Solano-Flores, G. (1993). Item structural properties as predictors of item difficulty and item association. *Educational and Psychological Measurement, 53*(1), 19-31.

ITEM STRUCTURAL PROPERTIES AS PREDICTORS OF ITEM DIFFICULTY AND ITEM ASSOCIATION

Guillermo Solano-Flores

Logical Test Design (Lit) is a technique for developing test items that evaluate procedural learning. This research investigated the ability of LTD to predict student performance in reading Roman numerals. The study (a) compared item structural properties identified by LTD with item size and expert judgment, in their ability to predict item difficulty, and (b) investigated the consistency of students' performance on item pairs varying on their LTD structural similarity. Two hundred and eleven sixth-grade students were tested on Roman numeral items varying on LTD-related and LTD-non-related variables. An LTD-related variable, *item iterativity,* the number of iterations through the algorithm for decoding a Roman numeral, was found to be the best predictor of item difficulty. Furthermore, *item association,* the proportion of students that consistently performed correctly or incorrectly on any two items, was greater for those items identified by LTD as structurally equivalent. However, that expert judgment, almost independently of iterativity, predicted difficulty about as well, suggested a cognitive component not captured by the model. Strengths and limitations of LTD for test development are discussed.

Even though test item construction lacks objectivity (e.g., Bormouth, 1970), and has been described as more art than science (Millman, 1980; Roid and Haladyna, 1982), research has focused more on test theory (Lord and Novick, 1968; Nunally, 1964, 1967; Payne, 1968; Thorndike and Hagen, 1969) than on test design (Osterlind, 1989). Even construction guidelines (e.g., Conoley and O'Neil, 1979; Ebel, 1972; Gronlund, 1971, 1988) deal more with form than with test item content. This lack of objectivity limits test construction in three ways. First, it is often impossible to offer a non-statistical, logical reason to justify the use of a given number of items in a test, although the number of items is a critical factor for the appropriate evaluation of a skill (Priestley, 1982). Second, item difficulty (DIF) is commonly estimated empirically (Wood, 1977), not according to the intrinsic properties of items. Item writers cannot determine and manipulate DIP without having the results of an item analysis. Third, test content is influenced by personal idiosyncratic factors, even when item writers are provided complete information about program objectives and content (Roid and Haladyna, 1978).

Concerns about the limitations in determining item content (Guttman, 1969) have given rise to the development of alternative ways for constructing items (e.g., Hively, 1974; Roid and Haladyna, 1982). For example, when structural properties of items are taken into consideration, performance on cognitive tasks can be more accurately predicted than when these properties are ignored (Anderson, 1985; Ashcraft, 1989). By taking into account knowledge structure, then, item writers may be able to identify, control, and manipulate variables in item construction (Embretson, 1985a).

Logical Test Design (LTD) is one technique that accounts for structural properties to develop items that test procedural learning (Solano-Flores, 1989). For LTD, a procedure is a set of sequences of physical or mental actions that must be performed to solve a class of problems or to acomplish a class of goals (Trakhtenbrot, 1963); different procedures address different problem classes. LTD assumes that a procedure user makes binary decisions to identify of what a particular problem consists (i.e., identifies problem subclasses) (Sternberg, 1969). This way of defining items for a particular problem class leads to an "item form" that defines a universe of

possible items (e.g., lively, Patterson, and Page, 1968), and is simpler to apply than graph analysis (e.g., Scandura, 1973).

Although LTD has proven useful in generating items in a variety of subject matters (arithmetic, chemistry, electronics, and commerce (Solano-Flores, 1991), it lacks empirical support for its ability to predict observed performance. This research investigated the ability of LTD to predict student performance. More specifically, the purpose of this study was to compare the ability of LTD-related and LTD-not related variables in predicting DIF. Prediction from LTD was expected to account for a larger proportion of DIF variance. The study also examined the consistency of student performance across item pairs varying in their structural similarity. Student performance was expected to be more consistent for those item pairs having the same structural properties than for those items having different structural properties.

Method

Subjects

Subjects were 211 male and female, sixth grade, Ill-to 13-year old students from four coeducational private schools in a middle-class neighborhood in Mexico City.

