DOCUMENT RESUME

ED 379 302 TM 022 651

AUTHOR Crehan, Kevin D.; And Others
TITLE A Comparison of Testlet Reliability for Polytomous Scoring Methods.
PUB DATE Apr 93
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Item Response Theory; *Measurement Techniques; *Reading Tests; *Scoring; Test Construction; Test Items; *Test Reliability; Thinking Skills
IDENTIFIERS Max Alpha Scoring; Number Right Scoring; *Polytomous Scoring; *Testlets

ABSTRACT

Among the measurement techniques receiving greater attention is the context-dependent item set or testlet. The context-dependent item set consists of a scenario and related test questions. This item format is generally believed to be able to tap higher level thinking. Unfortunately, this item form leads to inter-item dependence within item sets and inflated reliability estimates when items are treated as unrelated. In this study alternative ways to score item sets (number right and max-alpha scoring) are examined with respect to classical reliability and item response theory (IRT) information using both dichotomous and polytomous scoring models. Responses of 2,817 examinees to 17 items in a reading test served as the data set. The results are consistent with previous research showing inflated reliability estimates when context-dependent item sets are treated as stand-alone items. The evidence suggests that the testlet structure of the measure must be taken into account in determining test statistics and examinee scores. (Contains 32 references, 1 table, and 3 figures.) (Author/SLD)

A Comparison of Testlet Reliability for Polytomous Scoring Methods[1]

Kevin D. Crehan

University of Nevada, Las Vegas

Stephen G. Sireci

American Council on Education

Thomas M. Haladyna

Arizona State University West,

Pamela A. Henderson

University of Nevada, Las Vegas

[1]*The authors gratefully acknowledge the help provided by David Thissen in this effort. However, he is not responsible for errors of omission or commission present in this manuscript.*

Paper presented at the annual meeting of the *American Educational Research Association*, Session 25.42, Atlanta, GA, April, 1993.

Abstract

Among the measurement techniques receiving greater attention is the context-dependent item set or testlet. The context-dependent item set consists of a scenario and related test questions. This item format is generally believed to be able to tap higher level thinking. Unfortunately, this item form leads to inter-item dependence within item sets and inflated reliability estimates when items are treated as unrelated. In this study alternative ways to score item sets are examined with respect to classical reliability and IRT information using both dichotomous and polytomous scoring models. The results are consistent with previous research showing inflated reliability estimates when context-dependent item sets are treated as stand-alone items. The evidence suggests that the testlet structure of the measure must be taken into account in determining test statistics and examinee scores.

Introduction

The multiple-choice item format is frequently criticized for over emphasizing measurement of trivial recall level learning outcomes (Hoffman, 1964; Morgenstern and Renner, 1984; Stiggins, Griswold & Wikelund, 1989). Recent reform in schooling, coupled with advances in cognitive psychology, have promoted the measurement of higher levels of thinking (Nickerson, 1989; Toch, 1991). Several objectively scorable formats have been proposed and studied that purport to accomplish this end (Haladyna, 1992a; 1992b). Ebel (1951) suggested the context-dependent item set as a promising objective test format capable of measuring higher level learning outcomes. Wainer and Kiely (1987) have used the more concise term, testlet, to refer to the context-dependent item type. This format has a structure that consists of a scenario and a set of related test items. These items can be in any format, but most common examples are either conventional multiple-choice or multiple true-false (Frisbie, 1992). A review of current research on the context dependent item set reveals that the item set can be effectively applied to measuring various types of higher level thinking (Haladyna, 1992b).

