DOCUMENT RESUME

ED 199 266 TM 810 162

AUTHOR Thompson, Bruce; Barnitz, John
TITLE The Use of Latent Trait Models in Psycholinguistic Research.
PUB DATE 81
NOTE 22p.: Paper presented at the Annual Meeting of the Southwest Educational Research Association (Dallas, TX, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Comparative Analysis; Elementary Education; Grade 2; Grade 4; Grade 6; *Latent Trait Theory; Measurement Techniques; *Psycholinguistics; *Reading Achievement; *Reading Research; *Scoring Formulas
IDENTIFIERS *Rasch Model

ABSTRACT

Several applications of latent trait models in reading research are discussed, their nature reviewed, and two experiments presented as heuristic applications of the Rasch Model. Two studies of children's comprehension of selected anaphoric structures in prose were performed using both conventional and Rasch Model analyses. Twenty-six children (grades 2, 4, and 6) in the first study and 91 children (grades 2, 4, and 6) in the second study were asked to read 16 passages containing the pronoun "it." After reading each passage, the subjects responded to a question requiring identification of the pronoun's referent, and the responses were scored right or wrong. Both conventional and Rasch Model scoring procedures revealed that: (1) variances for the three grades were equal in the first study; and (2) no significant sex effect was identified in the second study. Both studies demonstrated that benefits can be derived by utilizing Rasch Model measurement procedures in reading research. Progress in reading research depends upon replication of findings across various studies, and the Rasch Model facilitates this process because item (or person) calibrations are sample (or item) free. (RL)

The Use of Latent Trait Models in Psycholinguistic Research

Bruce Thompson

John Barnitz

University of New Orleans

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, 1981.

Abstract

Recent literature on reading processes reflects improved use of research methodologies. Nevertheless, several improvements can still be achieved. For example, progress in realizing better understanding of psycholinguistic processes might be facilitated by the use of latent trait measurement models. The features of these models are discussed intuitively. Two heuristic applications of a latent trait model are presented. Subjects in both these experiments were elementary school children. Both studies examined the impacts of selected variations in the anaphoric structures of prose. Finally, some potential advantages of the latent trait measurement models are discussed.

The last decade has witnessed a dramatic growth in understanding of reading processes. Today, there is less of an inverse relationship "between the size and complexity of the lingiuistic unit being studied and the amount of research devoted to that unit" (Thorndyke, 1975, p. 1). An important benchmark of this progress is the theory advanced by researchers such as Anderson and Bower (1973), van Dijk (1973), Kintsch (1974), Frederiksen (1975), Rumelhart (1975), and Anderson, Reynolds, Schallert, and Goetz (1976).

Several factors have facilitated this progress. But these gains have been realized at least partly because reading researchers have become more sophisticated in the methodologies which they employ. In particular, improvements have occurred in the application of analysis of variance (ANOVA) procedures, procedures which are frequently applied in this area of inquiry. For example, Clark (1973) and later Coleman and Morris (1978) have influenced the selection of the error terms which are used to evaluate treatment effects. Marascuilo and Levin (1970, 1976) have pointed out the importance of avoiding Type IV errors when conducting tests of certain hypotheses.

Notwithstanding past improvements in methodological practice, however, some additional improvements remain desirable. Morrow and Frankiewicz (1979) have identified certain errors which continue to be made in some applications of ANOVA and ANOVA analogues. Also, the myth that analysis of covariance can always magically equalize non-equivalent control groups has not yet fully been dispelled (Campbell & Erlebacher, 1975). Finally, researchers have yet to give adequate attention to "power" considerations when reporting their work (Cohen & Hyman, 1979). An even more fundamental error, however, is manifest in the measurement which is performed in some reading research. Unfortunately, even the most elaborate test statistics can not rescue a study from the pitfalls of inappropriate measurement.

These limitations might be avoided if researchers made more use of "latent trait" measurement models. These models have been usefully applied in myriad content areas, including intelligence testing (Anderson, Kearney, & Everett, 1968), the preparation of Civil Service examinations (Durovic, 1970), and mathematics testing (Connolly, Nachtman, & Pritchett, 1974). Other example applications of latent trait models have been listed by Rentz and Rentz (1978). However, applications of latent trait models in reading research remain rare indeed. The few exceptions to this rule include studies by Rentz and Bashaw (1977) and by Andrich and Godfrey (1979).

