DOCUMENT RESUME

ED 247 256 TM 840 427

AUTHOR Subkoviak, Michael J.; Harris, Deborah J.
TITLE A Short-Cut Statistic for Item Analysis of Mastery Tests: A Comparison of Three Procedures.
PUB DATE Apr 84
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Elementary Secondary Education; *Item Analysis; Latent Trait Theory; *Mastery Tests; Pretests Posttests; *Test Construction; Test Items
IDENTIFIERS *Agreement Statistic (Subkoviak and Harris)

ABSTRACT

This study examined three statistical methods for selecting items for mastery tests. One is the pretest-posttest method due to Cox and Vargas (1966); it is computationally simple, but has a number of serious limitations. The second is a latent trait method recommended by van der Linden (1981); it is computationally complex, but has a number of theoretical advantages. The third method, the agreement statistic proposed in this paper, parallels the latent trait method in many respects; but it is computationally simple, like the pretest-posttest procedure. A total of 81 distinct data sets were simulated; and the three item selection methods were applied to each data set for the purpose of studying relationships among the methods. The correlation between the latent trait method and the agreement method was substantial, suggesting that the latter might be recommended as a practical alternative to the former for classroom use. The results for the pretest-posttest method tended to confirm its reputed limitations. (Author/BS)

A Short-Cut Statistic for Item Analysis of Masterw Tests: A Comparison of Three Procedures

Michael J. Subkoviak and Deborah J. Harris

University of Wisconsin-Madison

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 1984.

Abstract

This study examined three statistical methods for selecting items for mastery tests. One is the pretest-posttest method due to Cox and Vargas (1966); it is computationally simple, but has a number of serious limitations. The second is a latent trait method recommended by van der Linden (1981); it is computationally complex, but has a number of theoretical advantages. The third method, proposed herein, parallels the latent trait method in many respects; but it is computationally simple, like the pretest-posttest procedure. A total of eighty-one distinct data sets were simulated; and the three item selection methods were applied to each data set for the purpose of studying relationships among the methods. The correlation between the latent trait method and the one proposed herein was substantial, suggesting that the latter might be recommended as a practical alternative to the former. The results for the pretest-posttest method tended to confirm its reputed limitations.

Item Selection for Mastery Tests: A Comparison of Three Procedures

BACKGROUND AND PURPOSE

Mastery testing has become increasingly popular within the classroom during the last decade, for at least two reasons. First is the increased use of individualized educational programs (see, for example, Davis, 1983; Hambleton & Novick, 1973; Huynh, 1976). Second, is the movement to upgrade education and to hold teachers and school systems accountable for that which they claim to teach (see, for example, Bock, 1976).

Mastery testing presumes a series of well-defined tasks to be assessed and a cutoff which distinguishes those who have successfully mastered the aforementioned tasks, and those who have not. According to Glaser (1963) criterion-referenced test scores should maximize the difference between these two groups and minimize differences within groups. For a mastery test, this means selecting items that discriminate between masters and nonmasters, as opposed to within masters and within nonmasters. The general consensus appears to be that a good mastery test item is one which masters answer correctly and nonmasters answer incorrectly (see, for example, Edwards, 1970; Lord, 1980); and the index proposed herein is based on this concept. An abundance of statistics have been proposed for selecting items for a mastery test (Berk, 1980). Unfortunately, some of these statistics have conceptual flaws, while others are too complex for routine classroom use. Thus, this study proposes a simple item selection statistic for classroom use and compares it to two other statistics that have appeared previously in the literature (Cox & Vargas, 1966; van der Linden, 1981).

The Pretest-Posttest Statistic

In 1966, Cox and Vargas proposed a pretest-posttest statistic designed to select criterion-referenced test items. The pretest-posttest statistic, designated D_{pp}, is computed by subtracting the proportion of subjects who correctly respond to an item on a pretest from the proportion of subjects who correctly respond to the item on a posttest, thus entailing two test administrations. Cox and Vargas demonstrated in their 1966 study that the pretest-posttest statistic and a traditional norm-referenced statistic produced sufficiently different results to advocate use of the former in criterion-referenced testing.