Instrument

A test for reading 69 Roman numerals was constructed to examine LTD empirically. This topic was chosen because it is part of the official mathematics curriculum in elementary schools, and because training of subjects was not required.

Test items were generated according to LTD phases. In phase 1, a procedure for reading Roman numerals was represented in a flow chart (Figure 1). Particularly relevant are references by Farina (1970); Jonassen, Hannum, and Tessmer (1989); Martin and McClure (1985); Schriber (1969); Stern (1975); and Wheatley and Unwin (1972). Two types of elements are used: *operations* (square boxes), which prescribe actions (e.g., "Add the partial results") that change or transform the problem and contribute to its solution; and *decisions* (diamond-shaped boxes), which are binary questions (e.g., "Does the letter appear two or three times?") with one of two possible answers, YES or NO.

In phase 2, subclasses of Roman numerals were identified, based on two types of decisions distinguished by LTD: (a) *identification decisions* (A, B, and C, Figure 1) which identify the characteristics of a particular problem that require operations to be performed, and (h) *control decisions* (a, Figure 1), which signify that a solution has been attained. In a control decision, one option ends the application of the procedure; the other option leads back to a previous part of the procedure and thus prescribes another "turn" of actions in the algorithm. "XV", for example, is solved in two turns. The first turn solves the "X ..." portion of the problem, by means os this sequence of actions (Figure 1):

1, 2,3, A(YES), B(NO), C(NO), 6, α(YES)…

Option α(YES) prescribes another "turn", to solve the "…V" portion:

…7, A(NO), 6, α(NO), 8, 9.

Thus, the procedure ends when NO is the response to a, "More letters to analyze?"

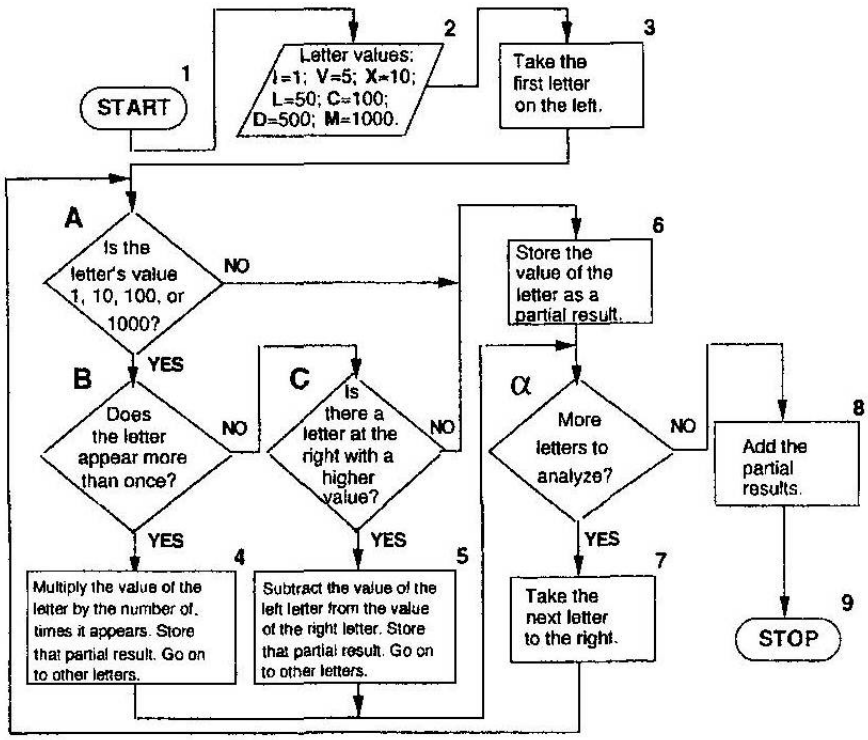

Figure 1. Flow chart of the procedure for reading Roman numerals. Operations are represented by rectangles and have been assigned numbers. Decisions are represented by diamonds and have been assigned letters. Latin letters correspond to identification decisions. Greek letters correspond to control decisions.

To determine different problem subclasses in the class of Roman numeral-reading problems, the flow chart shown in Figure 1 was decomposed according to LTD into the *spanning tree* shown in Figure 2 (Gibbons, 1985; Harary, 1969; McHugh, 1990). This tree was constructed by representing all of the different sequences of operations and decision options that lead to the STOP component of the algorithm. Trajectories involving option α (YES) are not included; each branch of the tree corresponds, then, with a particular subclass of problems.

