DOCUMENT RESUME

ED 127 363 TM 005 458

AUTHOR Bart, William M.; Airasian, Peter W.

TITLE An Empirical Examination of the Relationship Between Test Factor Structure and Test Hierarchical Structure.

PUB DATE [Apr 76]

NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.

DESCRIPTORS Comparative Analysis; Correlation; Factor Analysis; *Factor Structure; Goodness of Fit; *Item Analysis; Mathematical Models; Multidimensional Scaling; Probability; Response Style (Tests); Statistical Analysis; *Tests

IDENTIFIERS *Hierarchical Structure (Tests); Latent Trait Statistical Test Theory; Ordering Theory

ABSTRACT

The question of whether test factor structure is indicative of the test item hierarchy was examined. Data from 1,000 subjects on two sets of five bivalued Law School Admission Test items, which were analyzed with latent trait methods of Bock and Lieberman and of Christoffersson in Psychometrika, were analyzed with an ordering-theoretic method to locate item hierarchies. Though one item set was unifactor and the other bifactor, both item sets showed no hierarchies even after 10 percent of the subjects responsible for low response pattern frequencies were deleted. Factor structure of a test appears to reveal nothing about the test's hierarchical structure. (Author)

An Empirical Examination of the Relationship between Test Factor Structure and Test Hierarchical Structure

William M. Bart
University of Minnesota

Peter W. Airasian
Boston College

Paper presented at the
annual meeting of the
American Educational
Research Association,
San Francisco, 1976

A central issue in psychometrics is the analysis of the structure among the dichotomous items in a test. Test structure is conceived in two basic forms: 1) test factor structure which is the system of latent traits (factors or dimensions) which the items measure and 2) test hierarchical structure which is the network of prerequisite relations among the items. To study these types of test structure, two approaches to test data analysis have developed: 1) latent trait statistical theory and 2) ordering theory. The first approach provides information as to how many latent traits are needed to describe the correlational relationships among test items as well as provide information as to invariant item difficulties and person ability estimates. The second approach provides information as to the hierarchical structure that best describes the system of logical relationships among test items.

Latent trait statistical test theory has its roots in the model for dichotomous item responses developed by Lawley (1943). This model, in turn, formed the basis for the "normal ogive" model, which has been further developed by Lord (1952, 1953), Lord & Novick (1968), and Samejima (1969). Bock and Lieberman (1970) and Christoffersson (1975) have provided further refinements in the normal ogive model for the analysis of dichotomously scored items from a consideration of test factor structure.

Ordering theory, on the other hand, is a measurement model which has as its primary intent either the testing of hypothesized hierarchies among bivalent items or the determination of hierarchies among bivalued items. Methods of ordering theory have been described in a book of published reports in ordering theory (Krus, Bart, and Airasian, 1975). Applications of ordering theory have been made in the investigation of Piagetian task hierarchies (Bart and Airasian, 1974), instructional sequences (Airasian and Bart, 1975), attitudes toward school (Airasian, Madaus, and Woods, 1975), and other educational psychological topics.

Although any test of bivalent items has a factor structure and a hierarchical structure, the relationship between the two types of test structure has not yet been examined. Are the two types of structures redundant? Is there a correspondence between hierarchical structures and factor structures such as that items fitting a n-factor structure would have a hierarchy with n branches and vice versa? Or are the two types of structures unrelated by such a correspondence and thus are psychometrically distinct?

In behavioral sciences such as instructional psychology and the psychology of intelligence, there have developed models which posit psychological variables as continuous dimensional traits (e.g., Guilford, 1956) and models which posit psychological variables as classes of behaviors and skills which have internal logical structures (e.g., Piaget, 1950; Gagne, 1968). Along with those types of models have developed continuous measurement models such as factor analysis and discrete measurement models such as ordering theory which have contributed to the empirical examination of various examples of the two types of psychological models. To study the relationship between hierarchical structures and factor structures is to study the relationship between discrete and continuous measurement models and thus to examine the relationship between discrete and psychological theories. Herein lies the relevance of this study. Are the two types of behavioral science theories and their corresponding measurement models closely related or is there a conceptual chasm between them? As a step in the answer to that question, the following question will be investigated in this study: can the factor structure of a test tell us something about the hierarchical structure?

METHOD

Four methods of bivalent item analysis were compared on a single set of test data. Two methods come from latent trait statistical test theory: 1) the Bock-Lieberman (1970) unconditional maximum likelihood method of estimating parameters of the normal ogive model for dichotomously-scored item response patterns and 2) the Christoffersson (1975) generalized least square method for multiple factor analysis of dichotomized variables. Two methods come from ordering theory: 1) the Krus-Bart (1974) ordering-theoretic method of non-correlational multidimensional scaling of bivalued items and 2) an ordering-theoretic method, derived from the methods of Airasian and Bart (1973) and of Bart and Krus (1973), to identify the hierarchy of test items that best fits bivalued test item data so that item intransitivities cannot occur.