Since its introduction in 1966, the pretest-posttest statistic has become a prototype for selecting criterion-referenced test items. This is primarily due to its simplicity both conceptually and computationally. Unfortunately, the pretest-posttest statistic has some serious disadvantages, as several authors have pointed out (e.g., Berk, 1980; van der Linden, 1981). These limitations include among others: the need for two test administrations; problems related to administering the same item set twice; problems inherent to change scores; population dependency; and lack of sensitivity to the power of an item to discriminate at the cutoff score. Thus, while the pretest-posttest statistic is conceptually easy to understand and is manually calculable, the flaws embedded within it make its use as a method of item selection questionable in the context of mastery testing.

The Latent Trait Statistic

In 1981, van der Linden proposed a statistic which measures the power of an item to discriminate at a given cutoff point and which was recommended as a replacement for the pretest-posttest statistic as an item selection technique for mastery tests. This latent trait statistic is based on the concept of an item characteristic curve, which specifies the probability that an examinee with ability θ will correctly respond to an item with given difficulty, discrimination, and guessing parameters. In the usual case, the more ability a subject has, the more probable it is that he or she will correctly respond to the item. The item characteristic curve is most generally defined by:

$$P_i(+|\theta) = c_i + (1 - c_i)\{1 + \exp[-a_i(\theta - b_i)]\}^{-1} ; \qquad (1)$$

where $P_i(+|\theta)$ is the probability of responding correctly to item i given ability level θ; c_i is the item guessing parameter; a_i is the item discrimination parameter; and b_i is the item difficulty parameter. A popular simplification of Equation (1), known as the Rasch model, results by assuming $c_i = 0$ and $a_i = 1$ in (1):

$$P_i(+|\theta) = \{1 + \exp[-(\theta - b_i)]\}^{-1} . \qquad (2)$$

A mastery test requires that a criterion or cutoff score be specified to distinguish masters from nonmasters; and once this cutoff is selected, its associated value, θ_c, on the ability scale being measured can be determined (see van der Linden, 1981). Desirable items have characteristic curves, given by (1) or (2), with steep slope at θ_c, indicating that masters have a much higher relative probability of responding correctly than nonmasters. The slope is given by the derivative of the item characteristic curve at θ_c, and is designated $P_i'(+|\theta_c)$. Desirable items also have small scatter or variance of item responses at θ_c. Since item responses are dichotomous, the scatter at θ_c equals $P_i(+|\theta_c)[1 - P_i(+|\theta_c)]$. van der Linden (1981) proposed an item selection index, $I_i(\theta_c)$, which combines the slope and scatter of an item at θ_c as follows:

$$I_i(\theta_c) = \frac{P_i'(+|\theta_c)^2}{P_i(+|\theta_c)[1 - P_i(+|\theta_c)]} . \tag{3}$$

The index given by (3) is the value of the "item information function" at θ_c (Birnbaum, 1968); and, roughly speaking, it is a measure of a item's power to discriminate between masters and nonmasters. For the Rasch model given by (2), index (3) reduces to the simple form:

$$I_i(\theta_c) = P_i(+|\theta_c)[1 - P_i(+|\theta_c)] . \tag{4}$$

van der Linden (1981) performed an empirical study to explore the relationship between the pretest-posttest statistic D_{pp} and the Rasch statistic $I_i(\theta_c)$ given by (4). A physics unit was taught to 156 tenth grade subjects, and a 25 item multiple choice test was administered as a pretest and a posttest. D_{pp} and $I_i(\theta_c)$ statistics were obtained for each item, and the correlation between the two statistics was computed across items. The correlation was .23 for one cutoff point considered and was -.19 for another cutoff point. Thus, the basic conclusion was that the two statistics would tend to select very different subsets of "desirable" items from the item pool employed in the study.

