DOCUMENT RESUME

ED 135 145 EC 092 973

AUTHOR Hofmann, Richard J.
TITLE Illustrative Examples of the Development and Interpretation of Hierarchial Tests in the Field of Learning Disabilities.
SPONS AGENCY Miami Univ. Alumni Association, Oxford, Ohio. Faculty Development Fund.
PUB DATE Aug 76
NOTE 23p.; Paper presented at the International Scientific Conference of IFLD (3rd, Montreal, Canada, August 9-13, 1976); Best Available Copy

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Developmental Tasks; *Learning Disabilities; Prediction; *Test Construction; *Testing
IDENTIFIERS *Hierarchical Analysis

ABSTRACT

The author discusses the use of hierarchical tests with learning disabled (LD) children and presents four examples to explain basic characteristics of this type of test. It is explained that a hierarchical measurement provides two associated scores - a composite score and an error of prediction score. The examples are used to portray the use of a hierarchical test in analyzing cognitive processing in normal and LD children, the process of developing a hierarchical test to identify learning problems in young children, and methods of evaluating a standardized test for the properties of a hierarchical test. (CI)

Illustrative Examples of the Development and Interpretation of Hierarchical Tests in the Field of Learning Disabilities

Richard J. Hofmann[1,2]
Miami University
U.S.A.

Paper presented at the Third International Scientific
Conference of the International Federation
of Learning Disabilities - August, 1976
Montreal, Canada

[1] The support necessary for the preparation of this paper was provided by the Faculty Development Fund of the Miami University Alumni Association through the auspices of John Dolibois, Vice President of Alumni Affairs.

[2] Preparation of this paper was made while the author was serving as a Faculty Research Fellow at Miami University.

Illustrative Examples of the Development and Interpretation of Hierarchical Tests in the Field of Learning Disabilities*

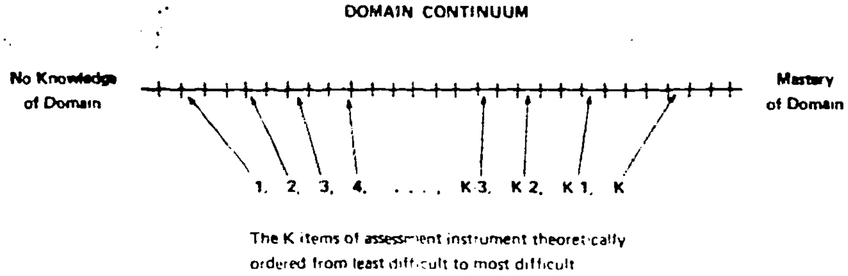
Richard J. Hofmann
Miami University

In the area of attitude measurement the Guttman (1950) scale has been recognized and used as a model for many years. In the area of cognitive assessment the Guttman scale has a great potential that has not yet been capitalized. To this end a new measurement model has been developed, a hierarchical test. When several hierarchical tests are intercorrelated they will have that elusive property referred to as "G". That is, when factor analyzed they usually will have hierarchical loadings on a factor, and if from the same content domain they will define just one factor. The objectives of this manuscript are to simply discuss hierarchical tests and to present several illustrative examples of their use with binary response data: i.e., right-wrong, yes-no, etc.

The hierarchical test has several characteristics not found in traditional assessment instruments. These characteristics are based upon the property that on such a test the individual item response patterns of a large majority of the responding individuals are highly predictable and orderly. With regard to the measurement, identification, and understanding of certain types of learning disabilities, the hierarchical test may provide insights and new measurement approaches. It is possible that curriculum may be developed hierarchically in terms of subordinate knowledge as determined normatively by the item content of a test.

*This term is used in a very general fashion throughout this manuscript.

Figure 1. Relationship of hierarchical test items to items of content domain.