In the normal ogive model used by Bock and Lieberman (1970), the probability of success at an item is viewed as a normally-distributed function of the difficulty and discriminating power of the item and the latent ability of the subject. Their method was an unconditional maximum likelihood one, because the data was regarded as coming from a sample of subjects from a specified population and because item parameters were estimated from integration of the probability function over the distribution of latent ability. Their method provided also a goodness of fit test for the normal ogive framework.

In the Christoffersson (1975) method, the three types of parameters considered in the previous paper were again considered. However, the marginal distributions of single and pairs of items are used to allow for a technique that permits more than 10-12 items to be analyzed which was the limit for the previous study. In addition, a goodness of fit test was offered.

In the Krus-Bart (1974) method, the total tested sample of subjects is divided into disjoint, mutually exclusive classes such that the response patterns of subjects in any class complies with a linear hierarchy among the items and that the numbers of response patterns for the classes decrease in a monotone manner. The number of 01 and 10 scores relating each item and the other items are counted in each class to produce a non-correlational equivalent to the factor loading relating an item and the underlying trait. In this method, each class of subjects determines a factor and its corresponding factor loadings. Bart and Krus (1973) recommended that hierarchies be built from the inter-item prerequisite relations. An item i is prerequisite to item j if and only if the frequency of 01 item response patterns for items i and j respectively is equal to or smaller than some pre-established tolerance level. If the tolerance level is greater than 0, then item intransitivities (i.e., if item i is prerequisite to item j and item j is prerequisite to item k, then item i is not prequisite to item k) are possible. However, transitivity is a basic property of a hierarchy; therefore, a method which allows no item intransitivities is preferable. The method used in this study to determine bivalent test item hierarchical structure had two phases: 1) item response patterns, whose frequencies were less than or equal to some minimum and whose combined frequencies were less than or equal to some pre-established percentage of the sample size, are deleted as suggested by Airasian and Bart (1973) and 2) inter-item prerequisite relations are derived from the remaining item response patterns using the method of Bart and Krus (1973) with a tolerance level of 0. This composite method allows no item intransitivities and incorporates no infrequently occurring response patterns.

The four-fold table relating two bivalent items used in the composite method can also be used to note a relationship between factor structure and hierarchy structure. If item i is a prerequisite to item j and if the only zero entry in the table is in the 01 cell for items i and j respectively, then the phi coefficient will be large to the extent that the 10 cell entry is small. If item i is logically equivalent to item j and if the 01 and 10 cell entries are zero, then the phi coefficient will be 1. Thus, logical equivalence between items implies a unifactor structure, but a prerequisite relation between items only suggests strongly a unifactor structure. With respect to other two item hierarchical structures, no firm correspondence between factor and hierarchical structures was identified.

To study the question "Can the factor structure of a test tell us something about the hierarchical structure?", a set of bivalued items having a unifactor structure and a set of items having a bifactor structure were located to determine whether the unifactor items engendered a test hierarchy with fewer items and thus more inter-item prerequisite relations than the test hierarchy for the bifactor items. Data for such groups of test items were used in the Bock and Lieberman study (1970) and were later reanalyzed by Christoffersson (1975). One group of test items consisted of five highly homogeneous Figure Classification items with KR_{20}= .880 from the Law School Admission Test (LSAT). From a conceptual analysis of the items, no inter-item prerequisite relations were hypothesized. The other group of test items consisted of five somewhat more heterogeneous Debate items with KR_{20}=.765 from the same LSAT. Again, inter-item prerequisite relations were not evident. All items were in the five alternative multiple choice format and were dichotomously scored. The sample for this data was 1000 subjects drawn from a larger sample of subjects applying for admission at various American universities. The data sample was stratified with respect to university and achievement level within universities. The raw, item response data are presented in Table 1. Positive features of this data as reported by Christoffersson (1975) include a unifactor structure for the first set of items and a bifactor structure for the second set of items. Other positive features are that the data involved few items, a large subject sample, and a systematic stratified sampling plan for response pattern selection. Such features are ideal for ordering theory, partly because hierarchy determination is very dependent on the number of patterns used in the analysis. The greater the sample size is in relation to the total number of possible item response patterns (2^n with n being the number of items), the greater the likelihood is that the response patterns actually involved in the test item hierarchy will have greater frequencies than the response patterns attributable to chance or error.

RESULTS AND DISCUSSION

Bock and Lieberman (1970) used a chi-square approximation for a likelihood-ratio statistic to determine that the Figure Classification items fitted a unifactor model (χ^2=21.28, d.f.=21, .40<p<.50), but that the Debate items did not (χ^2=31.59, d.f.=21, .05<p<.10). Christoffersson (1975) using a similar chi-square approximation also determined that the Figure Classification items fitted a unifactor model (χ^2=5.02, d.f.=5, .40<p<.50), and that the Debate items did not (χ^2=10.30, d.f.=5, .05<p<.10). Christoffersson suggested that the Debate items fitted much better a bifactor model (χ^2=.63, d.f.=1, .40<p<.45).