In phase 3, numbers from 1 to 2000 written in Roman notation were solved by the procedure in Figure 1, and analyzed to determine three properties: iterativity, multiple array, and multiple grade.

Iterativity (ITE) is the number of *iterations* or "turns" required for the correct solution of a problem. Thus, ITE of "X" is 1, whereas ITE of "XV" is 2. ITE of the 2000 Roman numerals analyzed ranged from 1 to 7.

Multiple array (MUL ARR) is a succinct description of the sequence of actions required to solve problems in a subclass correctly. As the options taken in solving a problem determine the sequence of actions performed, an item can be characterized in terms of the sequence of decisions that must be made to solve the problem correctly. Thus, "XV" can be described this way:

A(YES), B(NO), C(NO),for the first iteration;
A(NO), for the second iteration.

LTD treats identification decisions as binary variables: 1 = YES, —1 = NO. This numerical

representation produces a set of variable values, called an *array* (Wirth, 1976). An array (AER) is obtained for each iteration. Each different ARR corresponds to a specific subclass of problems and, therefore, to a branch of the spanning tree (Figure 2). Formally, a MUL ALRR is a set of ARR's. If parentheses are used to distinguish iterations, and if zeros are used to represent the absence of a binary variable in an array, "XV" can be characterized as:

	A	B	C	A	B	C
MUL ARR("xv") = [(1		-1	-1)	(1	0	0)]

Multiple grade (MUL GRA) is the number of different ARR's in a problem's MUL ARR. Any problem solved with more than one iteration combines or repeats AER's across iterations. As there are four possible ARR's in the spanning tree, MUL GRA values observed in the 2000 Roman numeral problems ranged from to 4.

Sixteen combinations of the seven values of ITE and the four values of MIUL GRA were found to be represented in the 2000 items. For each combination, four items were used, each having a different first ARR in its MUL ARR. For instance, each of the four items with ITE = 4, MUL GRA = 3, had one of the four different possible ARR's at the beginning of its MUL ARR (Figure 2). The four different ARR's were thus equally represented in the test. This control was not possible for some combinations of lit and MUL GRA, for which fewer then four items were found in the 2000 numbers analyzed. The result of this selection process was an intentional sample of 53 items, out of the 64 intended, for the DIF prediction portion of the study.

Of those 53 items, 16 items were randomly selected, one for each combination of ITE and MUL GRA. For each of these 16 items, another item with the same MUL ARR was generated. Thus, the final number of items in the test was 69,32 of which formed 16 pairs having the same MUL ARR and, consequently, the same ITE. Of the remaining 37 items, 18 were used to form 9 item pairs having same ITE and different MUL ARR, and one item was left without a pair and was considered only for the DIF prediction part of the study. Item pairs, which were independent, were used for the part of the study dealing with item structural similarity.

Analysis and Procedure

Four variables were compared as to their ability to predict DIF, defined as the percentage of students responding correctly to each of the 53 items originally sampled. Two of those variables were drawn from LTD theory, ITE and MUL GRA. The third variable was expert estimation (EE). A mathematics educator was given the 53 items on 3" x 5" cards which he sorted into seven categories according to the difficulty he thought they posed for the students, with I being very difficult and 7 being very easy. The fourth variable was simply numeral size (NS), the number of characters in each of the 53 Roman numerals. NS ranged from 1 to 11 characters.

Structural similarity was indexed with the phi coefficient. It reflected the consistency of students' performance across item pairs varying in structural similarity. Structural similarity was defined in terms of having or not having the same MUL ARR and the same ITE. Three kinds of item pairs were used, structurally-different (9 pairs having different MUL ARR and different ITE), structurally-similar (9 pairs, different MUL ARR, same ITE), and structurally-equivalent (same MUL ARR and, consequently, same ITE).

To control for practice effects, items were randomly ordered within the test. To ensure that DIF had not been influenced by the location of items in the test, the total number of incorrect responses was counted for each of three parts of the test: beginning, middle, and end. An analysis

of variance found these differences to be non-significant. (All statistical analysis reported herein were conducted at $\alpha = .05$).

