DOCUMENT RESUME

ED 219 418 TM 820 430

AUTHOR Yap, Kim Onn
TITLE Use of Item Dimensionality to Reduce Test Burden in Title I Projects.
PUB DATE Mar 82
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Bilingual Students; Cluster Analysis; Compensatory Education; *Item Analysis; Language Tests; *Multidimensional Scaling; Primary Education; Program Evaluation; *Scoring; *Test Items; Test Use
IDENTIFIERS Elementary Secondary Education Act Title I; Individual Differences Scaling (INDSCAL); *Item Dimensionality

ABSTRACT

A few multi-dimensional items were found to be more efficient than a larger number of uni-dimensional items in tests for participant selection and evaluation of such compensatory educational programs as Title I. A multi-dimensional scaling technique is described which derives scores for each item on three dimensions, and allows multiple scores to be derived from the same set of test items. A 35-item English language test was administered to first- and second-grade bilingual students and an inter-item correlation matrix was computed. In an analysis by the INDSCAL program, a three dimensional configuration of pronoun use, object identification and word-endings pronunciation was found. The scale scores, intercorrelations among nine score variables, and degrees of item saliences on the dimensions were calculated. Five items were classified as multi-dimensional (appearing in all three clusters of items based on dimensions) and seven were identified as uni-dimensional. The possibility of using item dimensionality to reduce the test burden in program evaluation is discussed with suggestions for improving the present procedure regarding scoring algorithms and salience measures. (Author/CM)

Use of Item Dimensionality to Reduce Test Burden in Title I Projects

Kim Onn Yap
Northwest Regional Educational Laboratory

A paper presented at the annual meeting of the American Educational Research Association, New York, March 19-23, 1982

Use of Item Dimensionality to Reduce Test Burden in Title I Projects

Kim Onn Yap

Northwest Regional Educational Laboratory

This paper examines the potential of using multi-dimensional test items to reduce test burden in education projects. Reducing test burden appears particularly important in such compensatory education projects as Title I where testing is conducted not only for purposes of measuring student achievement growth but also for selecting students to participate in the treatment. Quite often separate tests are used for measuring program effects and for selecting participants.

Classical test theories (Gulliksen, 1950; Guilford, 1954) have typically assumed uni-dimensionality of test items. More recently, item response theories (Rasch, 1960; Wright, 1979) have made it possible for item calibration to be less dependent on specific samples of tested subjects. The newer approach has, however, continued the tradition of uni-dimensionality assumptions.

The tenability of the uni-dimensionality assumption has been questioned by various researchers (e.g., Samejima, 1974; Sympson, 1978). In some cases the limitations of that assumption are obvious. Problem-solving items in a mathematics test, for instance, undoubtedly require reading comprehension ability in addition to computation skills. In such cases, a multi-dimensional latent space is obviously more reasonable.

Unlike classical test theories and item response theories, the present paper assumes that some test items have a multi- rather than uni-dimensional structure. In other words, an item measures several triats in varying degrees rather than a single trait or part of a single trait as is commonly assumed. For instance, an item measuring reading comprehension also measure word attack, word recognition and word comprehension skills to varying degrees. Thus multiple indices of reading competencies can be derived from a set of reading comprehension items. A two-dimensional item space is illustrated below:

Figure 1 here

The item space suggests that item A is more a measure of comprehension than vocabulary. Item B is more a measure of vocabulary than comprehension. Item C is as much a measure of comprehension as it is a measure of vocabulary. Yet all three items could have been included in a comprehension or a vocabulary subtest.

Procedure

Multi-dimensional scaling techniques (Carroll & Chang, 1970; Subkoviak, 1975; Kruskal & Wish, 1978) can be used to construct a multi-dimensional latent space for a test. The configuration can then be used to derive scale scores for each item on each of the dimensions represented in the test space. Stress values or other indices of goodness of fit (Kruskal & Wish, 1978) are used to determine the number of dimensions necessary for arriving at an adequate configuration.

Since the approach allows for multiple scores to be derived from the same set of test items, fewer items are required to obtain subscores (e.g., vocabulary and comprehension). Testing time can be reduced and testing can be done in a most cost-effective way.

The procedure was applied to a set of test data obtained from a compensatory educationa project. A 35-item English language test was given to 100 first- and 110 second-graders with a bilingual family background. The test questions were verbally presented to the students on a one-to-one basis. For example, while pointing to a picture of money the teacher may ask, "what is this?" To receive credit for their answers the students were to respond to the questions in complete sentences (e.g., That is money.) A inter-item correlation matrix was first computed from the test data, separately for first- and second-graders. These item intercorrelations were then used as proximity measures and were analyzed by INDSCAL (Carroll & Chang, 1970).