In a more recent study (Harwell, 1983), items selected by the Rasch statistic (4) were generally found to produce more reliable tests than those composed of items selected by the pretest-posttest statistic, when the two selection methods were compared across multiple data sets. Thus, the results of this study provide additional evidence for preferring the latent trait statistic to the pretest-posttest statistic for mastery test item selection.

Despite its apparent theoretical and empirical superiority, the latent trait statistic is not without its own disadvantages. Its computation requires the use of a computer and appropriate software; it requires a background in latent trait theory to understand and interpret; it generally requires relatively large sample sizes (van der Linden, 1981); and the special case of the Rasch model makes some stringent assumptions that probably are not met in practice (see, for example, Hambleton & Cook, 1977). Thus, while it is to be preferred on theoretical and empirical grounds, it is impractical for classroom use in certain respects.

The Agreement Statistic

While the latent trait statistic appears to be the best available index for selecting criterion-referenced test items in terms of conceptual compatability with the purpose of mastery testing, its computational and conceptual complexity is not entirely suitable for classroom use. Conversely, the pretest-posttest statistic is manually calculable, but is inappropriate for reasons previously noted. Thus, another statistic, designated $P(X_c)$, is proposed here as an item selection index, more aligned with the latent trait statistic than is the pretest-posttest statistic, but still manually calculable.

$P(X_c)$ is computed from Table 1, with mastery/nonmastery status (determined by total test score) and item response (correct/incorrect), as marginal categories.

Insert Table 1 about here

Specifically, $P(X_c)$ is defined as:

$$P(X_c) = \frac{a_{11} + a_{22}}{N} , \qquad (5)$$

where a_{11} is the number of masters passing the item, a_{22} is the number of nonmasters failing the item, and N is the total number of examinees. Thus, $P(X_c)$ can be interpreted as the probability of agreement between outcomes on an item and outcomes on the total test (Goodman & Kruskal, 1954). Ideal items would have $P(X_c)$ values equal to one. The practical lower bound of $P(X_c)$, when no relationship exists between mastery status and item response, may be computed as:

$$P(X_c)' = \frac{(a_{11} + a_{12})(a_{11} + a_{21}) + (a_{21} + a_{22})(a_{12} + a_{22})}{N^2} .$$

Purpose of the Study

For reasons noted previously (Harwell, 1983; van der Linden, 1981), the latent trait index was viewed as the theoretically preferred statistic in the present study. The basic purpose of the study was to determine whether final test forms selected by the latent trait statistic (4) and by the agreement coefficients (5) are sufficiently similar to advocate use of the latter in classroom applications. Given the current popularity of the pretest-posttest statistic, its relationship to indices (4) and (5) was also considered in the study.

METHOD

To compare the three item selection statistics, a simulation study was designed to examine a variety of test conditions. The computer program GENIRV (Baker, 1982) was used to generate dichotomous item response data for tests of 30, 50, and 100 items "administered" as pretests and posttests to samples of 30, 60, and 120 subjects.

A two-parameter latent trait model, with $c_i = 0$ in (1), was used to generate the data. Item difficulty parameters b_i were randomly selected from a uniform distribution with values ranging from -3.00 to +3.00; item discrimination parameters a_i were randomly selected from a uniform distribution with values ranging from +.30 to +1.25. These parameter values were kept constant across both pretest and posttest within each data set; but the ability level of the examinee group increased from pretest to posttest, as discussed next.

Subjects' abilities θ were randomly sampled from normal distributions with values ranging from -3.00 to +3.00. Since the pretest-posttest statistic is reliant on the difference between the pretest and posttest ability distributions, mean differences of 1, 2, and 3 standard deviations between the two distributions were simulated. More specifically, three levels of pretest knowledge were simulated by sampling abilities from normal distributions having respective means and standard deviations of 0 and 1; -1 and .75; and -2 and .50. The standard deviation was decreased as the pretest mean decreased to simulate a "floor effect", as often occurs in real data. Posttest abilities were then simulated by drawing random samples from a normal distribution with mean 1 and standard deviation 1.