Although construction of context-dependent item sets is more difficult and time consuming than other types of objective test items, this item format has an advantage over the "stand-alone" objective item formats in the ease with which higher-level thinking can be measured. However, a major problem exists with context-dependent item sets. Since all items within a set require appropriate analysis and interpretation of the introductory material for success in selecting the keyed response, a single misinterpretation can result in more than one incorrect response. This unfortunate feature allows the possibility of correlated errors of measurement within an item set. The potential for correlated errors has generally been ignored and context-dependent item sets have been scored as stand-alone items rather than separate item sets. This scoring procedure results in inflated reliability estimates due to irrelevant context-dependent item covariance that is not related to the underlying trait being measured (Thissen, Steinberg, & Mooney, 1989; Sireci, Thissen, and Wainer 1991). This concern has long been recognized. Kelley (1924) commented that the Spearman-Brown reliability coefficient would be too large when two or more exercises contain common features. Guilford (1936), Thorndike (1951), and Anastasi (1961) cautioned about inflated reliability estimates when items dealing with a single problem are scored as separate items. Cureton (1965) also argued that errors of measurement in classical test theory are correlated for items within context-dependent item sets.

Three strategies for dealing with the violation of local independence of context-dependent item sets were suggested by Wainer and Lewis (1990). One strategy was to use only a single item for each passage; however, this strategy is obviously inefficient. A second approach was to ignore the

interdependencies among the related items, fit a dichotomous IRT model and hope for the best. Thirdly, and preferably, is the context-dependent item set approach, which treats the passage and its related items as a single unit.

The second and third approaches were examined by Thissen, Steinberg, and Mooney (1989). In a comparison of dichotomous and polytomous IRT models, they found that scoring at the item level with a three-parameter logistic model resulted in overestimation of the precision of measurement as characterized by the test information function. The item level reliability was .08 or 12-13% higher than the context-dependent item set reliability. They proposed scoring with context-dependent item set response models where the focus is still at the level of the individual item but within a context-dependent item set format. Sireci, Thissen, and Wainer (1991) found that failing to consider the dependencies within four context-dependent item sets resulted in a 10-15% overestimation of reliability. Yen (1992) found less test information for scores which treated item sets as a unit rather than as separate items. She suggests scoring context-dependent item sets as separate item sets, thereby creating subscores as opposed to traditional stand-alone or individual item level scoring.

Since context-dependent item sets are in wide use in standardized achievement, competency, certification, and licensure testing, the question of appropriate scoring and reliability estimation is important. There has been little research comparing various scoring strategies to investigate differences between classical and IRT scoring models and between dichotomous and polytomous scoring models as they relate to the context-dependent item set. The problem of diminished reliability due to local dependence within item sets has not been investigated with polytomous scoring models. The research reported here compares several scoring methods on the reliability of context dependent item sets treated as dichotomous stand-alone items, polytomous stand-alone items, and as single score testlets.

Polytomous scoring models are based on the assumption of a systematic relationship between distractor performance and total test score (Levine & Drasgow, 1983; Lord, 1977). This relationship has been the basis for attempts to use the differential information represented in wrong answer choice to score test results. The term polytomous scoring has been used to describe strategies which use the information present in distractors. Polytomous scoring techniques are an alternate to dichotomous scoring methods under which item responses are scored as either right or wrong and total score is a function of the number of right answers.

Scoring Methods to be Investigated

Number Right

Number right is the sum of correct responses weighted one and incorrect responses weighted zero.

Max-alpha

Guttman (1941) proposed a polytomous scoring strategy for maximum performance measures resulting in an optimization of coefficient alpha, hence the method is often referred to as max-alpha. Max-alpha consistently yields higher internal consistency than dichotomous scoring(Haladyna & Sympson, 1988).

Max-alpha uses the concept of option mean, the mean of total test score for all examinees choosing an option. Each option's mean is used as an initial weight to score test results, the new total score is used to recompute option means, and the process is iterated to a criterion of stabilization in the change of coefficient alpha. This option weighting strategy results in both differential option and item weighting with more difficult items having higher weights assigned to their keyed response. Echternacht (1975) found that the initial option mean is very close to maximizing alpha and few iterations are needed. Therefore, using the initial option mean without iteration is a relatively simple way to obtain a good approximation to an iterated set of option weights.