DOMAIN CONTINUUM
No Knowledge of Domain
Mastery of Domain
1. 2. 3. 4. . . . . K-3. K-2. K-1. K
The K items of assessment instrument theoretically ordered from least difficult to most difficult

The nature of hierarchical tests. If the items defining an assessment instrument are similar to those portrayed by Figure 1 then they are all measuring different levels of a single content domain and might be said to be defining a hierarchical test. Such a test is composed of non-redundant binary items measuring different levels of mastery within the same content domain. That is, the item difficulty varies as opposed to a traditional homogeneous test assumed to be composed completely of items all measuring the same level of mastery, a classic mastery test. One particular compelling property of a hierarchical test is that the total score (number of correct or positive responses) that one obtains may be interpreted within a criterion-referenced framework or meaningful "product framework" from the view that a majority of the response patterns will be orderly and well behaved. To the extent that the test is a "perfect hierarchical test" the total score will actually define without error the response pattern (correct and incorrect item responses to each item) or processing, for each and every response. For example, if we assume that we have a perfect six item hierarchical test, a six item assessment instrument, and some individual obtains a socre of four correct responses, then this individual responded correctly to the four easiest items. If we understand both the content and construct validity of the domain associated with the test items we can make interpretations of an individual's score directly in terms of specified performance standards.

Clearly one will not usually have a perfect hierarchical test. For some respondents we will have less than perfect prediction of their item response pattern. This is to be expected just by chance, however for some respondents we will have extremely poor accuracy in predicting their response patterns. Whereas a typical test would have one score, the composite score, a hierarchical test will have two scores: a composite score determined by the summation of correct responses; and error score determined as the number of responses incorrectly predicted for an individual when attempting to predict their item response pattern given their composite score or the degree of "composite confusion". To the extent that an individual has a large error score, his item response pattern and probably his cognitive processing would be normatively atypical.

Example 1: Computational Example of Reproducibility, Error (Composite Confusion)

Assume that a group of k tasks (items, responses and so on) have been obtained. These items are ordered on the basis of empirical observation from easiest to hardest. Assuming the k items to be associated with a perfect Guttman scale and assuming the items to be ordered from easiest, item 1, to hardest; item k then no subject with j correct responses will respond to any item m where m is more difficult than j. Following a similar logic this same subject will respond correctly to any item i where i is as easy or easier than j.

Within the framework of binary responses 1 is an affirmative response and 0 is a negative response. With a perfect Guttman scale one would not anticipate any pattern of the nature (01) for a two item easy-hard sequencing. Such a pattern would be empirically illogical and disconfirming of the empirically based easy-hard sequencing of the two items. Generalizing this concept to k items an index called reproducibility has been developed to quantify how well the data conform to these assumptions. Reproducibility is just the proportion of responses correctly predicted for a group of n subjects on k tasks given their individual composite scores. The composite score for each individual is just the number of correct responses made by the individual as previously noted.

In Table 1 an artificial response matrix is presented. There are ten subjects and six items. The items have been ordered from left to right, difficult to easy. The subjects have been ordered from top to bottom, highest score to lowest score. Notice that all orderings are empirical or normative. The item difficulties from left to right are .3, .4, .5, .6, .7 and .8. Because there are tied composite scores the orderings within a score level are arbitrary. How well can the total response patterns of all subjects be reproduced? There are 10 subjects and six items, thus a total of 60 responses to predict. A total of 14 errors of prediction were made, thus 46 responses were correctly predicted or 77 percent accurracy. The reproducibility of the items is then .77. Alternatively 23 percent error occurred. This is a large percentage of error most likely it is more than one would tolerate.

Error might occur for any one of three reasons: (a) there may be a bad item(s) in the test such as item 4; (b) there may be several (never more than several) subjects for whom the item orderings are inapplicable; (c) the test is is just a poor test.

Table 1. Illustrative item response matrix.

	Item						Score	
Subject	2	6	4	1	5	3	Composite	Error[2]
D	$\underline{1}$[1]	1	$\underline{0}$	1	1	1	5	2
E	0	1	1	1	1	1	5	0
B	0	0	1	1	1	1	4	0
I	0	$\underline{1}$	$\underline{0}$	1	1	1	4	2
H	0	0	1	1	1	1	4	0
A	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{0}$	$\underline{0}$	$\underline{0}$	3	6
J	0	0	$\underline{1}$	1	1	$\underline{0}$	3	2
G	$\underline{1}$	0	0	0	$\underline{0}$	1	2	2
F	0	0	0	0	1	1	2	0
C	0	0	0	0	0	1	1	0
Number of "1" responses	3	4	5	6	7	8	33	
Prediction Errors	3	2	4	1	2	2		14