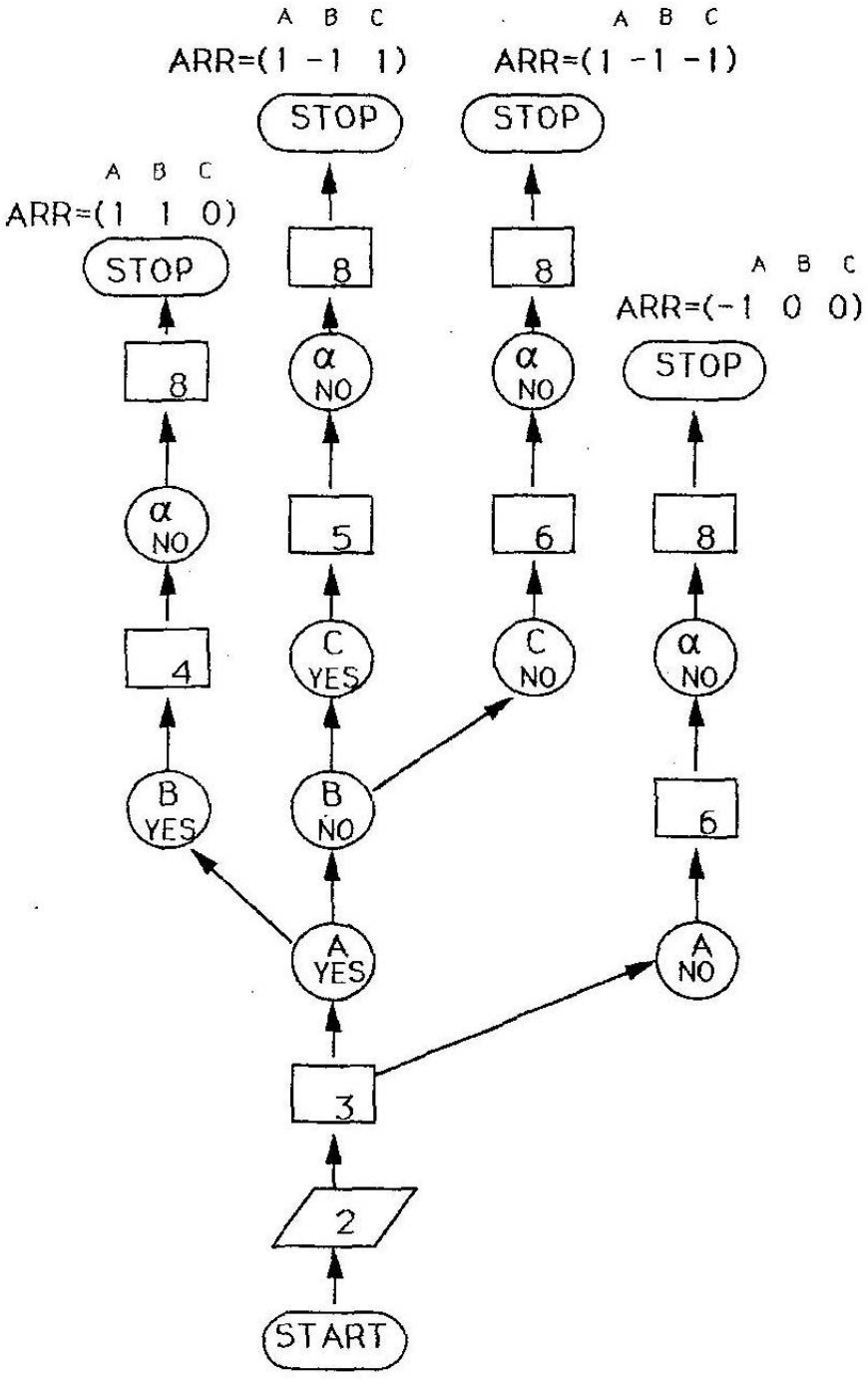

Figure 2. Spanning tree of the procedure for reading Roman numerals.

TABLE 1
Multiple Regression Coefficients in Item Difficulty Prediction and Pearson Correlations

Variable	*r*	*b*	Pearson correlations			
			ITE	MUL GRA	EE	NS
ITE	−.583	−4.721*	—	.559	.237	.807
MUL GRA	−.336	0.794		—	.459	.521
EE	−.425	−1.576*			—	.308
NS	−.278	1.816*				—

*$p < .05$.