Crossing the three simulation factors (number of items, number of examinees, and pretest-posttest mean differences) resulted in 3 x 3 x 3 = 27 different conditions; and each condition was independently replicated 3 times, resulting in a total of 3 x 27 = 81 test data sets being generated. For each test data set, cutoff scores were set at 75% and 85% of the test items correct. For each data set and each cutoff, three statistics were then computed: pretest-posttest, latent trait (4), and agreement (5).

Once these statistics were computed, the items within each data set were ranked on the basic of each statistic according to the order in which they would be selected for a mastery test. Due to the apparent superiority of the latent trait statistic (Harwell, 1983; van der Linden, 1981), it was used as a basis for comparison. Spearman rank order correlations were thus computed between the ranks of the pretest-posttest and the latent trait statistic, and also between the ranks of the agreement and the latent trait statistic. In order to determine if a statistically significant difference existed between these two correlations across the various condition simulated, a split plot ANOVA was performed (see Kirk, 1982).

While running the split plot analysis would establish the existence of a significant difference in the way the pretest-posttest statistic and the agreement statistic correlate with the theoretically preferred latent trait statistic, additional analyses were required to determine if the agreement statistic is a suitable substitute for the latent trait statistic. For example, the agreement statistic might correlate more highly with the latent trait statistic than the pretest-posttest statistic, yet still not correlate highly enough to be an adequate substitute for the latent trait statistic. Therefore, two additional analyses were performed to determine if the overlap between the latent trait and agreement statistics was large enough to advocate the use of the latter, in classroom settings. First, the individual correlations were examined to determine the degree of similarity between these two statistics. The second supplemental analysis involved determining the amount of overlap between item sets selected by the two statistics, when 50% of the initial item pool was selected for a final test form.

RESULTS

The cell means and standard deviations of correlation values across 3 replications of the various test conditions are presented in Table 2. In the split plot analysis of the data, the effect for correlation-by-method was significant at the $\alpha = .05$ level, and $\hat{\eta}^2$ for this effect was equal to .87, meaning 87% of the variance in the data was explained by the two item selection methods that were correlated with the latent trait method. It can thus be concluded that a significant difference does exist in the way the agreement statistic and pretest-posttest statistic correlate with the latent trait statistic. In all cases, the mean correlation between the latent trait statistic and agreement statistic exceeded the correlation of the latent trait statistic and the pretest-posttest statistic in Table 2.

Insert Table 2 about here

With 87% of the variance accounted for by the primary effect of interest in the ANOVA, relatively little variance was left to be accounted for by the other effects and interactions. While some of these effects, such as those due to number of items and to number of examinees, were statistically significant, the associated variance accounted for by these effects was of little practical significance. The reader interested in these secondary details is referred to Harris (1983).

An examination of the individual correlations summarized in Table 2 revealed that the average correlation between the latent trait statistic and the agreement statistic was .91, suggesting that the latter may generally be a reasonable substitute for the former. In contrast, the average correlation between the latent trait statistic and the pretest-posttest statistic was -.17. An interesting finding in the data was that a majority of the correlations between the pretest-posttest statistic and the latent trait statistic were negative, meaning that the items the latent trait statistic tended to select first are items the pretest-posttest statistic tended to select last. A tentative explanation for this result follows.

Recall that the pretest-posttest statistic is computed by subtracting the proportion of examinees who respond correctly to an item on the pretest from the proportion who respond correctly on the posttest. As such, difficult posttest items tend to be selected last by this method, because the pretest-posttest difference tends to be small. Conversely, difficult posttest items tend to be selected first by the latent trait method, because the cutoffs θ_c in the present study correspond to high ability levels. Thus an item which discriminates well at θ_c (which is the basis by which the latent trait statistic selects items) would tend to be a difficult posttest item in the present study. Therefore, difficult posttest items would tend to be selected first by the latent trait statistic and last by the pretest-posttest statistic.

The final analysis involved computing the proportion of overlap between item sets selected by the latent trait and agreement statistics, when 50% of the items in the initial item pool were selected for a final test form. The results are summarized in Table 3. The average proportion of overlap across all conditions in Table 3 was .94.