Test Configuration

Results of the analysis suggested that a three-dimensional configuration appeared to adequately represent the latent space of the test. Specifically, the configuration accounted for over 71 percent of the variance (multiple R=.845). The three major dimensions may be labeled as follows:

Dimension 1: Proper use of pronouns

Dimension 2: Correct identification of objects

Dimension 3: Proper pronunciation of word endings

Items having the highest saliences (generally .2 or above) on each of the dimensions were identified. Twelve itmes were found to have the highest saliences on proper use of pronouns, 13 on correct identification of objects and 10 on proper pronunciation of word endings.

Multi-Dimensional Items

A close examination of the saliences and graphical representation of the test items provided by INDSCAL suggested three clusters of items. Note that although these item clusters are formed on the basis of the three major test dimensions described earlier, they do not correspond on a one-to-one basis with the three dimensions. Items which appeared in all three clusters were identified as multi-dimensional items. Items which were found in only one of the three clusters were considered uni-dimensional items. Based on this criterion, five items were classified as multi-dimensional and seven as uni-dimensional items.

Traditional psychometric qualities of these items are shown in Table 1 and 2. The indices suggested that, by and large, the multi-dimensional items were superior to the uni-dimensional items in terms of p value and item-test correlation. Specifically, the multi-dimensional items had p values more closely clustered around the .6 and .7 region. They also had higher item-test correlations. This was true with respect to both the first and second grade data. Futhermore, if subtests were formed using the five multi-dimensional and seven uni-dimensional items, respectively, the multi-dimensional subtest would have a higher correlation with the total test than the uni-dimensional subtest. (See Table 3). The difference was statistically significant ($p < .01$) in the case of the second-grade sample and the combined sample.

Tables 1,2 and 3 here

Saliences provided by INDSCAL for each of the items were used to derive subscores for the first- and second-graders. The procedure is illustrated as follows:

$$PS = T$$

Where P is a matrix of item difficulty (p = 1 or 0); S is a matrix of saliences provided by INDSCAL; and T is a matrix of derived scale scores. The size of each of the matrices is determined by the number of test items, the number of major dimensions of the test configuration, and/or the number of subjects.

A computer program was written to derive scale scores and to provide other indices to examine the usefulness of the scale scores. The algorithm provided a total of nine score variables:

Variable 1: Scale score for proper use of pronouns based on 12 items having highest saliences on the dimension

Variable 2: Scale score for correct identification of objects based on 13 items having highest saliences on the dimension

Variable 3: Scale score for proper pronunciation of word endings based on 10 items having highest saliences on the dimension

Variable 4: Raw score for proper use of pronouns based on 12 items having highest saliences on the dimension

Variable 5: Raw score for correct identification of objects based on 13 items having highest saliences on the dimension

Variable 6: Raw score for proper pronunciation of word endings based on 10 items having highest saliences on the dimension

Variable 7: Scale score for proper use of pronouns based on five multi-dimensional items

Variable 8: Scale score for correct identification of objects based on five multi-dimensional items

Variable 9: Scale score for proper pronunciation of word endings based on five multi-dimensional items

The computer program also provided intercorrelations among the nine variables. The correlation coefficients are s ..n in Table 4 and 5.

Tables 4 and 5 here

As expected, the results showed that there were high correlations between the derived scale scores and raw scores based on items having high saliences on the respective dimensions. The correlation coefficients ranging from .88 to .96 no doubt reflected the goodness of fit between the test configuration provided by INDSCAL and the raw data.

Note that since the configuration can be rotated to reverse the sign of the saliences, only the magnitudes of the correlation coefficients are of interest. Their signs are of no relevance in this context.

The relationships between scale scores based on high salience items on a particular dimension and scale scores derived from the five multi-dimensional items were shown to be moderately high, the correlation coefficients ranging from .61 to .89. Similarly, there appeared to be substantial correlations between scale scores derived from the multi-dimensional items and raw scores based on high salience items on the respective dimensions. These correlations coefficients ranged from .44 to .80.

Discussion

Results of the present study suggest that item saliences generated by INDSCAL are potentially useful for identifying multi-dimensional items and for deriving various scale scores. More importantly, the results indicate that a few multi-dimensional items may in fact be more efficient than a larger number of uni-dimensional items for participant selection and program evaluation in education. Scale scores derived from item saliences provided a potentially useful measure of treatment effects. All this points to the possibility of using item dimensionality to reduce test burden in educational projects. The pay-off appears particularly

pertinent in compensatory education programs in which testing typically consumes an inordinately large amount of time which otherwise could be spent on instruction.