The test was administered in group form; no time restrictions were imposed. Directions given to students asked them to do their best and to attempt each item.

Results and Discussion
Prediction of Item Difficulty

To examine the contention that LTD variables predict DIF more nearly accurately than conventional variables, first zero-order correlations were compared. The prediction contention was then tested by multiple linear regression, in which the following model was examined:

$$\text{DIF} = f(\text{ITE, MUL GRA, EE, NS})$$

DIF of the 53 items examined varied from 73.46 to 100 (Mean = 87.55, S.D. = 6.71). Correlation and multiple regression coefficients are shown in Table 1. ITE correlated highest with DIF (r = — .583), followed by EE (r = — .425), then NS (r = — .278). MUL GRA was not significant. (The reversal of sign for the regression coefficient of NS suggests that NS operated as a suppressor variable in the regression.) Of the four variables considered, only ITE and EE were important predictors of DIF. It should be noted, however, that these variables were correlated very low (r = .237), a circumstance which suggests that lit and EE heed different item characteristics.

In justifying his ranking of items, the expert said that much of the Roman numeral items' DIF was influenced by the students' lack of familiarity with some rather uncommon symbols like "D." According to this reasoning, "D" might be more difficult than "MCMXCI." This argument makes evident an assumption of LTD, that subjects store and retrieve every piece of information (every symbol, in this case) with complete accuracy, regardless of how frequently it takes part in the universe of problems, whereas an expert can distinguish differences in recall efficiency, a feature that accounts for the low correlation between ITE and EE.

To assess how the absence of recall accuracy in LTD theory affects predictive accuracy, two analyses were carried out. The first analysis was conducted to see whether the means of items having and not having an uncommon symbol ("D") were significantly different. In the second analysis, DIP was predicted by Equation I after dropping off those items having the uncommon ("D") symbol. It was expected that lit prediction of DIF would improve and that FE prediction of DIF would worsen when "D" items were removed.

TABLE 2

Mean Values and Significance *Test for Differences between Items Having and Not Having a Rare Symbol ("D")*

Variable Group of items	DIF	lit	MUL CR4	EE	NS
Without "D"	90.378	3.156	2.156	4.094	5.250
n = 32	(6.320)	(1.370)	(0.987)	(2.022)	(2.489)
With "D"	83.236	4.619	2.619	4.254	6.429
n = 21	(4.778)	(1.396)	(1.024)	(1.470)	(2.315)
Significant difference	Yes	Yes	No	No	No

For the first analysis, the same 53 items used for DIP prediction were partitioned into two groups, those having a "D" (21 items) and those not having a "D" (32 items). Mean differences of DIP, lit, MUL GRA, EE, and NS between both groups of items were tested (Table 2). DIP and lit differed significantly for items with and without the rare symbol. "D" items had, on average, a significantly higher ITE because the solution of the symbol "D" required an iteration. In support of the expert's argument, these items were significantly more difficult than items not having a "D," a difference to which LTD was not sensitive. It should be observed, however, that EE did not vary significantly for "D" items and for items not having a "D." A possible explanation for this outcome is that, although the expert was able to point out that recall efficiency may vary for each symbol, he was unable to apply this notion when estimating the difficulty of items in which "D" appeared in combination with other symbols.

For the second analysis, zero-order correlations and multiple regression coefficients were compared after removing "D" items. The results are shown in Table 3, which can be compared with those in Table 1 examine coefficient variations. When "D" items were removed, ITE and EE yielded again the highest correlations with DIF *(r* = - .535 and *r* = -.476, respectively). The effect of MUL GRA was again not significant, as NS operated again as a suppressor (b = 1.570). These correlations were almost the same as those in Table 1. The decrease in the correlation between EE and MUL GRA (from *r* = .459 to *r* = .396) and the increase in the correlation between EE and ITE (from *r* = .237 to *r* = .321) suggest that, when dealing with items not having a "D," ITE characteristics became increasingly salient in the expert's judgments.