Insert Table 3 about here

The results of this analysis, coupled with the preceding examination of correlation values in Table 2, suggest that the agreement statistic and the latent trait statistic perform similarly enough for the former, due to its conceptual and computational simplicity, to be tentatively recommended for further study and for possible classroom use.

DISCUSSION

The purpose of this study was to compare three methods of item selection which might be considered for mastery tests. The pretest-posttest statistic is computationally and conceptually simple, but also has serious limitations. The latent trait statistic is a desirable item selection method both in its theoretical alignment with the purpose of mastery tests (distinguishing masters from nonmasters at the cutoff score) and in yielding more reliable tests than those constructed by the pretest-posttest statistic (Harwell, 1983); however, the latent trait statistic is complex. Thus, the agreement statistic was proposed as a possible alternative for classroom use.

Items selected by the agreement statistic were found to correlate highly with items selected by the latent trait statistic; the mean correlation, across all simulation conditions, was .91; and the average proportion of overlap between selected item sets was .94. The item characteristic curve in Figure 1 provides a basis for understanding the close relationship between the latent trait statistic and the agreement statistic (Baker, personal communication). With ability on the horizontal axis and the probability of correct response given ability θ on the vertical axis, an item characteristic curve is plotted. The cutoff score θ_c divides the ability scale into masters (M) and nonmasters (NM).

The item characteristic curve is divided into four areas I-IV, which may be identified with the four cells in Table 1. In Figure 1, the area below the item characteristic curve is viewed as corresponding to those examinees correctly responding to the item. Thus, area I corresponding to those examinees who are masters (have ability levels to the right of the θ_c) and who correctly respond to the item. In Table 1, a_{11} is likewise the number of examinees who are classified as masters in terms of their total test score and who respond correctly to the item. Similarly, it can be seen that II corresponds to a_{12}, III to a_{21}, and IV to a_{22}. By viewing the item characteristic curve in Figure 1 it may also be noted that the steeper the slope at θ_c, the better the item discriminates between masters and nonmasters, and the larger the areas I and IV become. Since the latent trait statistic selects items on the magnitude of the information function at θ_c (see 3 or 4) and the agreement statistic selects items on the magnitude of $(a_{11} + a_{22})/N$, it may be seen from the item characteristic curve that the two statistics will be selecting much the same items.

Insert Figure 1 about here

REFERENCES

Baker, F. B. (1982). GENIRV: A program to generate item response vectors. Unpublished manuscript, University of Wisconsin-Madison.

Baker, F. B. (1983, August). Personal communication.

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

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Bock, R. D. (1976). Basic issues in the measurement of change. In D. M. De Gruijter & L. J. Van der Kamp (Eds.), Advances in psychological and educational measurement. New York: Wiley.

Cox, R. C., & Vargas, J. S. (1966, February). A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Davis, G. A. (1983). Educational psychology. Reading, MA: Addison-Wesley.

Edwards, A. L. (1970). The measurement of personality traits by scales and inventories. New York: Holt, Rinehart and Winston.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. Journal of the American Statistical Association, 49, 732-764.

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

Harris, D. J. (1983). Item selection for mastery tests: A comparison of three procedures. Unpublished doctoral dissertation, University of Wisconsin-Madison.

Harwell, M. R. (1983). A comparison of two item selection procedures in criterion-referenced measurement. Unpublished doctoral dissertation, University of Wisconsin-Madison.

Huynh, H. (1976). Statistical consideration of mastery scores. Psychometrika, 41, 65-78.

Kirk, R. E. (1982). Experimental design. Belmont, CA: Brooks/Cole.