Bock's Nominal Model

Bock's (1972) nominal model has been applied effectively to testlet-based tests by Thissen, Steinberg, and Mooney (1989), Wainer and Lewis (1990), Sireci, Thissen, and Wainer (1991), Wainer, Sireci, and Thissen (1991), and Yen (1992). Bock's nominal model is useful for analysis of testlet-based tests because it requires that the assumption of conditional independence hold only between testlets, rather than between the items that comprise them. To use Bock's nominal model, the items within a testlet are treated as a single, polytomous item. A "testlet score" is computed for examinees by summing the total number of items within the testlet that were answered correctly. Thus, the testlet scores (or the "responses" to this polytomous "item") range from 0 to m, where m equals the number of items within the testlet[1].

In this study we followed the use of Bock's (1972) model by Thissen Steinberg, and Mooney (1989), and others. Assuming there are J testlets, indexed by j, where j=1,2, ... J, there are mj questions. The polytomous response for each testlet would range from 0 to mj. The trace lines (or item characteristic curves) for score $x = 0,1, \ldots mj$ for testlet j is

$$T_{jx} = \frac{\exp[a_{jx}\theta + c_{jx}]}{\sum_{k=0}^{m_j} \exp[a_{jk}\theta + c_{jk}]} \qquad (1)$$

where θ is the latent variable being measured, and $\{ak,, ck\}$ $k = 0,1, \ldots mj$ are the item category parameters (Thissen, et al, 1989). The additional constraints

$$\sum_k a_{jk} = \sum_k c_{jk} = 0$$

are imposed to identify the model, and the model is reparameterized using centered polynomials of the associated scores to represent the category-to-category change in the ak s and cks:

$$a_{jk} = \sum_{p=1}^{P} \alpha_{jp} (k - \frac{m_j}{2})^p \qquad (2)$$

[1]Wainer and Kiley (1987) proposed alternatives to using the number correct score for scoring testlets. These alternatives use the actual responses to each item within a testlet, rather than scoring the items dichotomously and adding up the number correct. However, due to the large number of response patterns that would result from polytomous scoring of testlets, these are impracticable for most test data (and available IRT software) and have not yet been investigated.

and

$$c_{jk} = \sum_{p=1}^{P} \gamma_{jp} (k - \frac{m_j}{2})^p \qquad (3)$$

where the parameters $\{\alpha_p, \gamma_p\}j$, $p = 1,2, ..., P$, for $P \leq mj$ are the free parameters to be estimated from the data.

Multiple-choice model

The multiple-choice model (MC) was developed by Thissen and Steinberg (1984). Thissen and Steinberg (1986), and Thissen, Steinberg, and Fitzpatrick (1989) demonstrate the relationship of the MC model to Bock's (1972) nominal model and Samejima's (1969, 1979) graded model. The MC model differs from the nominal model in that an additional, latent response category is added to each item to represent the responses of examinees who "don't know" the correct answer and "guess." Thissen and Steinberg (1984, 1986) and Thissen, Steinberg, and Fitzpatrick (1989) provide the equations for the MC model and describe its empirical constraints. The MC model has been used to allow for polytomous scoring of multiple-choice items so that information in the distractors (incorrect response options) can be used to estimate examinee proficiency. Thus, the MC model allows for polytomous scoring of multiple-choice items using IRT.

3PL

The one-, two-, and three-parameter logistic IRT models have been thoroughly described and investigated (e.g., Hambleton, 1989). To investigate IRT scoring, the three-parameter logistic IRT model (3PL) was used. The mathematical form of the 3PL item characteristic curve is

$$p(\Theta) = \frac{c + (1 - c)}{1 + \exp[-a(\Theta - b)]}$$

where $p(\Theta)$ is the probability of choosing the correct answer as a function of Θ; b is the difficulty level of the item, a is the slope of the item characteristic curve (ICC) at the point $\Theta = b$, and c is the lower asymptote of the ICC. The item parameters a, b, and c are commonly referred to as the discrimination, difficulty, and lower-asymptote (or guessing) parameters, respectively. The 3PL IRT model is appropriate only for dichotomously-scored multiple-choice items.