[1] underscored responses are errors of prediction

[2] the term error refers to error of prediction

It is especially interesting to note that the reproducibility is nothing more than the compliment of the average precent error for each subject. Inasmuch as the model of a hierarchical test is usually normatively based it provides an opportunity to identify subjects who are normatively atypical. To wit, subject A in Table 1 obtained a raw score of 3 on the test. The average score is 3, thus one might not be concerned about a raw score of 3 but note that subject A obtained an error score of 6 when the average error score is 1.4. Clearly subject A is normatively atypical on this test and worthy of additional investigation. Such additional investigation would of course be initially based upon the content domain of the test. Certainly to the extent that the test has a low reproducibility and a number of subjects with errors of prediction such an interpretation is not warranted as the test does not conform well enough to the hierarchical model. The virtue of a hierarchical test is minimal composite confusion or errors of prediction. It seems that all tests purporting to use a composite score for any type of decision or regression analysis should be free of or have at least minimal composite confusion.

Example 2: Comparative Analysis of Normal and Learning Disabled Errors of Prediction on a Hierarchical Test of Seriation by Sense Modality--Cognitive Processing.

In a recent unpublished (yet to be completed) study 18 ten year old children from learning disability classes were compared to 74 children from normal classes (seven to ten years of age) with regard to cognitive processing used in conjunction with various sense modalities. Sixteen tasks were devised such that the children were required to sort a group of objects from smooth to rough (tactile), a second group from light to heavy (kinesthetic), a third group from white to dark (visual), and a fourth group of objects from short to long. Such tasks are properly referred to as seriation tasks in the sense of Inhelder and Piaget (1964). After each initial sorting the children were given three additional objects logically associated with the sorted group and asked to insert these additicnal objects into their proper positions within the sorted group. These tasks were all logically equivalent but because of the degree of sense discrimination required and the various sense modalities used they varied in difficulty.

The reproducibility of the 16 tasks for the 73 children from the normal classrooms was .81. Similarly the reproducibility for the learning disabled children for the same hierarchical test was .81! Clearly the tasks defined a hierarchical test of modality seriation. Of primary importance were the distributions of errors of prediction for the normal and learning disabled children. If the distributions were significantly different from each other this would suggest that the cognitive processing of the learning disabled children was different from the cognitive processing of the normal children used to normatively establish the hierarchical test.

On this test there was a possible maximum of 16 errors. The errors on any hierarchical test will always occur in multiples of two thus the range of error pairs on this test was from 0 to 8. The error distributions for this test are reported in Table 2. Eliminating the last column of Table 2 a chi square test of independence was conducted to determine if the frequencies in any of the paired error categories tended to occur with greater or less probability for either the normal or learning disabled children. The resulting chi square [$\chi^2(4)=3.70$, $p>.05$] was not significant thereby suggesting that the frequency

Table 2. Score error frequency distributions from hierarchical test of sense modality seriation.

Child Type	Score Error Pairs[1]				
	0	2	4	6	8[2]
Normal	7	37	21	8	1
Learning Disabled	3	5	8	2	0

[1] Score errors always occur in multiples of two, thus the distribution is described in multiples of two.

[2] Because 1 subject falls in this category it was ommited from analysis.

of errors of prediction occurred independently of the categories learning disabled and normal.

It was concluded on the basis of these findings that the cognitive processing of the two groups was the same. Note that reference was never made to the number of correct responses on the test as this would have addressed a different question, level of knowledge. Such a question is ideally addressed by a hierarchical test but simply was not part of the study just described. (Interestingly enough there was a difference with regard to level of knowledge).

The substantive implications of these findings are beyond the scope of this paper, however it should be clear from this example that one particularly compelling use of a properly defined hierarchical test is the comparative analysis of several well defined groups such as normal and learning disabled children.

Example 3: Constructing a Normatively-Based Hierarchical Test for the Early Identification of Learning Problems.