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

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

Table 1

Contingency Table for Computing $P(X_c)$

	master	nonmaster
correct	a_{11}	a_{12}
incorrect	a_{21}	a_{22}

Table 2

Means and Standard Deviations of Correlation Values[a]

			75% cutoff		85% cutoff	
Items	Examinees	Mean Difference	Agreement & Latent Trait	Pre-Post & Latent Trait	Agreement & Latent Trait	Pre-Post & Latent Trait
30	30	1	.59(.52)	.05(.46)	.94(.07)	-.08(.43)
		2	.83(.12)	-.20(.38)	.94(.08)	-.30(.33)
		3	.87(.07)	-.45(.31)	.98(.01)	-.63(.26)
	60	1	.81(.09)	.30(.22)	.98(.02)	.15(.09)
		2	.77(.08)	.06(.27)	.98(.01)	-.11(.22)
		3	.90(.01)	-.64(.07)	.98(.01)	-.73(.10)
	120	1	.93(.00)	.16(.28)	.99(.00)	.03(.25)
		2	.90(.02)	-.03(.07)	.99(.01)	-.22(.08)
		3	.88(.06)	-.45(.25)	.99(.00)	-.35(.59)
50	30	1	.76(.19)	.15(.12)	.98(.01)	.15(.07)
		2	.90(.02)	-.13(.06)	.95(.05)	-.11(.16)
		3	.87(.07)	-.45(.31)	.98(.01)	-.63(.26)
	60	1	.91(.03)	.27(.08)	.97(.01)	.13(.05)
		2	.86(.07)	.08(.13)	.98(.01)	-.07(.07)
		3	.90(.01)	-.64(.07)	.98(.01)	-.73(.10)
	120	1	.91(.05)	.51(.23)	.97(.03)	.36(.28)
		2	.85(.03)	.28(.23)	.98(.01)	.03(.23)
		3	.88(.06)	-.45(.25)	.99(.00)	-.35(.59)
100	30	1	.74(.13)	.02(.07)	.99(.01)	-.32(.50)
		2	.76(.13)	-.33(.27)	.98(.02)	-.43(.23)
		3	.81(.07)	-.54(.11)	.99(.00)	-.66(.07)
	60	1	.81(.10)	.37(.13)	.92(.06)	-.02(.34)
		2	.91(.01)	-.12(.50)	.99(.00)	-.25(.47)
		3	.90(.01)	-.34(.19)	.99(.00)	-.46(.21)
	120	1	.92(.01)	.24(.09)	.99(.00)	.13(.10)
		2	.91(.01)	-.25(.37)	.99(.00)	-.27(.39)
		3	.87(.04)	-.38(.06)	.97(.05)	-.52(.05)

[a]Means and standard deviations (in parentheses) were computed over three replications of each condition.