Table 1*

LSAT: Observed frequencies for response patterns

Pattern	Figure Classification		Pattern	Debate	
Item 1 2 3 4 5	frequency	cumulative frequency	Item 1 2 3 4 5	frequency	cumulative frequency
0 1 0 1 0	0	0	0 0 0 1 0	1	1
0 1 1 0 0	0	0	0 0 1 0 0	3	4
0 0 1 0 0	1	1	0 0 1 1 0	3	7
0 0 1 0 1	1	2	0 1 0 1 0	3	10
0 1 0 0 0	1	3	0 1 0 0 1	5	15
0 0 0 1 0	2	5	1 1 0 0 0	6	21
0 1 1 1 0	2	7	0 0 0 1 1	7	28
0 0 0 0 0	3	10	0 1 0 1 1	7	35
0 0 1 1 0	3	13	0 1 1 0 0	7	42
0 1 1 0 1	3	16	1 0 0 0 0	7	49
1 0 1 0 0	3	19	1 1 0 1 0	8	56
0 0 1 1 1	4	23	0 1 1 1 0	8	64
0 0 0 0 1	6	29	0 1 0 0 0	10	74
0 1 0 0 1	8	37	1 0 0 1 0	11	85
1 0 0 0 0	10	47	0 0 0 0 0	12	97
0 0 0 1 1	11	58	1 0 1 0 0	14	111
1 1 1 0 0	11	69	1 0 1 1 0	15	126
1 0 0 1 0	14	83	0 0 1 1 1	17	143
1 0 1 1 0	15	98	1 1 1 0 0	18	161
0 1 1 1 1	15	113	0 0 0 0 1	19	180
0 1 0 1 1	16	129	0 0 1 0 1	19	199
1 1 0 0 0	16	145	0 1 1 0 1	23	222
1 1 0 1 0	21	166	1 1 0 0 1	25	247
1 0 1 0 1	28	194	0 1 1 1 1	28	275
1 1 1 1 0	28	222	1 1 1 1 0	32	307
1 0 0 0 1	29	251	1 0 0 1 1	34	341
1 1 0 0 1	56	307	1 1 0 1 1	35	376
1 1 1 0 1	61	368	1 0 0 0 1	39	415
1 0 1 1 1	80	448	1 0 1 0 1	51	466
1 0 0 1 1	81	529	1 0 1 1 1	90	556
1 1 0 1 1	173	702	1 1 1 0 1	136	692
1 1 1 1 1	298	1000	1 1 1 1 1	308	1000

*Adapted from data table analyzed by Bock and Lieberman (1970).

The test data were analyzed with the Krus-Bart (1973) method of multidimensional scaling in which linear orders among items determined by sub-samples of subjects are conceived as factors or latent traits for the items. These results are reported in Table 2.

The Figure Classification item data were found to establish one prominent linear order accounting for 59.4% of the subjects. The Debate items were found to have one prominent linear order accounting for 56.5% of the subjects. Second linear orders accounted for 13.4% and 14.2% of the subjects for the two data sets respectively. The first four linear orders accounted for 90.5% and 85.3% of the subjects for the two data sets respectively. Though there are no statistical tests attached to the Krus-Bart scaling method, the analyses of the two data sets present a similar picture of one prominent first linear order and markedly less prominent secondary linear orders. However, the Krus-Bart method revealed no clear-cut uniorder or biorder structure to either data set.

The data sets were then analyzed with the composite method to determine item hierarchies through the determination of inter-item prerequisite relations. Item response patterns with the lowest frequencies whose combined frequencies were equal to or less than 3% of the sample, or 30 subjects, were deleted and inter-item prerequisite relations were sought; none were found. The procedure was repeated with the removal of 10% of the sample size, or 100 subjects, which determined cutoff frequencies of 15 and 12 for the two data sets respectively; again no prerequisite relations were found - i.e., no item was located to be prerequisite to any other item. Only after 129 subjects, or 12.9% of the sample, and 21 distinctly different response patterns, whose frequencies were less than 17, were deleted from analysis, was even one inter-item prerequisite relation located among the Figure Classification items. Similarly, only

Table 2

Latent structures for LSAT items using Krus-Bart method

Figure Classification Items	Factors I	II	III	IV	V	VI	VII	VIII	IX	X	Row Totals
1	562	221	145	112	0	30	22	0	0	0	1092
2	173	225	0	70	47	0	22	30	0	4	571
3	0	1	149	28	15	39	22	6	11	4	275
4	335	0	80	120	80	39	0	0	8	4	666
5	422	173	136	0	120	0	0	30	11	0	876
Column Totals	1492	620	510	330	246	108	66	66	30	12	3480
subject frequency	594	134	112	65	48	18	11	11	5	2	1000