The purpose of this paper is to discuss several applications of latent trait models in reading research. Specifically, the paper reviews on an intuitive level the nature of latent trait models, and presents two experiments as heuristic applications of one latent trait method, the Rasch model.

Overview of latent trait models

Latent trait theory proposes that the abilities of tested subjects are latent in their test item responses, but can be estimated by specifying the nature of the relationship between observed performance and the unobserved traits which are presumed to underlie performance. Several latent trait models have been delineated, including models proposed by Lord (1952) and by Birnbaum (1968). However, probably the most widely known and most frequently applied latent trait model is the model proposed by Rasch (1960), and it is the Rasch model which is discussed and applied in this paper. A more complete discussion of latent trait theory and other latent trait models is available elsewhere (cf. Hambleton, Swamiwathan, Cook, Eignor, & Gifford, 1978).

The logic of the Rasch model is quite straightforward. As Wright and Stone (1979, p. xiii) explain, the model assumes that success on any test item is "entirely governed by the difference betweeen the ability of the person and the difficulty of the item. Nothing more. The more able the person, the better their [sic] chances for success with any item. The easier the item, the more likely any person is to solve it. It is as simple as that." The mathematics necessary for estimating the latent difficulty of each test item and the latent ability of each tested subject are not quite so simple, although reasonable approximations of estimates can be calculated by hand if the researcher does not have access to an appropriate computer program (cf. Wright & Stone, 1979, chapter 2). However, three aspects of the model are noteworthy.

The model is orderly. Other measurement approaches can posit that the more able of two persons is always more likely to succeed on any given item, or that any given person is always more likely to succeed on the easier of any two items. However, the Rasch model requires that these assumptions be made, and more importantly provides test statistics which can be employed to evaluate deviations from expected performance by either persons or items.

The model is also objective. When conventional measurement procedures are used, item difficulty estimates are not invariant across different samples of subjects, and the ability scores assigned to subjects are not invariant across different tests. However, the Rasch model does generate both *sample-free* item calibrations and *test-free* person ability scores. The importance of these kinds of estimates at first may be difficult to comprehend, but the magnitude of this contribution has been recognized by researchers such as Loevinger (1965, p. 151), who noted that "Rasch must be credited with an outstanding contribution to one of the two central psychometric problems, the achievement of nonarbitrary measures."

Finally, the calibrations generated by the model are truly <u>interval</u>. Reading researchers frequently require subjects to read prose or to memorize words or symbols. Each subsequent task, e.g.—recall, closure, etc., is typically then scored "1" for a success or "0" for a failure. Next, scores are summed across items for each subject in order to arrive at an aggregate unit of analysis.

One problem with this process is that the difficulties of the items are presumed to be at least approximately the same. This assumption means that the item scores can legitimately be summed to provide a total test score. Unfortuantely, most researchers rarely test how well this assumption applies for a given data set. The importance of the violation of this additivity assumption will be demonstrated in at least one of the two experiments reported here.

Heuristic applications of the model

Two studies of children's comprehension of selected anaphoric structures in prose were performed to demonstrate some applications of the Rasch model in reading research. Both conventional and Rasch model analyses were performed in both studies so that a comparison of methods would be facilitated. Different children served as subjects in the two studies. However, the subjects in both studies were native English speaking working class children in grades two, four, and six. Subjects were excluded from the study if their standardized reading test scores were substantially below average.

Both studies investigated the comprehension of pronouns embedded in passages with different structures. This area of investigation is currently receiving considerable attention (cf. Richek, 1977). For example, Bormuth, et al. (1970) presented fourth grade students with short passages containing pronouns embedded in different structures, and then identified a hierarchy of difficulty for the various structures. However, Lesgold (1974) challenged this heirarchy in a study which produced somewhat different results. Of course, some variation in findings should be expected, since the background knowledge of subjects and the semantic content of passages can interact and override the influence of syntactic passage features (cf. Rumelhart, 1977).

At least three variations in the presentation of a pronoun in a passage can be identified. First, a pronoun's referent can either precede or follow the pronoun. Chomsky's (1969) research suggests that forward structures are easier for young children to comprehend orally. Second, a pronoun's referent may either be within the same sentence or be within another sentence. Third, a pronoun's referent may either be a noun phrase or a longer clause or sentence.

