DOCUMENT RESUME

ED 103 469 TM 004 314

AUTHOR Ambrosino, Robert J.; And Others
TITLE Items and Instruction Evaluated Using Partitioning Procedures.
PUB DATE [Aug 74]
NOTE 16p.; Paper presented at the Annual Meeting of the American Psychological Association (82nd, New Orleans, Louisiana, August 1974)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Achievement Tests; Cluster Analysis; *Cluster Grouping; Goodness of Fit; *Grouping Procedures; Student Evaluation
IDENTIFIERS Hierarchical Clustering Analysis; Partitioning Procedures; Test Items

ABSTRACT

Two studies were undertaken to demonstrate the usefulness of partitioning procedures for studying test items. Achievement test items in five content areas of educational measurement were used as stimuli to be sorted by groups of students with varying levels of sophistication with the content, with the hypothesis that sorting by classes with greater sophistication would agree more with simulated target sortings than sortings by classes with less sophistication. These sortings were analyzed using partitioning procedures. Results from both studies indicated that degree of sophistication in measurement was overall a potent variable in the sorting. In addition, several misconceptions among the students concerning the content under study were revealed. It was noted that a moderate number of students enrolled in upper-level measurement courses demonstrated what amounted to errors in knowledge in their sortings. It was concluded that the partitioning procedures were useful for studying how items are perceived by students and for determining how students organize content. (Author)

ITEMS AND INSTRUCTION EVALUATED USING PARTITIONING PROCEDURES

Robert J. Ambrosino
Albany Regional Medical Program

Robert F. McMorris   Lorraine K. Noval
State University of New York
at Albany

Paper presented at the eighty-second annual convention of the American Psychological Association, New Orleans, August, 1974.

ITEMS AND INSTRUCTION EVALUATED USING PARTITIONING PROCEDURES

Abstract

Two studies were undertaken to demonstrate the usefulness of partitioning procedures for studying test items. Achievement test items in five content areas of educational measurement were used as stimuli to be sorted by groups of students with varying levels of sophistication with the content, with the hypothesis that sortings by classes with greater sophistication would agree more with simulated target sortings than sortings by classes with less sophistication. These sortings were analyzed using partitioning procedures. Results from both studies indicated that degree of sophistication in measurement was overall a potent variable in the sorting. In addition, several misconceptions among the students concerning the content under study were revealed. It was noted that a moderate number of students enrolled in upper-level measurement courses demonstrated what amounted to errors in knowledge in their sortings. It was concluded that the partitioning procedures were useful for studying how items are perceived by students and for determining how students organize content.

ITEMS AND INSTRUCTION EVALUATED USING PARTITIONING PROCEDURES

A variety of empirical approaches have been used for studying test items, ranging from item discrimination indices to latent trait methods. Typically, such approaches have relied on data from testees' answers to the items. The empirical data for the present study, however, are based on sortings of items, where each respondent clustered the items according to his own perceptions. These sortings were analyzed using partitioning procedures.

In this study achievement test items were used as stimuli to be sorted by groups of students having differing levels of sophistication with the content. It was hypothesized that the sortings by members of those classes with greater sophistication would agree more with simulated target sortings than would sortings by members of classes with less sophistication. Other intents of the study included evaluating the methodology as a procedure for studying how items are perceived by students and for determining how students organize content.

Method

Classes were used with varying levels of sophistication in measurement: high school students, undergraduates enrolled in an educational psychology course (EPSY 200) and in a pupil evaluation

course (EPSY 440), and graduate students enrolled in a pupil evaluation course (EPSY 540), an educational and psychological measurement course (EPSY 640), and in a more advanced measurement seminar (EPSY 744). For the first of two applications of the methodology, 151 students sorted the items.

Thirty multiple-choice achievement test items were used in the content areas of correlation, validity, reliability, and standard error of measurement. Item statistics available from previous testings indicated a moderate range of item difficulty and discrimination coefficients. Also, test items were initially selected with reference to Bloom's (1956) Taxonomy of Educational Objectives; four of the six major categories in the cognitive domain were represented in this selection.

Each student was supplied with an envelope containing test items on individual slips of paper, several paper clips, and a piece of paper on which the student was requested to indicate his basis for sorting. The student was instructed to sort the items into between three and nine categories and to indicate the basis for sorting that he used.

The sortings were analyzed using the methods of latent partition analysis (Wiley, 1967) and hierarchical clustering analysis (Hartigan, 1972; Johnson, 1967) in a manner similar to that described by Pruzek and Pfeiffer (1973) and Pruzek, Stegman and Pfeiffer

(1972). The reader is referred to the latter report for a discussion of the algebra involved in the clustering procedures used in this study. In essence, the goal was to measure the goodness of fit of any single partition of the 30 items to a fixed target partition, which corresponded to the investigators' hypothesis about the cue system which the sorters should most likely use in partitioning the items.