Method

Instrument

The test data analyzed in this study were part of a 40 multiple-choice item test of reading proficiency used in a credentialing examination. Each of the 40 items were linked to one of seven reading passages. There were five response options for each item. To reduce problems associated with estimating large numbers of item parameters, the two longest reading passages were selected for analysis. The first passage (testlet) contained nine items, the second contained eight items. The responses of 2817 examinees to these 17 items served as the data set for all analyses.

IRT-based Analyses

Three IRT models were fit to the data for the 17 items. First, a 3PL model was fit to the scored (dichotomous) item data. Priors of .20 were used for the lower asymptote parameters. Second, the MC model was fit to the "raw" (polytomous) item responses. An unconstrained MC model was used initially and constraints were introduced as recommended by Thissen, Steinberg, and Fitzpatrick (1989). The final MC model for the results reported here, fixed the d_k parameters (parameters indicating the proportion of examinees who "don't know," and choose option k) at .20, and constrained several a_k parameters (location parameters for each option) to be equal to or greater than the location parameter for "option" d_1. The third IRT model fit to the data was the nominal model. For this analysis, the items within each testlet were scored dichotomously, and the number correct score for all the items comprising the testlet was used as the testlet score. Thus, the first testlet had 10 response categories ranging from 0 to 9, and the second testlet had 9 response categories ranging from 0 to 8. A fully unconstrained nominal model was initially fit to the data. Subsequently, the lowest-ordered polynomials for the a_ks and c_ks were found that did not illustrate a statistically significant change in fit to the data from the fully unconstrained model. This final reduced-polynomial model imposed second-order constraints for the a_k parameters of testlet 1, third-order constraints for the a_k parameters of testlet 2, and fourth-order constraints for the c_k parameters of both testlets.

All IRT models were fit to the data using MULTILOG, version 6.0 (Thissen, 1991). MULTILOG is a very general IRT program that allows for the fitting of dichotomous and polytomous IRT models and easily incorporates parameter constraints such as those required in this study.

Max-alpha

Polytomous option mean based scores were determined treating the data as originating from 17 stand-alone items and as based on two testlets.

Number Right

Number right scores based on all seventeen items and two total scores for the separate testlets were determined.

Results

The trace lines for testlets 1 and 2, resulting from the nominal model, are presented in Figures 1 and 2, respectively. Responses 0 through 9 are plotted as curves 1 through 10 in Figure 1, responses 0 through 8 are plotted as curves 1 through 9 in Figure 2. An inspection of these figures illustrates that getting two of the nine items correct in testlet 1 indicates about the same level of proficiency as getting no items correct (response curves for 1, 2, and 3 are virtually parallel), and that getting one item correct is about the same as getting no items correct for testlet 2. Given that five response options were available for each item, this finding is not surprising. The response curves for the other options are ordered as expected - the response curves associated with a greater number of correct answers are associated with higher levels of theta than are response curves associated with fewer items answered correctly. These results are consistent with Thissen, Steinberg, and Mooney (1989), and Wainer, Sireci, and Thissen (1991).

The (marginal) reliability estimates associated with the three IRT models indicated that, when the local item dependence was accounted for by the model, the reliability was reduced (see Table 1). The reliabilities for the MC model and 3PL model were similar, but were substantially higher than the reliability estimate generated by the nominal model. Because the nominal model represented the appropriate model (i.e., did not violate assumptions of local independence) this lower reliability is more accurate. These results are consistent with those of Sireci, et al, (1991) and Yen (1992).

