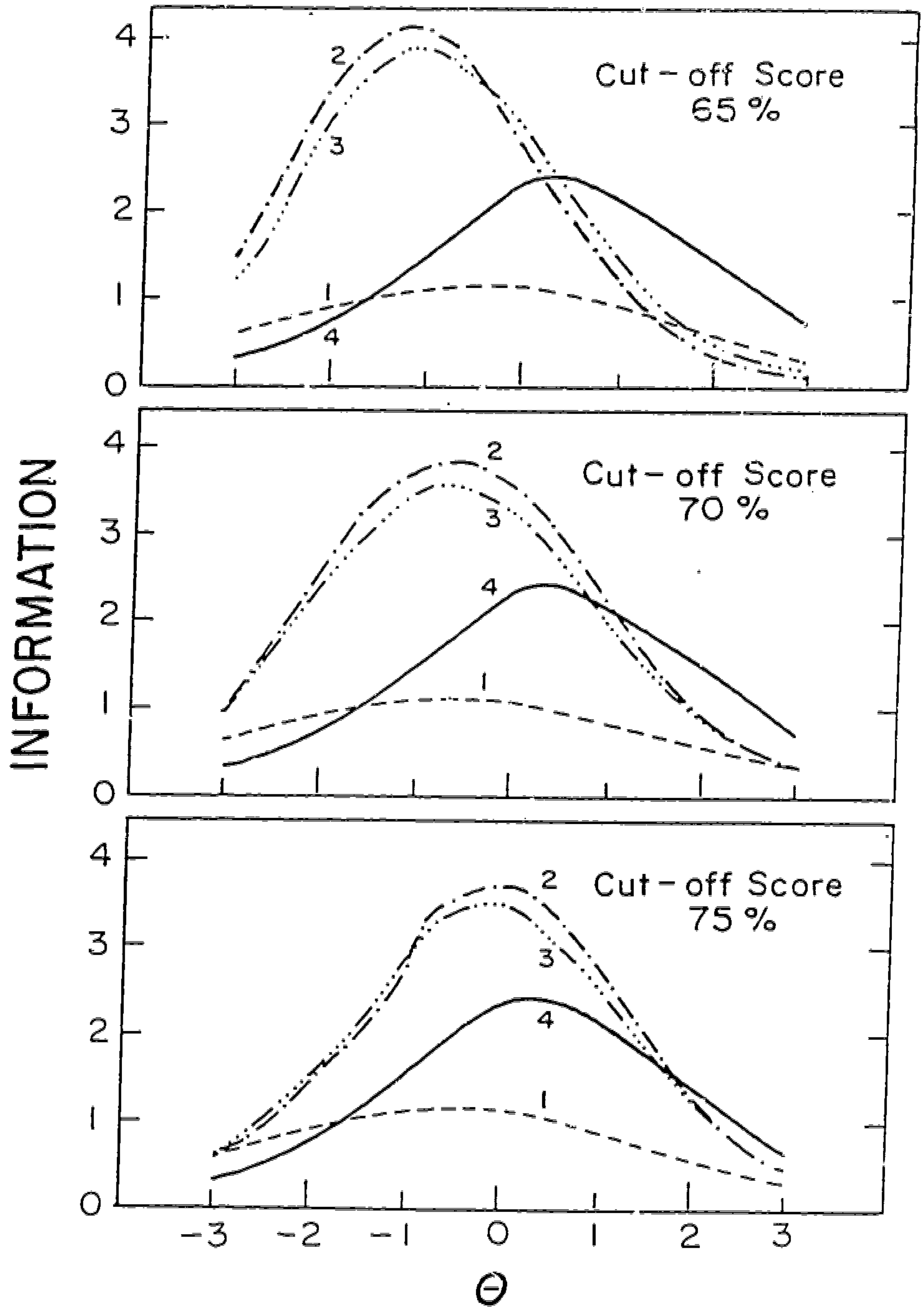

Figure 1. Test Information Functions for the 20 Item Tests

Key : 1 – Random, 2 – Optimal, 3 – Optimal – Content, 4 – Classical