Although the wording and content of the passages used in the studies varied, in both studies the subjects were asked to read 16 passages containing the pronoun "it." After reading each passage, the subjects responded to a question requiring the identification of the pronoun's referent, and the responses were scored right-wrong according to whether the correct referent was identified or a distractor item was choosen. Both studies utilized two passages representing each of the structure combinations presented in Table 1.

Insert Table 1 about here.

Experiment I

The subjects in the first experiment were 26 children from each of the grades, grades two, four, and six. The global null hypothes[illegible] the study was that there would be no statistically significant difference[illegible] [illegible]g the three mean test scores of the children in the three different grade levels. After the data were collected, the data were analyzed to determine if any items or any subjects behaved in a manner which deviated substantially from Rasch model expectations. No subjects and no items were identified (α=.05) as model "misfits," i.e.--deviated substantially from expected behavior. Consequently, sample-free item difficulties and test-free person ability estimates could be and were derived using all 16 test items and all 78 subjects.

In order to provide a direct comparison between conventional and Rasch model scoring procedures, the tests were scored in two ways. The tests were scored by counting the number of right answers each person selected. The tests were also scored by cumulating the sample-free item difficulty estimates for each item which each person correctly answered, after a constant was added to the difficulty estimates so that none were negative.

For both scoring procedures, a prelimanary null hypothesis that the variances for the three grades were equal was tested. For the conventional scoring procedure the preliminary null hypothesis was not rejected (Bartlett's $\underline{F}$=.3, p>.05). For the Rasch model scoring procedure the preliminary null hypothesis was not rejected ($\underline{F}$=.8, p>.05). These results suggested that ANOVA's could be conducted without violating the homogeneity of variance assumption.

Since the grade-way was quantitative and the levels within the way were equally spaced, a priori polynomial contrasts were applied to identify whether or not any observed differences among the three means reflected either a linear or a non-linear trend. ANOVA keyouts from both analyses are presented in Table 2. The Table 2 keyout illustrates that the results of the two procedures can lead to different conclusions.

<u>Insert Table 2 about here.</u>

On a substantive level, the sample-free item difficulty scoring procedure suggests that between grades two and four children improve in their ability to interpret the pronoun "it." This finding is consistent with past research. However, after the fourth grade there is apparently less motivation for children to focus on highly specific syntactic features of the prose which they read. This finding is consistent with a belief that as children become more proficient at using syntactic rules, they focus more on an interactive combination of the syntactic and the content features of prose (Pearson & Kamil, 1978). Of course, this result may be a sampling artifact which would not be replicated in a longitudinal study. The external validity of this result remains to be explored in future research.

Experiment II

The subjects in the second experiment were 91 second, fourth, and sixth graders. Of the 91 subjects, 44 were boys and 47 were girls. The null hypothesis of the study was that the mean ability score of the boys would not be significantly different from the mean score of the girls. This hypothesis was of limited substantive interest, but will facilitate discussion of some additional features of latent trait methods. After the data were collected, the data were analyzed to determine if any items or subjects behaved in a manner which deviated substantially ($\alpha = .05$) from Rasch model expectations. No items were identified as "misfits," but one subject did deviate substantially from expected performance ($\underline{t}=2.1$, $p<.05$).

Table 3 presents the expected and the actual performance of the "misfitting" subject on the 16 test items. The items are listed in order of their sample-free difficulty estimates. Since the subject made seven correct responses, it should be expected that the seven easiest items would have been correctly answered while the remaining nine items would have been missed. Instead, this individual missed two of the three easiest items and correctly answered the two most difficult items. Wright and Stone (1979) might call this a combined "sleeping," i.e.-- warm-up, and "guessing" pattern. Because the subject deviated substantially from expected performance, the subject was excluded from further analysis.

Insert Table 3 about here.

Table 3 also illustrates that [illegible]-free item difficulty estimates can differ from the sample-bound item difficulty estimates. Some items with identical sample-bound difficulty estimates have different sample-free values, and vice versa. Of course, the magnitude of these differences will vary from study to study, but clearly the different estimates will not necessarily be similar.