Though the 3PL and MC models were similar in terms of reliability, some differences in the test information functions (TIFs) were noted. The TIFs for the three IRT models are presented in Figure 3. The difference between the TIF for the 3PL model and the nominal model are as expected. The dramatically different shape of the MC model illustrates the difference between the two types of data used: the 3PL and nominal models used dichotomously-scored data, while the MC model used the polytomous data. Thus, though the MC model overestimated the

"true" reliability, it seems to provide more information (precision of measurement) over the lower end of the proficiency scale. It is unfortunate that a polytomous (pattern scoring) testlet model was not applied to these data to evaluate the potential gain in information from using the "raw" item responses.

In summary, treating the data as two item sets rather than as 17 stand-alone items resulted in a substantial reduction in reliability. For number right scoring the alpha reliabilities were .69 and .77, a difference of .08 or an 11.6% over estimation in reliability. Using the Spearman-Brown prophesy formula this translates to a 33% reduction in test length. For the 3pl IRT model the marginal reliabilities were .71 and .78, a difference of .07 or a 10% reduction in reliability and a 31% reduction in test length. Polytomous option-mean scoring did no better. The alpha reliabilites were .67 and .79, a difference of .12 or an 18% decrease in reliability translating to a 46% decrease in test length.

Conclusions

The results are generally consistent with Thissen, Steinberg, and Mooney (1989), Sireci, Wainer, and Thissen (1991) and Yen (1992). The most notable addition to previous results was the inclusion of a polytomous scoring method for stand alone items. The inflated reliability of the MC model is not surprising. However, polytomous scoring of testlets, using raw item responses, is likely to increase information at the lower end of the ability scale. This is important since many certification decisions are made at this level.

It is unfortunate that the nominal model used to investigate testlet reliability did not use the raw item response data. Though the nominal model used here was polytomous, some important information may be lost in the dichotomous scoring of items that comprise the testlet. The raw item responses could not be used for the testlet analyses because response-pattern testlets (based on right-wrong only scoring) would have $2^9=512$ response categories for the nine-item testlet, and $2^8=256$ response categories for the eight-item testlet. Using all of the five response options would result in $5^9=1,953,125$ and $5^8=390,625$ response categories, respectively, for these two testlets! Though Thissen and Steinberg (1988) used response-pattern testlet scoring for two dichotomously-scored items, the problem of applying these analyses to more than a few polytomous items is obvious.

One potential way to use the raw responses in IRT-based testlet analyses is to weight each response option a priori, and compute the testlet score based on a sum of the weights assigned to each option. In this manner, the actual response option chosen for each item would impact upon the testlet (and theta) score. For example, a MC model could be applied to the raw item response data, and the location parameters for each response option could

be used to compute the testlet scores (i.e, testlet score would be the sum of the location parameters for the options chosen, or some transformation thereof). Using the option mean or percentile procedure to weight the options may also prove useful. Such analyses would not be impracticable and would provide the data necessary for the subsequent nominal model analyses.

References

Anastasi, A. (1961). *Psychological testing (2nd ed.).* New York: Macmillan.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Cureton, E. E. (1965). Reliability and validity: Basic assumptions and experimental designs. *Educational and Psychological Measurement, 25,* 327-346.

Ebel, R. L. (1951). Writing the test item. In E. F. Lindquist (Ed.), *Educational Measurement, (1st ed., pp. 185-249).* Washington, DC: American Council on Education (ACE).

Echternacht, G. (1975). The variances of empirically derived option scoring weights. *Educational and Psychological Measurement, 36,* 301-309.

Frisbie, D. A. (1992). The status of multiple true-false testing. *Educational Measurement: Issues and Practices, 5,* 21-26.

Guilford, J. P. (1936). *Psychometric methods (1st ed.).* New York: McGraw-Hill.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.) *Prediction of Personal Adjustment. Social Science Research Bulletin, 48,* 321-345.

Haladyna, T. M. (1992a). Context dependent item sets. *Educational Measurement: Issues and Practices, 11,* 21-25.