TABLE 3

Multiple Regression Coefficients in Item Difficulty Prediction and Pearson Correlations after Removing Items Having a Rare Symbol ("D")

			Pearson correlations			
Variable	*r*	*b*	ITE	MUL GRA	EE	NS
ITE	−.535	−4.344*	—	.649	.321	.821
MUL GRA	−.439	0.202		—	.396	.469
EE	−.476	−1.181*			—	.309
NS	−.257	1.570*				—

*$p < .05$.

As items having an uncommon symbol were significantly more difficult (Table 2), ITE could be expected to predict DIP more nearly accurately when "D" items are not included in the analysis. Interestingly, that prediction did not occur. Apparently, the slight change in the correlation between JTE and DIF (from *r* = -*.583* to *r* = - *.535)* indicated that LTD prediction of DIP was not affected by how common each symbol was. A more reasonable possibility, however, was related to the fact that, in most of the cases, the presence of the symbol "D" in a problem constituted itself a portion of the problem, and a whole iteration was required to solve

this portion. Removing "D" items from the sample just decreased by one unit the average items' ITE.

LTD, then, lacked an important cognitive component, recall accuracy, that an expert is capable of acknowledging (though perhaps not of applying) in predicting DIP. A more nearly complete model should take into account this psychological factor.

Structural Similarity and item Association

To examine the consistency in performance across structurally different, similar, and equivalent item pairs, phi coefficients were calculated for the 16 item pairs. Overall item association for each category of item pairs was calculated as the percentage of item pairs having a significant phi coefficient.

Significant phi coefficients were found for 44.5%, *55.5%,* and 69% of the structurally different, similar, and equivalent item pairs, respectively. Although a chi-square test found these percentage differences not to be significant, they varied in the predicted direction.

Conclusions

This research has found, first, that LTD captures only part of the critical item characteristics that influence performance. Although ITE was found to be the most valid predictor of DIP, an expert was able to take into account a psychological factor, recall, to which LTD was not sensitive. It is acknowledged that other experts may vary in their ability to predict item difficulty and may base their judgments on item properties other than those used by the expert in this study. Nonetheless, this comparison showed that LTD and expert predictions might be based, in part, on different item characteristics. To accomplish a more rational prediction of performance based on LTD, this cognitive component should be considered.

Second, results from this research suggest that performance consistency is influenced by item structural similarity. Structural properties influenced performance even though they did not necessarily correspond with the items' appearances. For example, although "LXII" and "XXXVIII" have four and seven characters, respectively, their ITE is the same (ITE = 3). Likewise, "CCCLII" and "XXXVIII" look very different, but they are structurally equivalent. Thus, item structural properties may be more valid predictors than other, more conspicuous, item properties.

Though this research employed a particular class of problems (Roman numerals) and a particular algorithm, LTD principles are applicable to procedures involving at least one decision about the application of an operation. Further research is needed to establish how generalizable these findings are to other problem classes. Meanwhile, the finding that ITE affected DIP and that item structural similarity influenced item consistency suggests some item properties that might be taken into account when writing, selecting, and banking test items. Bejar (1985) has observed that test developers should make an effort to identify the characteristics of *unpretested* items and to generate items according to a given set of specifications, and Embretson (1985b) has suggested that a test developer can be thought of as a person who controls and manipulates variables that constitute the formal properties of items. Results of this research suggest that LTD, or a revised version of it, provides one possible, viable theory for doing so.

The author is indebted to Dr. Richard J. Shavelson, whose valuable comments and expert advice made possible the conclusion of this research, and to Dr. Gail P. Baxter for her comments to the original manuscript. Requests for reprints should be addressed to the author at the Graduate School of Education, University of California, Santa Barbara. Santa Barbara, CA 93106.

REFERENCES

Anderson, J. R. (1985). *Cognitive psychology and its implications.* New York: W. I-I. Freeman.

Ashcraft, M. H. (1989). *Human memory and cognition.* Glenview, Illinois: Scott, Foresman.

Bejar, 1.1. (1985). Speculations on the future of test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics.* New York: Academic Press.

Bormouth, J. R. (1970). *On the theory of achievement test items.* Chicago, Illinois: University of Chicago Press.

Conoley, J. 0. and O'Neil, H. F., Jr. (1979). A primer for developing test items. In H. P. O'Neil, Jr. (Ed.), *Procedures for instructional systems development.* New York: Academic Press.