In order to test the null hypothesis of the experiment, person ability scores were first estimated in the conventional manner, i.e.--by counting each person's number of correct responses. The Rasch test-free person ability scores were used for the alternative scoring procedure. In this study, the homogeniety of variance assumption was not violated when either of the two scoring procedures was used, and so one-way ANOVA's were performed. No significant sex effect was identified using the conventional scoring procedure (F=.2, p>.05), nor was a significant sex effect identified using the Rasch test-free person ability scores (F= [illegible], p>.05).

Discussion

The two experiments demonstrate several important benefits which can be derived by utilizing Rasch model measurement procedures in reading research. At least two of these benefits merit particular emphasis. Progress in reading research depends in the final analysis upon replication of findings across various studies. The growing emphasis on replication has been reflected in some recent essays (cf. Carver, 1978), and developing methodologies for empirically integrating research studies (cf. Iverson & Walberg, 1979). The use of the Rasch model facilitates this process, because item (or person) calibrations are sample (or item) free, and consequently can be more sensibly combined across studies.

Figure 1 provides a huerisuic demonstration of such an integration. The figure integrates the difficulty estimates for the different passages across the two different samples. Of course, the person-free difficulty estimates may themselves have important implications for psycholinguistic theory. For example, Figure 1 suggests that forward referent order and noun-referent structures are easier for children to interpret than backward referent order and sentence-referent structure, respectively.

<u>Insert Figure 1 about here.</u>

The importance of the misfit statistics of the Rasch model is also noteworthy. Reading researchers frequently eliminate subjects who do not meet minimal ability criteria. Even when this is done, some subjects whose test performance reflects either "sleeping" or "guessing" or both will unfortunately be included in conventional analyses. Similarly, conventional analyses will not identify "misbehaving" test items unless an items's behavior is genuinely bizarre, e.g.— everybody misses the item. However, the Rasch model integrates expectations about item and person behavior, and provides test statistics for evaluating deviations from expectations.

To date, "the major factors that have hindered widespread use of these methods are the lack of familiarity on the part of practitioners and the lack of user oriented computer programs" (Hambleton, et al., 1978, p. 503). However, these difficulties can now be at least partially overcome by consulting one of the recently published texts on latent trait measurement (cf. Wright & Stone, 1979), and by acquiring one of the recently developed computer programs which implement these models. In summary, latent trait measurement models appear to have some potentially helpful applications in psycholinguistic inquiry; these potentials have not yet been fully realized.

References

Anderson, J., & Bower, G. Human associative memory. Washington: Winston, 1973.

Anderson, J., Kearney, G.E., & Everett, A.V. An evaluation of Rasch's structural model for test items. British Journal of Mathematical and Statistical Psychology, 1968, 21, 231-238.

Anderson, R.C., Reynolds, R.G., Schallert, D.L., & Goetz, E.T. Frameworks for comprehending discourse (Technical Report No. 12). Urbana: Center for the Study of Reading, University of Illinois, 1976. (ERIC Document Reproduction No. ED 134 935)

Andrich, D., & Godfrey, J.R. Hierarchies in the skills of Davis' Reading Comprehension Test, Form D: An empirical investigation using a latent trait model. Reading Research Quarterly, 1979, 14, 182-200.

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

Bormuth, J.R., Manning, J., Carr, J., & Pearson, P.D. Children's comprehension of between and within sentence syntactic structure. Journal of Educational Psychology, 1970, 61, 349-357.

Campbell, D.T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In E.L. Struening & M. Guttentag (Eds.), Handbook of evaluation research (Vol. 1). Beverly Hills: Sage, 1975.

Carver, R.P. The case against statistical significance testing. *Harvard Education Review*, 1978, *48*, 378-399.

Chomsky, C. *The acquisition of syntax in children from five to ten*. Cambridge: M.I.T. Press, 1969.

Clark, H. The language as fixed effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 1973, *12*, 335-359.

Cohen, S.A., & Hyman, J.S. How come so many hypotheses in educational research are supported? (A modest proposal). *Educational Researcher*, 1979, *8*, 12-16.

Coleman, E.B., & Morris, G. Generalization tests: A terminology that focuses attention on fixed-effects restrictions. *Journal of Reading Behavior*, 1978, *10*, 377-392.