Haladyna, T. M. (1992b). The effectiveness of several multiple-choice formats. *Applied Measurement in Education, 5,* 73-88.

Haladyna, T. M., & Sympson, J. B. (1988). Empirically based polychotomous scoring of multiple-choice test items: A review. Paper presented in C. E. Davis (Chair), New Developments in Polychotomous Item Scoring and Modeling. Symposium conducted at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hoffman, B. (1964). *The tyranny of testing.* New York: Collier.

Kelley, T. L. (1924). Note on the reliability of a test: A reply to Dr. Crumm's criticism. *The Journal of Educational Psychology, 15,* 193-204.

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43,* 675-685.

Lord, F. M. (1977) Optimal number of choices per item: A comparison of four approaches. *Journal of Educational Measurement, 14,* 33-38.

Morgenstern, C. F. & Renner, J. W. (1984). Measuring thinking with standardized science tests. *Journal of Research in Science Teaching, 21,* 639-648.

Nickerson, R. S. (1989). New directions in educational assessment. *Educational Researcher, 18,* 3-7.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 4,* Part 2, Whole #17.

Samejima, F. (1979). A new family of models for the multiple-choice item. Office of Naval Research Report 79-4. Knoxville, TN: University of Tennessee.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237-247.

Stiggins, R. J., Griswold, M. M., & Wikelund, K. R. (1989). Measuring thinking skills through classroom assessment. *Journal of Educational Measurement, 26,* 233-246.

Thissen, D. (1991). *Multilog.* Chicago, Il.: Scientific Software, Inc.

Thissen, D. & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika, 49,* 501-519.

Thissen, D. & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104,* 385-395.

Thissen, D., Steinberg, L. & Fitzpatrick, A. R. (1989). Multiple-choice Models: The distractors are also part of the item. *Journal of Educational Measurement, 26,* 161-176.

Thissen, D., Steinberg, L. & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26,* 247-260

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement.* Washington, DC: American Council on Education.

Toch, T. (1991). *In the Name of Excellence: The Struggle to Reform the Nations Schools, Why it's Failing and What Should be*

Done. N.Y.: Oxford Press.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-202.

Wainer, H. & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1-14.

Wainer, H. Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197-219.

Yen, W. M. (1992). Scaling performance assessments: Strategies for managing local item dependence. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Table 1

Reliabilities for each scoring method for the items treated as 17 stand-alones and as 2 testlets.

Unit of Analysis	Model	Reliability	Type
17 items-dichotomous	NR	.77	alpha
17 items-raw respns	OptMn	.79	alpha
17 items-dichotomous	3pl	.78	marginal
17 items-raw respns	MC	.77	marginal
2 testlets	NR	.69	alpha
2 testlets	OptMn	.67	alpha
2 testlets	Nominal	.71	marginal