Manifest partitions for each class were analyzed with respect to an a priori target partition based on the content area covered by the item. The following item-content distribution was hypothesized: correlation - 9 items, validity - 7 items, reliability - 8 items, standard error of measurement - 4 items, and the relationship between validity and reliability - 2 items.

In this study the q_{st} statistic was used as a measure of goodness of fit for these data. A small value of this statistic, which has a range from 0 to 1, is taken as evidence that the target in question can reasonably be regarded as having been the model in some sense for an individual's manifest partition (Pruzek, Stegman and Pfeiffer, 1972, p. 7).

Results: Study A

Table 1 contains mean q_{st} values as well as standard deviations for each class, derived using the target partition based on item

content. Average q_{st} values are simply unweighted means computed across all class members, and are taken as a summary index of goodness of fit for each class.

Insert Table 1 about here

As can be seen, the average q_{st} values are highest for the two groups with least sophistication (Grade 11 and EPSY 200), and lowest for the EPSY 744, the most sophisticated group. Results were nearly identical for the three groups with some sophistication, i.e., EPSY 440, 540, and 640.

Table 2 includes results obtained from a comparison of inter-group q_{st} mean values using Duncan's New Multiple Range Test (Duncan, 1955; Cramer, 1956). The reader will note that significant differences were observed for eleven of the fifteen comparisons. There were no significant differences observed between $\bar{q}_{st}$s based on the Grade 11 and EPSY 200 data, and for comparisons made among $\bar{q}_{st}$s based on the EPSY 440, EPSY 540, and the EPSY 640 data.

Insert Table 2 about here

Many students responded to labels such as the term "reliability" in the item stems as an aid to sorting, as could be seen from the summaries of the sortings from self-reported replies to our request

for the basis for sorting, as well as from informal discussions with students who had completed the task. For items were such labels were not available, the sophistication of the group was a more potent variable in the sorting. Cues within the alternatives did not seem to have been important to the sorters.

Method: Study B

It was judged that results of the initial sortings were confounded by the presence of labels in the item stems, and the original set of items was revised to minimize such cueing by labels. Specifically, nineteen of the thirty items were revised, with intention to alter only the cues in the stems. Care was taken in this revision not to alter significantly the original item difficulty and discrimination levels and to maintain as closely as possible the original distribution of items as they related to Bloom's (1956) Taxonomy of Educational Objectives.

The process was replicated using a similar sample of students with varying levels of sophistication in measurement. Included in the 135 students were high school students and members of four out of the five university courses represented in Study A.

Results: Study B

The data were first analyzed using the a priori target specified for Study A. Table 3 contains means and standard deviations

of the q_{st} values for each class using this target. A detailed examination of individual partitions revealed in some classes that many sorters based their partitions on other than conventional sorting strategies, such as length of item stem, key answer, and the like. Partitions such as these were classified as outliers and were excluded from further analysis. Specifically, thirty-eight of the 135 sortings were classified as outliers.

The average q_{st}s contained in Table 3 are consistently higher across the various classes than those contained in Table 1. Since the composition of the classes and curricular content were fundamentally the same for each experiment, it was concluded that these differences were largely attributable to the cueing by labels discovered in the initial experiment. With the exception of the Grade 11 data the manifest partitions more nearly approximated the target partition as the sophistication of the class increased, i.e., for these data sophistication of the group appeared overall to be a potent variable in the sorting, even within the relatively homogeneous subset of classes.

Insert Table 3 about here

A comparison of intergroup q_{st} mean values was also made for these data. Table 4 includes results obtained from a comparison of intergroup $\bar{q}_{st}$s using Duncan's New Multiple Range Test. As can

be seen, significant differences were observed for seven of the ten comparisons. No significant differences were observed between $\bar{q}_{st}$s based on Grade 11 and EPSY 200 data, Grade 11 and EPSY 440 data, and EPSY 540 and EPSY 640 data.

Insert Table 4 about here

Data for each class were reanalyzed using derived targets generated by the initial clustering procedures, with the intent of further refining the results. The a priori target and the derived hierarchical clustering target were practically identical for each class, as were the mean q_{st} values derived using these targets. Results of these comparisons, which failed to improve the accuracy of initial results, are not included in this report.

A moderate number of sorters based their partitions on Bloom's (1956) Taxonomy of Educational Objectives: Cognitive Domain. A second target partition based on this classification scheme was constructed for analyzing this subset of data. Analysis of these data using this subsequent target resulted in an extremely poor fit, however, and further presentation of the findings is not included in this report.

Some misconceptions among the students concerning the content were suspected. Comments made by Ss relative to their sorting

strategies were reviewed and two-way contingency tables comparing the a priori target partition and the derived hierarchical clustering analysis partitions were constructed for each class with the purpose of detecting these errors.

To illustrate, for several groups an item based on expectancy tables was not associated with the validity items as expected. One might question then whether the concept of expectancy tables was adequately understood.