It should be noted that this paper discusses only some initial steps in the study of the use of item dimensionality to make testing more efficient in program planning and evaluation. The procedure used to identify effective multi-dimensional items is rudimentary and could undoubtedly be improved. Also, ways might be found to develop algorithms to select items which will maximize the relationships between the derived scale scores and the criterion--be it the total raw score or some external measure. Furthermore, it is not unlikely that saliences and configurations of test items are dependent on characteristics of those taking the test. This offers further opportunity of using item dimensionality to enhance the efficiency of testing by tailoring saliences to specific subgroups. Different saliences might be applied to the same items when groups of diverse characteristics are involved in the testing situation. Obviously, the validity and usefulness of these suppositions await future studies in this area.

References

Carroll, J.D. & Chang. J.J. Analysis of individual differences in multi-dimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. Psychometrika, 1970, 35, 283-319.

Guilford, J.P. Psychometric Methods. New York: McGraw-Hill, 1954.

Gulliksen, H. Theory of mental test. New York: Wiley & Sons, 1950.

Kruskal, J.B., & Wish, M. Multidimensional scaling. Beverly Hills: Sage Publications, 1978.

Rasch, G. Probabilistic models for some intelligence and attainment Tests. Copenhagen Denmark: Danmarks Paedogogiske Institute, 1960.

Samejima, F. Normal ogive model on the continucus response level in the multi-dimensional latent space. Psychometrika, 1974, 39, 111-121.

Subkoviak, M.J. The use of multi-dimensional scaling in educational research. Review of educational research, 1975, 45(3), 387-423.

Sympson, J.B. A model for testing with multi-dimensional items. In D.J. Weiss (Ed.) Proceedings of the 1977 computerized adaptive testing conference. University of Minnesota, 1978.

Wright, B.D., & Stone, M.H. Best test design-Rasch measurement. University of Chicago: MESA Press, 1979.