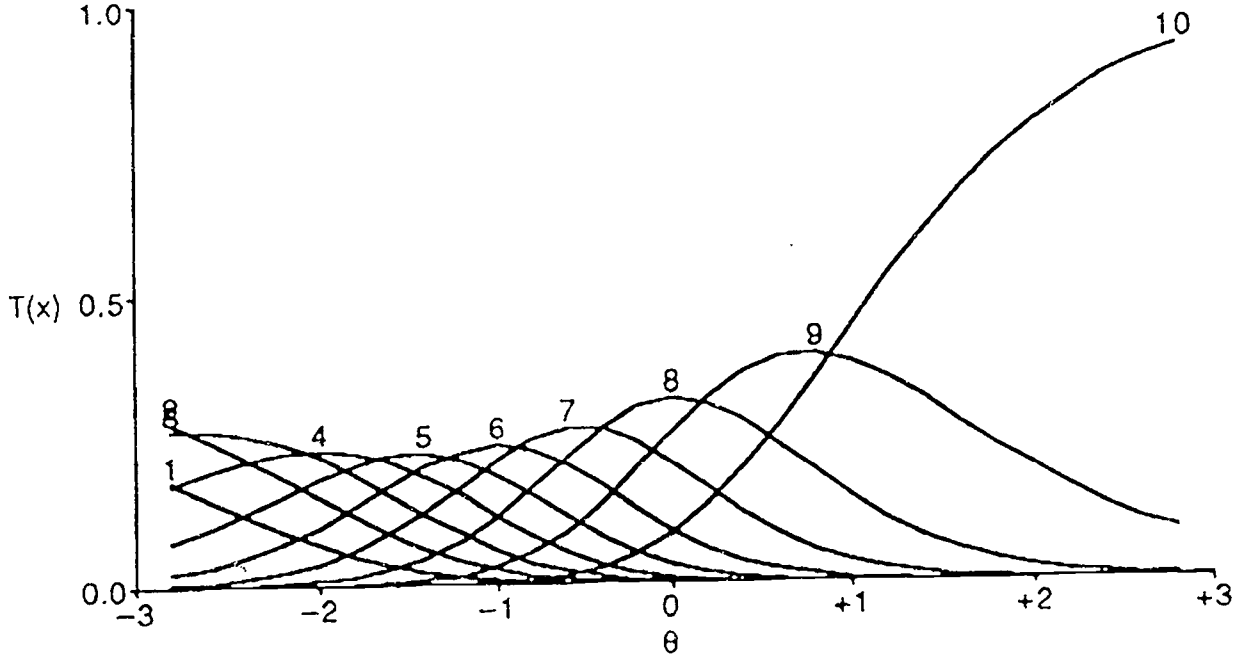

FIGURE 1

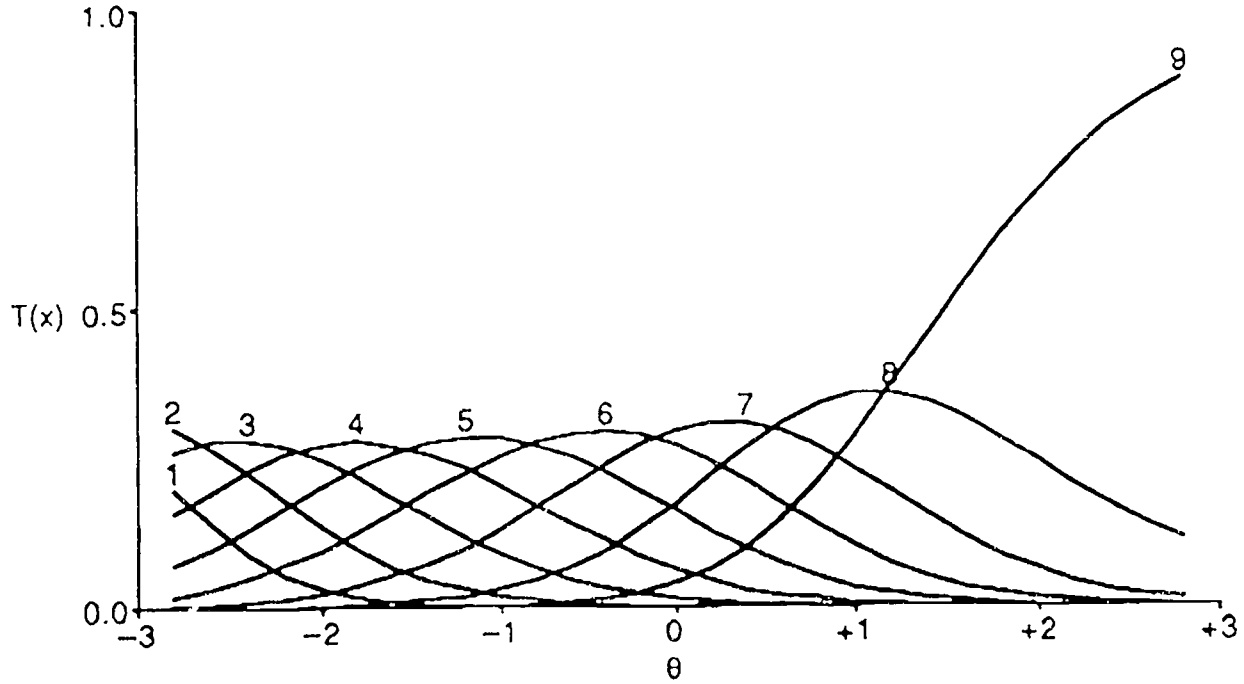