Assume that one has a number of items that purportedly represent what is felt to be a content domain. Furthermore, one would like very much to have the items define a test having the previously mentioned features, but there is some question about which items do or do not really belong to the content domain, or worse yet (Example 4), one thinks that he knows which items belong to the domain and how they define the content domain---but this knowledge is in error. Until very recently there seemed to be no way of knowing which items did or did not belong and there seemed to be no way of knowing whether or not one's *a priori* assumptions about the content domain were in error.

Recently a rather large school district developed an extensive pre-school inventory (50 items). It was clear to them that the instrument was not a single factor instrument. To use a score based upon the total number of correct responses would have resulted in a test with low validity. Initially one might assume that a factor analysis would aid in defining more valid subtests. Unfortunately each item was either correct or incorrect, binary, thus a factor analysis of the data would have resulted in what is typically referred to as difficulty factors, e.g., easy items cluster together, difficult items cluster together, extreme items cluster together and so on. The items will cluster together not because of their content but because of their difficulty, thus the subtests might or might not be more valid than the total test. This is an especially severe problem as one of the major objectives of the instrument was to serve as a screening device to identify children with potential learning problems, i.e., early identification. When properly developed such instruments use as a criterion some later measure of learning. If it is known that there are validity problems initially it seems senseless to conduct a longitudinal study.

The most logical approach seemed to be one of identifying hierarchical subtests. To this end a new multivariate procedure has been developed, Multiple Hierarchical Analysis (Hofmann, Note 1). The multivariate model will not be discussed in this manuscript let it suffice that the model identifies latent Guttman scales in the data and then determines the best real data approximations to these latent scales. The real data approximations are just hierarchical tests! However, these hierarchical tests are not composed of all of the items in the test battery rather, those items that appear not to belong to the content domain associated with the hierarchical test are excluded. As a result several hierarchical subtests are derived from the original test battery. These subtests may have certain items in common and there may be certain items excluded from all of the subtests. Typically the scores on one such hierarchical test will be correlated with the scores on another hierarchical test derived from the same battery of items. This is not a severe problem as the correlations tend not to be high. Thus it is possible to take a group of logically homogeneous [illegible] using the Multiple Hierarchical Analysis model "cull-out" those [illegible] are not part of a hierarchical test domain, the remaining items forming, normatively, a hierarchical test or a group of hierarchical subtests. Such cleaned up hierarchical tests might properly be referred to as normatively-based hierarchical tests. When such a test is determined from a single content area it might be referred to as a <u>normatively based criterion referenced test</u>.

Using the multiple hierarchical analysis model in conjunction with the responses of 1236 children ages 4.5 to approximately 6.0 to the 50 items, ten hierarchical subtests were identified. Of considerable importance were the first two hierarchical subtests which were composed of 41 of the original 50 items. The first subtest is composed of 31 items while the second subtest is composed of an additional ten items not on the first subtest.

The 31 item hierarchical subtest has an astonishingly high Kuder-Richardson 20 reliability (Ferguson, 1971) of .97. The reproducibility of this subtest is .88. The error of prediction distribution is reported in Table 3. The second hierarchical subtest is somewhat of a disappointment with a reliability of .484, when corrected to a reliability equivalent to a 31 item test the reliability becomes .74, and a reproducibility of .81. These figures are not necessarily poor but relative to the 31 item hierarchical test they leave much to be desired.

Because the first subtest has such fine properties serious consideration should be given to the additional testing or retesting of those children with

Table 3. Frequency of errors of prediction distribution for 1236 children on 31 item and 10 item hierarchical subtests.

	Score Error Pairs									
	0	2	4	6	8	12	14	16	18	Total
31 Items	104	438	397	188	66	23	17	2	1	1236
10 Items	335	612	276	14	0	0	--	--	--	1236

error scores of 6, 7 and 8. Clearly these errors suggest that their response patterns are drastically different from the response patterns of the 1216 children. Presently we are preparing to gather the results of the first year of achievement for these 1236 children. Certainly a linear regression is planned for use with the well behaved composite scores which range from a low of approximately 5 to [illegible] approximately 30. Also being planned as an alternative to linear regression is an expectancy table approach. Although the achievement data have not yet been obtained it may be informative to illustrate how an expectancy table is developed as an alternative to linear regression.