Debate Items	Factors I	II	III	IV	V	VI	VII	VIII	IX	X	Row Totals
1	355	202	103	0	110	0	30	0	0	14	814
2	136	0	143	74	68	0	0	37	29	23	510
3	238	90	0	131	122	34	39	37	0	0	691
4	0	191	35	28	32	69	39	16	14	23	447
5	431	158	85	131	0	55	0	0	29	0	889
Column Totals	1160	641	366	364	332	158	108	90	82	60	3361
subject frequency	565	142	76	70	67	25	18	15	12	10	1000

after 126 subjects, or 12.6% of the sample, and 17 distinctly different response patterns, whose frequencies were less than 15, were deleted from analysis, was even one inter-item prerequisite relation located among the Debate items. Both sets of items indicated no hierarchical structure, thus confirming the hypothesis of no prerequisite relations derived from conceptual analysis of the items, because no inter-item prerequisite relations were indicated even when 10% of the most infrequently occurring response patterns were deleted. Further information on the hierarchy analysis could be generated, but those results would be pale in comparison to the main result that, in this case, bivalent items which had either a unifactor or a bifactor structure showed no hierarchical structure. If one assumes that absence of a factor structure implies absence of a hierarchical structure, then only from the ordering-theoretic scaling of the items which indicated that the two sets of items had no clear-cut factor structure, could one have expected no clear-cut hierarchy for either set of test items.

SUMMARY

The purpose of this study was to determine whether bivalent items complying to a unifactor structure produce an item hierarchy with fewer b ranches and more inter-item presequisite relations than bivalent items complying to a bifactor structure. No such difference was found; in fact, both sets of items were found to have no hierarchical structure. Highly homogeneous items testing for one latent trait were found to be logically independent of each other in terms of indicating no prerequisite interdependencies. Less homogeneous items fitting a two latent trait model were also logically independent of each other. Partly because we would expect many cases of multifactor tests engendering no hierarchical structures, this study indicates that the factor structure of a test does not necessarily indicate anything about the hierarchy of a test.

Certain unresolved issues emanate from this study: 1) is there a conceptual chasm between latent trait statistical test theory and ordering theory, between the quest for latent traits and the quest for item hierarchies? 2) is there some connection that will allow information from one approach to be converted into information within the other approach? It would be helpful if response patterns with their frequencies could be reported in any test results, to allow alternative item analyses. But much test information in education is in the form of completed analysis results. Thus, it would be helpful to educators if such test results could be converted into psychometric information relevant from another perspective. Such efforts at increasing the informativeness and multi-usefulness of test results would inexorably be tied to gains in the synthesis of psychometric methods.

REFERENCES

Airasian, P., Madaus, G., and Woods, E. Scaling attitude items: a comparison of scalogram analysis and ordering theory. Educational and Psychological Measurement, 1975, 35, 809-819.

Airasian, P. and Bart, W. Ordering theory: a new and useful measurement model. Educational Technology, 1973, 13, 56-60.

Airasian, P. and Bart, W. Validating a prioti instructional hierarchies. Journal of Educational Measurement, 1975, 12, 163-173.

Bart, W. and Airasian, P. The determination of the ordering among seven Piagetian tasks by an ordering-theoretic method. Journal of Educational Psychology, 1974, 66, 277-284.

Bart, W. and Krus, D. An ordering theoretic method to determine hierarchies among items. Educational and Psychological Measurement, 1973, 33, 291-300.

Bock, R. and Lieberman, M. Fitting a response model for n dichotomously scored items. Psychometrika, 1970, 35, 179-197.

Christoffersson, A. Factor analysis of dichotomized variables. Psychometrika, 1975, 40, 5-32.

Gagnè R. Learning hierarchies. Educational Psychologist, 1968, 6, 1-9.

Guilford, J. The structure of intellect. Psychological Bulletin, 1956, 53, 267-293.

Krus, D. and Bart, W. An ordering-theoretic method of multidimensional scaling of items. Educational and Psychological Measurement, 1974, 34, 525-535.

Krus, D., and Bart, W., and Airasian, P. Ordering theory and methods. Los Angeles: Theta, 1975.

Lawley, D. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 1943, 61, 273-287.

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

Lord F. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-75.

Lord, F. & Novick, M. Statistical theories of mental test scores. Reading Massachusetts: Addison-Wesley, 1968.

Piaget, J. Logic and psychology. London: Routledge, 1950.

Samejima, F. Estimating latent ability using a pattern of graded scores. Psychometrika Monograph Supplement No. 17. William Byrd Press, 1969.