FIGURE 2

Test Information Functions

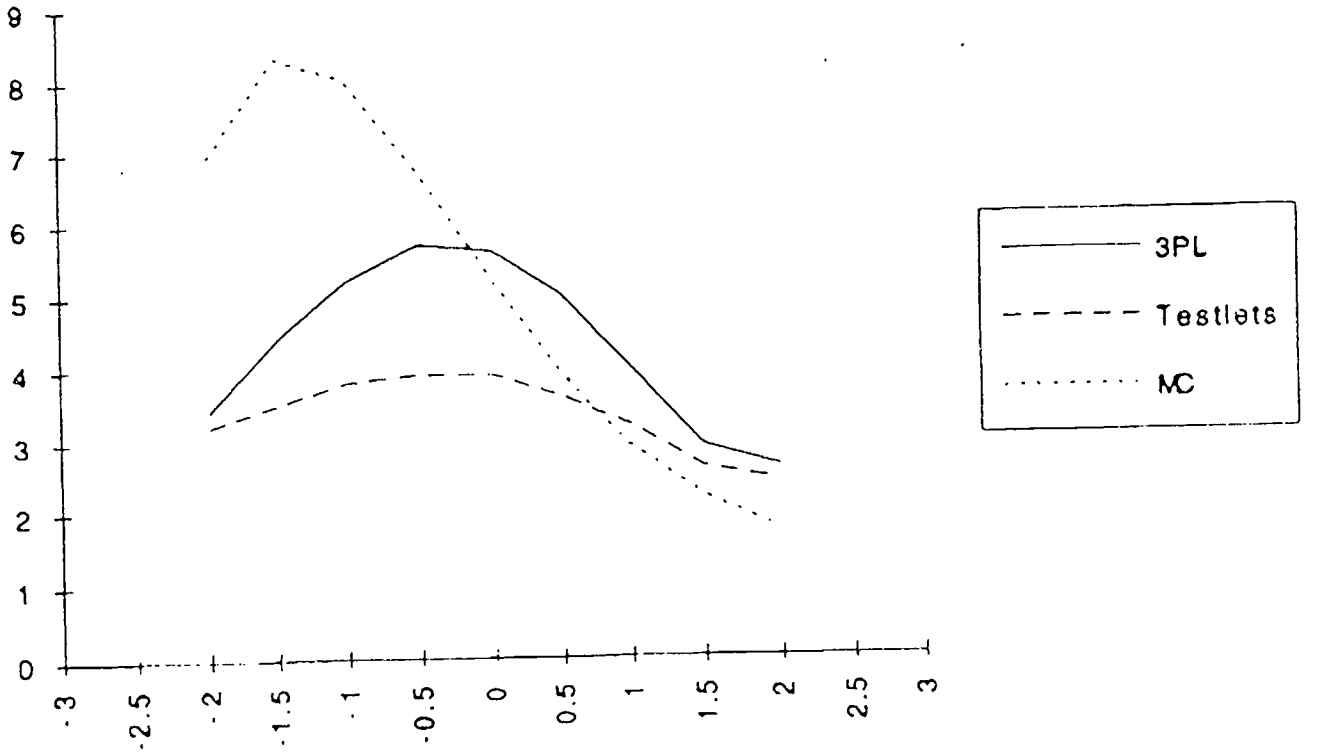

FIGURE 3