Two other misconceptions may be noted as illustrative. Several persons sorted items based on reliability into a category which they labeled correlation. It appears for these sorters that a limited conceptualization of the notion of reliability had been formed. In a similar fashion, others sorted items based on criterion-related validity into the same correlation category.

Summary and Discussion

Two studies were undertaken to demonstrate the usefulness of partitioning procedures for studying test items. Achievement test items in the content areas of correlation, validity, reliability and standard error of measurement were used as stimuli to be sorted by groups of students with varying levels of sophistication with the content, with the hypothesis that the sortings by members of those classes with greater sophistication would agree more with

simulated target sortings than would sortings by members of classes with less sophistication.

Findings from the first study indicated that sophistication of the group in measurement was a reasonably potent variable in the sorting. Subjects frequently responded to labels in the item stems as an aid to sorting, however, and thus failed to systematically apply their knowledge of the content to the sorting task.

The original set of items was revised to minimize such cueing by labels and the experiment was replicated using a similar sample of subjects. Results from the second study confirmed that degree of sophistication in measurement was overall a potent variable in the sorting.

Inspection of two-way contingency tables comparing the a priori target partition and the derived hierarchical clustering analysis partitions revealed several misconceptions among the students concerning the content under study. In this context, it was noted that a moderate number of students enrolled in upper-level measurement courses demonstrated what amounted to errors in knowledge in their sortings. Further, some content topics were apparently not well understood.

Numerous misconceptions involving the use of Bloom's (1956) Taxonomy of Educational Objectives: Cognitive Domain as a basis

for sorting the items were also noted. The majority of students who used this paradigm as a sorting strategy appeared to have mastered a knowledge of the category labels but failed to demonstrate an indepth understanding of the Taxonomy.

The procedures used in this study proved to be useful for studying how items are perceived by students and for determining how students organize content. Results such as those reported above seem to have value as a means of feedback to an instructor regarding the way in which his students perceive a given test and the corresponding course content. Such information has the potential for improving the teaching-learning process.

Further studies might include an investigation of the relationship between the goodness of fit of sorting data and selected organismic variables such as aptitude and achievement.

References

Bloom, B.S. (Ed.) Taxonomy of educational objectives: cognitive domain. New York: David McKay, 1956.

Cramer, C.Y. Extension of multiple range tests to group means with unequal numbers of replications. Biometrics, 1956, 12, 307-310.

Duncan, D.B. Multiple range and multiple F tests. Biometrics, 1955, 11, 1-42.

Hartigan, J.A. Direct clustering of a data matrix. Journal of the American Statistical Association, 1972, 67, 123-129.

Johnson, S.C. Hierarchical clustering schemes. Psychometrika, 1967, 32, 241-254.

Pruzek, R.M. and Pfeiffer, R.A. An illustration of an approach to analyzing partitioned data in the context of educational measurement. Paper presented at the annual convocation of the Northeastern Educational Research Association, Boston, November, 1972.

Pruzek, R.M.; Stegman, C.A. and Pfeiffer, R.A. On the analysis of partitioned data. Paper presented at the annual meetings of the American Educational Research Association, Chicago, March, 1972.

Wiley, D.E. Latent partition analysis. Psychometrika, 1967, 32, 183-192.

TABLE 1

Means and Standard Deviations for q_{st}s Derived Using Target Partition Based on Item Content: Study A

Class	N	$\bar{q}_{st}$	SD (q_{st})
Grade 11	39	.269	.068
EPSY 200	32	.280	.087
EPSY 440	16	.202	.072
EPSY 540	21	.200	.104
EPSY 640	26	.203	.068
EPSY 744	12	.096	.069

TABLE 3

Means and Standard Deviations for q_{st}s Derived Using Target Partition Based on Item Content: Study B

Class	N	$\bar{q}_{st}$	SD (q_{st})
Grade 11	19	.340	.052
EPSY 200	22	.367	.087
EPSY 440	16	.308	.056
EPSY 540	15	.226	.075
EPSY 640	25	.194	.075

TABLE 2

Duncan's Values For Intergroup
$\bar{q}_{st}$ Comparisons: Study A

Class	Grade 11	EPSY 200	EPSY 440	EPSY 540	EPSY 640	EPSY 744
Grade 11	-					
EPSY 200	.064	-				
EPSY 440	.329*	.360*	-			
EPSY 540	.359*	.402*	.008	-		
EPSY 640	.399*	.410*	.004	.014	-	
EPSY 744	.734*	.765*	.390*	.406*	.434*	-

* $p < .05$.

TABLE 4

Duncan's Values For Intergroup
$\bar{q}_{st}$ Comparisons: Study B

Class	Grade 11	EPSY 200	EPSY 440	EPSY 540	EPSY 640
Grade 11	-				
EPSY 200	.012	-			
EPSY 440	.133	.254*	-		
EPSY 540	.452*	.592*	.322*	-	
EPSY 640	.679*	.835*	.503*	.138	-

* $p < .05$.