5

Table 3

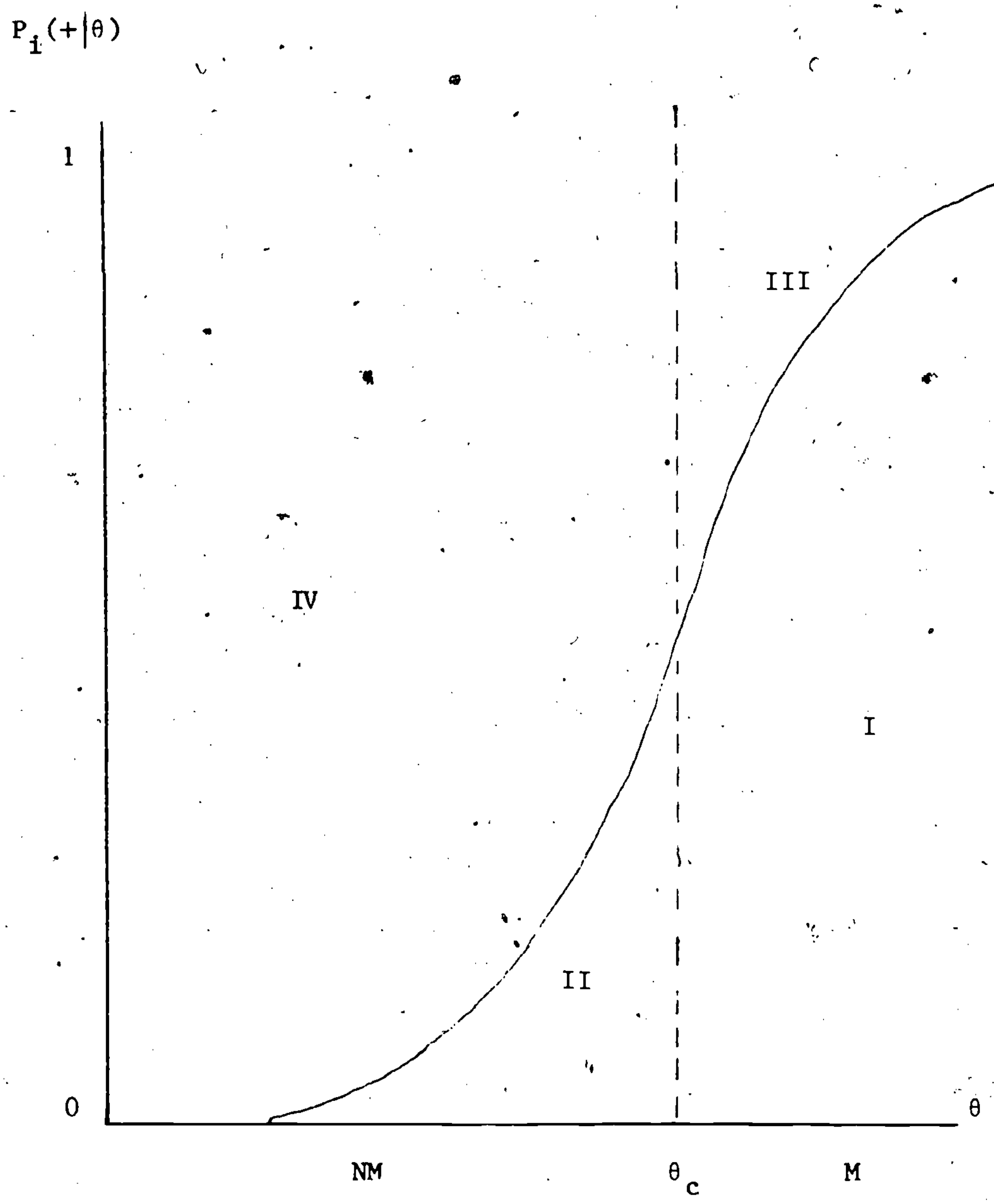
Means and Standard Deviations: Proportion of Overlap for Items Selected by Latent Trait and Agreement Statistics[a]

Items	Examinees	Mean Difference	75% cutoff	85% cutoff
30	30	1	.73(.35)	.91(.10)
		2	.89(.08)	.95(.04)
		3	.91(.10)	.95(.04)
	60	1	.95(.04)	.95(.04)
		2	1.00(.00)	.98(.04)
		3	.98(.04)	1.00(.00)
	120	1	.84(.08)	.98(.04)
		2	.95(.04)	.98(.04)
		3	.93(.00)	.98(.04)
50	30	1	.88(.00)	.96(.04)
		2	.91(.06)	.96(.04)
		3	.91(.02)	.92(.07)
	60	1	.93(.05)	.96(.00)
		2	.91(.09)	.99(.02)
		3	.95(.06)	.99(.02)
	120	1	.97(.02)	.96(.04)
		2	.93(.05)	.97(.02)
		3	.91(.10)	.97(.02)
100	30	1	.92(.06)	.99(.02)
		2	.87(.11)	.93(.03)
		3	.93(.02)	.98(.02)
	60	1	.95(.03)	.98(.02)
		2	.93(.06)	.98(.01)
		3	.96(.02)	.97(.04)
	120	1	.95(.02)	.97(.01)
		2	.91(.06)	.96(.02)
		3	.98(.00)	.97(.01)

[a] Means and standard deviations (in parentheses) were computed over three replications of each condition.

Figure 1

Item Characteristic Curve Relating Agreement Statistic and Latent Trait Statistic


$P_i(+|\theta)$
1
0
III
IV
I
II
θ
NM
θ_c
M