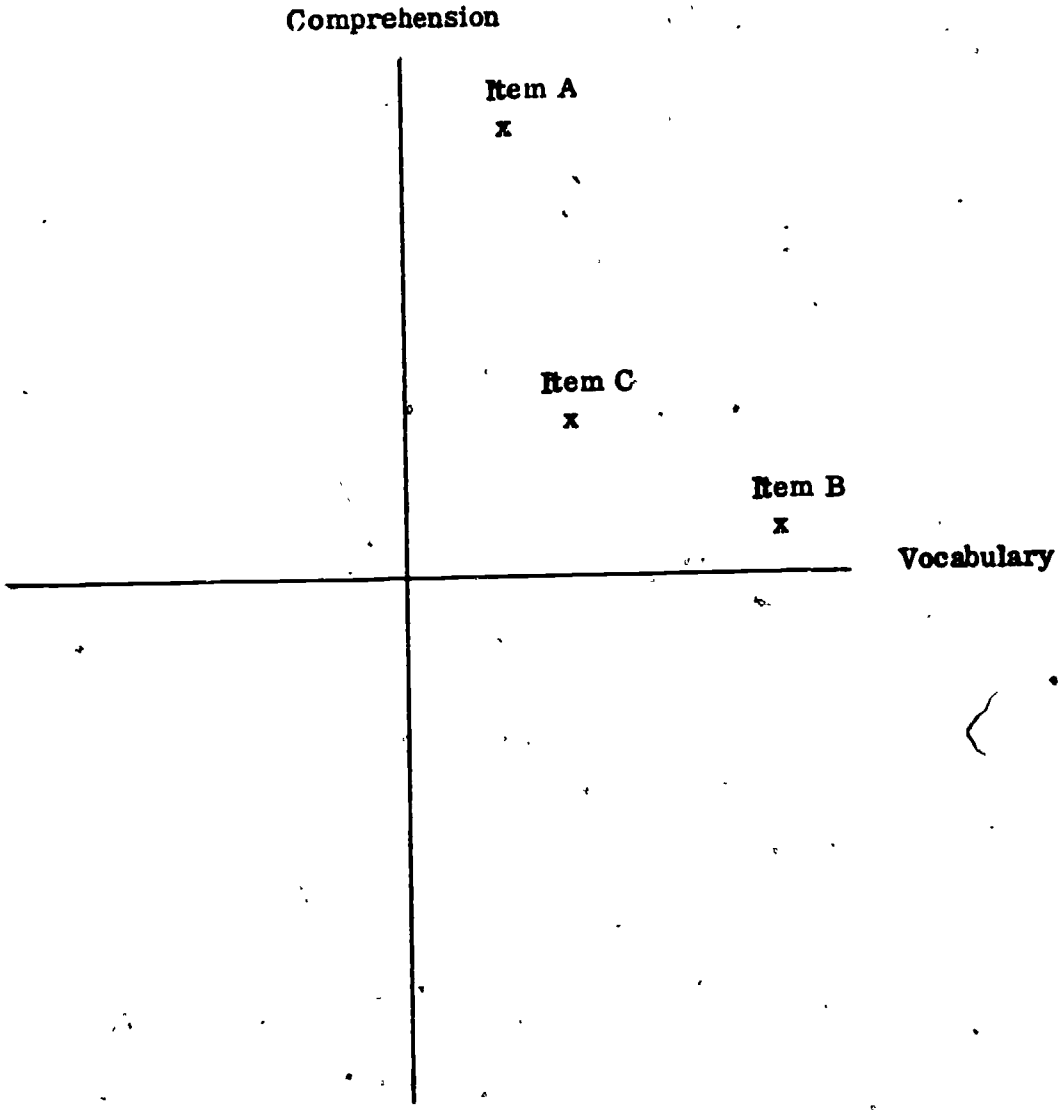

Figure 1. A Two-Dimensional Item Space.

Table 1

Psychometric Characteristics of Multi-Dimensional Items

Item	Sample			
	First Grade		Second Grade	
	P	rit	P	rit
1	.25	.40	.40	.50
2	.34	.47	.54	.59
3	.23	.45	.45	.56
4	.82	.49	.83	.58
5	.13	.43	.26	.51

Note. p = Item difficulty

rit = Item-test correlation

Table 2

Psychrometric Characteristics of Uni-Dimensional Items

Item	Sample			
	First Grade		Second Grade	
	P	rit	P	rit
1	.25	.35	.29	.37
2	.07	.27	.20	.40
3	.05	.15	.15	.47
4	.61	.55	.74	.45
5	.09	.39	.30	.37
6	.46	.13	.57	.27
7	.07	.23	.08	.49

Note. p = Item difficulty

rit = Item-test correlation

Table 3

Test-Subtest Correlations

Correlation	Sample		
	First Grade (N=100)	Second Grade (N=110)	Combined (N=210)
r_1	.758	.854	.829
r_2	.672	.742	.738
r_3	.446	.541	.539
$r_1 - r_2$	.086	.112**	.091**
	(t=1.51)	(t=3.01)	(t=3.10)

** $p < .01$

Note. r_1 is correlation between total test and subtest consisting of five multi-dimensional items.

r_2 is correlation between total test and subtest consisting of seven uni-dimensional items.

r_3 is correlation between the two subtests.

Table 4

Intercorrelations Among Score Variables

for First Graders (N=100)

	var. 1	var. 2	var. 3	var. 4	var. 5	var. 6	var. 7	var. 8
var. 2	-.90							
var. 3	.77	-.81						
var. 4	-.96	.88	-.67					
var. 5	-.67	.38	-.66	.61				
var. 6	-.73	.66	-.92	.59	.48			
var. 7	.78	-.56	.61	-.66	-.36	-.70		
var. 8	-.40	.70	-.67	.38	.71	.46	-.19	
var. 9	.59	-.66	.87	-.51	-.51	-.78	.60	-.74

Note. Since the configuration can be rotated to reverse the sign of the saliences, only the magnitude of the correlation coefficient is of interest. The sign is irrelevant.

Table 5

Intercorrelations Among Score Variables

for Second Graders (N=110)

	var. 1	var. 2	var. 3	var. 4	var. 5	var. 6	var. 7	var. 8
var. 2	-.86							
var. 3	.73	-.81						
var. 4	-.96	.85	-.60					
var. 5	-.69	.94	-.74	.66				
var. 6	-.70	.66	-.94	.53	.55			
var. 7	.61	-.27	.45	-.44	-.14	-.56		
var. 8	-.35	.74	-.69	.34	.80	.49	.08	
var. 9	.49	-.67	.89	-.38	-.65	-.79	.35	.80

Note. Since the configuration can be rotated to reverse the sign of the saliences, only the magnitude of the correlation coefficient is of interest. The sign is irrelevant.