Assume as an independent variable the error of prediction score. As a dependent variable one might consider ranges of achievement as opposed to specific scores or subjective judgements of teachers. Assume that the dependent variable is a teacher's subjective judgement of a child's achievement. A linear regression approach would most likely utilize the composite score on the test and a numerical index of achievement. The cell entries are hypothetical but in practice they would represent the frequency of children obtaining the particular error of prediction associated with the row and the teacher rating associated with the column. Dividing any row entry by a row total will define the probability of a child who obtained the row error of prediction receiving the column rating.

Implicit in this table is a major hypothesis of this paper---mainly that children who are normatively atypical in their performance on a particular cognitive test will be normatively atypical in their school performance or be learning disabled. An error score of zero is only indicative of the lack of confusion in an individual's composite score. Table 4 simply implies that

Table 4. Illustrative example of an expectancy table.

		Poor	Below Average	Average	Above Average	Superior	Row Total
	0	21	21	21	21	20	104
	2	88	88	88	87	88	438
Score	4	79	80	80	79	79	397
Error	6	45	42	39	30	32	188
Pairs	8	33	20	10	0	0	66
	10	18	2	2	1	0	23
	12 or more	18	1	1	0	0	20

a child with an error score of zero has an equal probability of being categorized into any one of the five teacher rating categoreis. This follows logically as level of achievement should ordinarily predict teacher's ratings and it should occur independently of error prediction. (Strictly speaking there will be fewer errors of predictions associated with very low and very high composite scores.) If the total table were converted to probabilities, based on row totals, it would be found that the greatest probability of being rated poor is associated with a high error of prediction. Alternatively within a linear regression framework these same 20 individuals would be the ones for whom the greatest errors of linear prediction would occur.

In time it is hoped that the adequacy of this prediction model will be established. Clearly the accurracy of this model from a learning disability framework is dependent upon the content domain of the test and its logical relationship to achievement.

Example 4: Can We Make A Silk Purse From a Sow's Ear?

As previously noted the use of a composite score implicitly assumes a hierarchical test. How well is this assumption met with real-life standardized data? A subtest of a prominent American standardized test was evaluated with regard to certain properties of a hierarchical test.

Utilizing the response patterns of 83 second grade children (a total population from one school), five of whom were labled as learning disabled, a reproducibility of .72 was obtained for the 32 items. The consequences of such a low reproducbiliity are best characterized by the error frequencies in Table 5.

Table 5. Frequencies of errors of prediction on a prominent 35 item standardized subtest.

	Score Error Pairs								
	0	2	4	6	8	10	12	14	Total
Frequency	0	3	3	5	20	23	20	9	83

Had one attempted to predict the response patterns for these children at least 9 of the children would have missed seven items that were predicted as being correct for them and they would have responded correctly to seven items that were predicted as being correct for them, 14 errors for each of the nine children. In the previous example only three percent of the sample had 12 or more errors of prediction whereas on this subtest 35 percent had 12 or more errors of prediction, yet this instrument has only four more items. On the average it is possible to predict 72 percent of any child's response pattern. The degree of composite confusion is immense on this particular instrument. Although the validity must be low for an instrument with such great composite confusion ironically the Kuder-Richardson 20 reliabiltiy estimate for this instrument is .78.

The publishers claim that this subtest measures four different components. The component subtests were analyzed with the following reproducibilities .74, .78, .79 and .73 with corrected relaibilities of .80, .71, .80 and .78 respectively. These subtests show little improvement over the original subtest as the two largest reproducibilities are associated with subtests of seven and five items respectively.

In an attempt to identify latent scales the Multiple Hierarchical Analysis model was applied to the data. The analysis defined 13 hierarchical subtests with ten subtests showing a greater reproducibility than the four subtests defined by the publishers. The subscale item content ranged from a low of two items to a high of seven items. The subscales are summarized in Table 6.

Although one must be skeptical regarding the use of a two or three item subtest it does not seem at all unreasonable to use a 6, 7, 8 or 9 item subtest especially given the large percentage of low errors of prediction.

Table 6. Summary table for the 13 normatively-based subtests determined from a prominent American standardized subtest.