Connolly, A.J., Nachtman, W., & Pritchett, E.M. *Key math diagnostic arithmetic test*. Circle Pines, MN: American Guidance Service, 1974.

Durovic, J. *Application of the Rasch model to Civil Service testing*. Paper presented at the annual meeting of the Northeastern Educational Research Association, 1970. (ERIC Document Reproduction Service No. ED 049 305)

Frederiksen, C. Effects of context-induced processing operations on semantic information required from discourse. *Cognitive Psychology*, 1975, *7*, 139-166.

Hambleton, R.K., Swamiwathan, H., Cook, L.L., Eignor, D.R., & Gifford, J.A. Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research*, 1978, *48*, 467-510.

Iverson, B.K., & Walberg, H.J. Home environment and learning: A quantitative synthesis. Paper presented at the annual meeting of American Educational Research Association, 1979. (ERIC Document Reproduction Service No. 168 714)

Kintsch, W. The representation of meaning in memory. Potomac, MD: Erlbaum, 1974.

Lesgold, A.M. Variability in children's comprehension of syntactic structures. Journal of Educational Psychology, 1974, 66, 333-338.

Loevinger, J. Person and population as psychometric concepts. Psychological Review, 1965, 72, 143-155.

Lord, F.M. A theory of test scores. Psychometric monograph, 1952, No. 7.

Marascuilo, L.A., & Levin, J.R. Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type IV errors. American Educational Research Journal, 1970, 7, 397-421.

Marascuilo, L.A., & Levin, J.R. The simultaneous investigation of interaction and nested hypotheses in two-factor analysis of variance designs. American Educational Research Journal, 1976, 13, 61-65.

Morrow, J.P., J.r., & Frankiewicz, R.G. Strategies for the analysis of repeated and multiple measure designs. Research Quarterly, 1979, 50, 297-304.

Pearson, P.D., & Kamil, M.L. Basic processes and instructional practices in teaching reading (Reading Education Report No. 7). Urbana: Center for the Study of Reading, University of Illinois, 1978. (ERIC Document Reproduction Service No. ED 165 118)

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institute, 1960 (To be reprinted by the University of Chicago Press, 1980).

Rentz, R.R., & Bashaw, W.L. The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-180.

Rentz, R.R., & Rentz, C.C. Does the Rasch model really work? A Discussion for practitioners. Princeton, NJ: ERIC Clearinghouse of Tests, Measurements, and Evaluations, 1978. (ERIC Document Reproduction Service No. ED 169 137)

Richek, M.A. Reading comprehension of anaphoric forms in varying linguistic contexts. Reading Research Quarterly, 1977, 12, 145-165.

Rumelhart, D. Notes on a schema for stories. In D. Borrow & A. Collins (Eds.), Representations and understanding: Studies in cognitive science. New York: Academic Press, 1975.

Rumelhart, D.E. Toward an interactive model of reading. In S. Dornic (Ed.), Attention and performance VI. New York: Academic Press, 1977.

Thorndyke, P.W. Cognitive structures in human story comprehension and memory. Unpublished doctoral dissertation, Stanford University, 1975. (ERIC Document Reproduction Service No. ED 12? 587)

van Dijk, T. Text grammar and text logic. In J.S. Petofl & H. Reiser (Eds.), Studies in text grammar. Dodrecht: Reidel, 1973.

Wright, B.D., & Stone, M.H. Best test design. Chicago: MESA Press, 1979.

Table 1

Structure Combinations

Order	Distance	Referent Type	Acronym
Forward	intrA	Noun phrase	FAN
Forward	intrA	Sentence/clause	FAS
Forward	inteR	Noun phrase	FRN
Forward	inteR	Sentence/clause	FRS
Backward	intrA	Noun phrase	BAN
Backward	intrA	Sentence/clause	BAS
Backward	inteR	Noun phrase	BRN
Backward	inteR	Sentence/clause	BRS

Note. Hereafter the passages are arbitrarily each numbered "1" or "2." Thus, "FAN1" refers to number one of two passages with a Noun phrase referent presented in an intrA-sentence Forward referent-order (FAN) structure.