Ebel, II. (1972). *Essentials of educational measurement.* Englewood Cliffs, New Jersey: Prentice Hall.

Embretson, S. E. (Ed.) (1985 a). *Test design: Developments in psychology and psychometrics.* New York: Academic Press.

Embretson, S. E. (1985 b). Introduction to the problem of test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics.* New York: Academic Press.

Farina, M. V. (1970). *Flowcharting.* Englewood Cliffs, New Jersey:Prentice-Hall.

Gibbons, A. (1985). *Algorithmic graph theory.* Cambridge: Cambridge University Press.

Gronlund, N. E. (1971). *Measurement and evaluation in teaching.* New York: Macmillan.

Gronlund, N. E. (1988). *flow to construct achievement tests.* Englewood Cliffs, New Jersey: Prentice-Hall.

Guttman, L. (1969). Integration of test design and analysis. In *Proceedings of the 1969 invitational conference on testing problems.* Princeton, New Jersey: Educational Testing Service.

Haraiy, P. (1969). *Graph theory.* Reading, Massachusetts: Addison-Wesley.

Hively, W. (1974). Introduction to domain referenced testing. *Educational Technology,* 14, 5-9.

Hively, W., Patterson, H. L., and Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement,* 5(4), 275-290.

Jonassen, D. H., Hannum, W. H., and Tessmer, M. (1989). *Handbook of task analysis procedures.* New York: Praeger.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Massachusetts: Addison Wesley.

Martin, J. and McClure, C. (1985). *Diagramming techniques for analysts and programmers.* Englewood Cliffs, New Jersey: Prentice-Hall.

McHugh, J. A. (1990). *Algorithmic graph theory.* Englewood Cliffs, New Jersey: Prentice Hall.

Millman, J. (1980). Computer-based item generation. In R. A. Berk (Ed.), *Criterion-referenced measurement: the state of the art.* Baltimore: The Johns Hopkins University Press.

Nunally, J. C. (1964). *Educational measurement and evaluation.* New York: McGraw-Hill, 1964.

Nunally, 1. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Osterlind, S. J. (1989). *Constructing test items.* Boston: Kluwer Academic Publisher.

Payne, D. A. (1968). *The specification and measurement of learning outcomes.* Lexington, Massachusetts: Xerox.

Priestley, M. (1982). *Performance assessment in education and training: alternative techniques.* Englewood Cliffs, New Jersey: Education Technology Publications.

Roid, G. H. and Haladyna, T. M. (1978). A comparison of objective-based and modified-Bormouth item writing techniques. Educational and Psychological Measurement, 35, 19-28.

Roid, G. H. and Haladyna, T. M. (1982). *A technology for test-item writing.* New .York: Academic Press.
Scandura, J. (1973). *Structural learning, Vol. 1: Theory and research.* New York: Gordon Breach and Associates.
Schriber, T. J. (1969). *Fundamentals of flowcharting.* New York: Wiley.
Solano-Flores, G. (1989). *Estimación empirica de Ia capacidad predictiva de Ia técnica de diseno logico de erámenes.* Unpublished Master's dissertation. Mexico City, Mexico: Univcrsidad Nacional Autónoma de Mexico.
Solano-Flores, G. (1991). *Diseno lOgico de e.xámenes.* Mexico:Trillas.
Stern, N. B. *(1975). Flowcharting: A tool for understanding computer logic.* New York: Wiley.
Sternberg, S. (1969). The discovery of processing stages: extensions of Donder's method. In W. G. Koster (Ed.), Attention and performance II. *Acta Psychologica,* 30, 276—315.
Thorndike, R. L. and Hagen, E. (1969). *Measurement and evaluation in psychology and education.* New York: Wiley.
Trakhtenbrot, B. A. (1963). *Algorithms and automatic computing machines.* Chicago: University of Chicago Press.
Wheatley, D. M. and Unwin, A. W. (1972). *The algorithm writer's guide.* London: Longman.
Wirth, N. *(1976).Algorithms + data structures = programs.* Englewood Cliffs, New Jersey: Prentice-Hall.
Wood, R. (1977). Multiple-choice: a state of the art report. *Evaluation in Education,* 1, 191-280.