Subtest	Score Error Pair Frequency						
	0	2	4	6	REP[2]	Number of Items	Normalized[3] Reliabilit
1	36	45	2	0	.83	7	.86
2	47	35	1	-[1]	.82	5	.90
3	22	38	22	1	.77	9	.90
4	38	41	4	0	.80	6	.89
5	61	22	0	-	.87	4	.90
6	36	42	5	0	.79	6	.91
7	35	45	3	0	.82	7	.89
8	63	19	1	-	.87	4	.83
9	77	6	0	-	.92	2	.91
10	33	47	3	-	.75	5	.87
11	65	18	0	-	.86	3	.44
12	71	12	0	-	.90	3	.73
13	60	23	0	-	.86	4	.88

[1] Dashes are used when the number of errors is not possible.

[2] REP refers to subtest reproducibility.

[3] Normalized reliability refers to reliabilities corrected by the Spearman-Brown prophesy (Ferguson, 1971) to a magnitude that would be associated with a 35 item test.

When utilizing subtests composed of a restricted number of items one must keep in mind the consequences of an associated restricted variance for the composite scores if they are used in a linear regression or any parametric analysis. Although it has not been mentioned thus far because it has not been a problem in the examples, it is possible that regardless of the number of items in a subscale there may be a restricted variance with the composite scores if the subtest items are homogeneous with regard to difficulty. Contrary to much traditional psychometric literature it is desirable to have heterogeneous item difficulty on a test if it is to have the properties of a hierarchical test.

After all of the efforts to obtain the subtests, three of the learning disabled children were not but two were associated with extreme errors of prediction on certain subtests. Most likely there were not enough learning disabled children identified in the sample to allow a reasonable analysis of the errors of prediction.

Finally in response to the subtitle of this section---maybe, but it will require effort.

Summary

In this manuscript a new type of measurement model was discussed, the hierarchical test. Unlike traditional tests which result in a composite score the hierarchical test was shown to have two associated scores; a composite score and an error of prediction score. Utilizing an artificial data set as a first illustrative example most of the basic characteristics of a hierarchical test were identified and their computations were discussed verablly. Three additional real-life examples were presented. Under the assumption that most researchers have a working knowledge of composite scores and their research uses the discussion of such scores [illegible] minimal. The discussion within the illustrative examples emphasized [illegible] importance and use of the error or prediction scores. The examples demonstrated: the use of a hierarchical test in the comparative analysis of certain cognitive processing of learning disabled and normal children; a method establishing a hierarchical test for the early identification of children with learning problems; how one might go about testing a standardized test for the properties of a hierarchical test.

Space does not permit the extensive use of illustrations however there are several additional uses of hierarchical tests worthy of brief mention. The items of such a test might specify a learning hierarchy in the sense of Gagné, facilitating the assessment of an individual's position within the specified hierarchy. In specifying learning hierarchies the items of such an instrument would also allow one to utilize chaining concepts establishing item level empirical prerequisites for learning within the test domain; possibly facilitating an empirically based aptitude interaction model. Alternatively the items of such a test would facilitate the advancement of the state of knowledge with regard to task analysis. In addition to all of this the composite score of a hierarchical test may be interpreted with considerable validity within a traditional normative framework.

Finally it is possible that the composite scores of a hierarchical test would be predictive perhaps using expectancy tables or linear regressions, of levels of achievement while error scores might be predicative of specific categories of learning disabilities depending upon the content domain. The hierarchical test approach to the identification of learning disabilities may provide a means to better classification, better understanding and to improved program development for learning disabled children. Finally the proposed measurement approach may provide increased understanding of the processing characteristics of various types of learning disabled children as well as providing information for the planning of individual education programs. Testing of the model is just beginning.

Reference Note

Hofmann, R.J. *Multiple hierarchical analysis*. (Manuscript in preparation for Spring, 1977).

References

Ferguson, G. *Statistical analysis in psychology and education*. New York: McGraw-Hill, 1971.

Guttman, L. The basis for scalogram analysis. In S. Stouffer, et al., *Measurement and prediction*. Princeton: Princeton University Press, 1970.

Inhelder, B. and Piaget, J. *The early growth of logic in the child*. New York: W.W. Norton and Company, Inc., 1964.