Table 2

ANOVA Keyouts for Grade-level Hypothesis

Method	Source	Sum of Squares	*df*	Mean Squares	*F*
Conventional	Linear	111.1	1	111.1	15.6***
	Non-linear	13.6	1	13.6	1.9
	Within	534.2	75	7.1	
Rasch	Linear	.5	1	.5	.3
	Non-linear	8.1	1	8.1	5.1*
	Within	118.2	75	1.6	

*$p<.05$

**$p<.01$

***$p<.001$

Table 3

"Misfitting" Person's Performance

Item Acronym[a]	Actual Performance[b]	Expected Performance	Item Difficulties Rasch	Item Difficulties Conventional[c]
BAN1	0	1	-2.43	.93
FRN2	1	1	-1.79	.91
FRS2	0	1	-1.36	.87
FRN1	1	1	-1.24	.87
FAN1	1	1	-.74	.81
FAS1	0	1	-.27	.74
FAS2	1	1	-.14	.64
BRN1	0	0	-.14	.71
FAN2	0	0	.30	.64
BAN2	0	0	.41	.62
BAS2	1	0	.47	.62
BRS1	0	0	.58	.58
BRN2	0	0	.96	.50
BRS2	0	0	1.50	.40
FRS1	1	0	1.73	.36
BAS2	1	0	2.15	.28

[a] See Table 1 for acronym derivatives.

[b] Scored "1"=right; "0"=wrong.

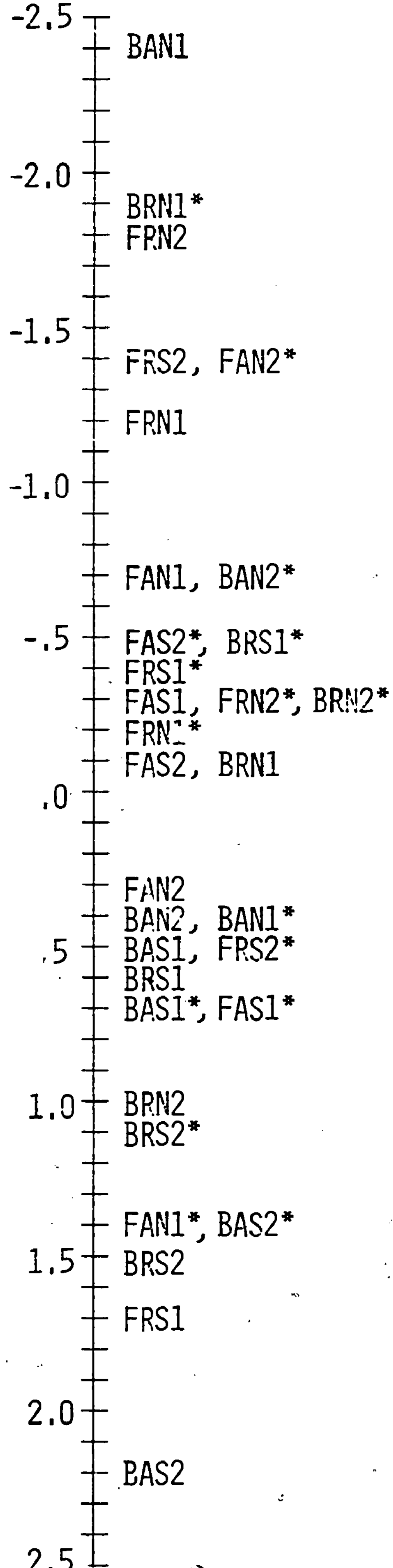
[c] $\underline{n}$ of subjects correctly answering item divided by $\underline{n}$ of subjects.

Figure Caption and Note

Figure 1

Structures Arrayed Along Sample-free Difficulty Continuum

Note. See Table 1 for acronym derivatives. The passages presented in Experiment I are identified by asterisks.


-2.5
BAN1
-2.0
BRN1*
FRN2
-1.5
FRS2, FAN2*
FRN1
-1.0
FAN1, BAN2*
-.5
FAS2*, BRS1*
FRS1*
FAS1, FRN2*, BRN2*
FRN1*
FAS2, BRN1
.0
FAN2
BAN2, BAN1*
.5
BAS1, FRS2*
BRS1
BAS1*, FAS1*
1.0
BRN2
BRS2*
FAN1*, BAS2*
1.5
BRS2
FRS1
2.0
BAS2
2.5