DOCUMENT RESUME

ED 041 943 TM 000 035

AUTHOR Lewy, Arieh
TITLE Item Analysis Based on Classroom Performance.
PUB DATE Mar 70
NOTE 9p.; Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, Minn., March 1970

EDRS PRICE EDRS Price MF-$0.25 HC-$0.55
DESCRIPTORS *Achievement Tests, Correlation, *Group Norms, Group Tests, *Item Analysis, *Statistical Analysis, Statistics, Tests

ABSTRACT

Statistical procedures employed in item analysis are based on the performance of the individual on a given test. If, however, one desires to assess the efficiency of some treatment, the researcher is interested in the performance of groups of students who received the treatment rather than in the performance of any one individual. In such cases statistics based on individual performance do not constitute a proper basis for item selection and for interpretation of test results. In this context, therefore, the discrimination power of an item should be conceived as its capacity to discriminate among classes and not among individuals. This approach is not only logically sound, but also increases the quantity and quality of the information available. Three new parameters are introduced: item-test correlation based on class data, item standard deviation based on class data, and the intraclass correlation coefficient of an item. An item analysis of a fourth grade arithmetic test is used as an example. Other class data parameters are suggested for future research. (DG)

ITEM ANALYSIS BASED ON CLASSROOM PERFORMANCE

Dr. ARIEH LEWY

Tel-Aviv University

Department of Education

The Background

Statistical procedures employed in test analysis and especially in item analysis are based on the performance of the individual in a given test. Item characteristics which are of interest for the evaluation expert and which frequently serve as a basis for item selection are: the difficulty level of an item, its correlation with the scores on the total test, its correlation with other items of the same test and with one or with several external criteria scores. Several standard computer programs (5) report also the standard deviations of each item, which in most multiple choice tests are curvilineal function of the item difficulty and thus do not convey any new information which is not present in the item difficulty index itself. That test statistics are based on measurements of individual performance is possible due to the fact that the roots of statistical test theory are in clinical testing, which, in turn, has its major interest in individual differences. Whenever one wishes to develop an instrument which will maximalise the discrimination among individuals with regard to certain trait or capability such procedures may be most efficient. But the concern of educational testing and measurement is not only the assessment of individual performance. Tests are used within the framework of research studies come to assess the efficiency of some treatment, like teaching method, curriculum material, organizational setting etc. employed in classroom situations to groups of students. In such cases the researcher is interested in the performance of groups of students who underwent some treatment rather than in the performance of any individual. In such cases statistics based on individual performance do not constitute a proper basis for item selection and for interpretation of test results. Whenever tests

---

Read at the Annual Meeting of the American Educational Research Association, March 5, 1970, Minneapolis

are used for the purpose of assessment of some aspects of classroom teaching item analysis has to be performed on the basis of differences between achievements of classes.

This approach can be justified on the basis of logical considerations but additionally it possesses empirical advantages by providing a variety of information which is not available whenever individual data are used for item analysis.

Logical basis for using class-performance data

The fact that in classroom experimentation the unit of observation is not the individual but the class has been pointed out by Lindquist's classical book on Experimental Design[6]. Lindquist calls attention to the fact that performance of students withing the same class are not independent one of the other, since they are affected by circumstances prevailing in the class. Therefore the number of independent replications is equal to the munber of classes included in the sample and not to the number of individuals comprised in the sample. This idea has been further developed by Lord[7] and utilized in several educational studies like Bock[1], but surprisingly has never been applied to item analysis and item selection procedures. Since the natural setting of test administration is the classroom one has to consider the whole class as a unit and compare performances of classes rather performances of individual. Accordingly questions usually asked in the context of item analysis have to be reformulated. With this view in mind the discrimination power of item e.g. has to be conceived as its capability to discriminate among classes and not among individuals. For this reason it is proposed that the discrimination power of an item should be defined as the correlation between the average achievement of students in each class on a particular item with the average performance of the class on the whole test. Thus far the logical reasons for the suggested modification of item analysis have been described and a new definition for the concept "item discrimination" power has been suggested. But this approach not only increases the logical precision of a concept widely used, but also increases the amount and the quality of information which is processed by item analysis.

Empirical Advantages

The increase in amount and the improvement of the quality of information is illustrated by the presentation of results obtained from item analysis of a fourth grade arithmetic test.

The Test

A 42 item test has been constructed to measure achievement of fourth grade pupils in arithmetic[4]. The test covered several topics taught in the 4th grade according to the local syllabus. In addition to topics taught in all schools the test contained two subsets of items which are of special interest for the purpose of this paper. The first of them is Fractions. This topic has been taught in grade four of some of the schools which study arithmetic according to a new experimental program, while in other schools it is taught only in grade five. Thus some classes studied the topic Fractions before taking the test while others did not systematically study this topic. The second subset of items labelled in this paper, Theory, deals with general aspects of mathematics and contains questions like "Each number which can be divided by 4, can be divided by 2 too" etc., a topic which has not been dealt with directly in classes and may be considered as a kind of measure of numerical ability. And finally there is a third set which contains a single item - item 12 - labelled error which in this case refers to the fact that the correct response has been erronerously provided for the computer analysis and thus the results of this item have to be disregarded, but for analytical purposes it turned out to be valuable for the present topic. The test has been administered to 107 classes comprising 3057 students with an average of 28 students per class. The classes represented the whole range of differences in ability levels existing in the country.

The results of item analysis

Some results of the analysis are presented in table 1. This table contains information based traditionally on individual data, and information based on class data. It can be seen that the test is a typical one, its items represent different difficulty levels.

Their distribution is as follows:

Percent of students giving correct response	No. of item
20 - 29	4
30 - 39	2
40 - 49	9
50 - 59	7
60 - 69	8
70 - 79	7
80 - 89	8

On most of the items 40 - 80 percent of correct responses were obtained, but some items are rather difficult and others quite easy.

Item-test correlation

The traditionally reported item total test biserial correlation coefficient is listed in column (2) of table 1. In column 4 the correlation coefficients among the same variables based on class performance data is listed. The first striking observation is that correlations based on class performance are considerably higher than those based on individual performance data. However a constant difference between the two types of correlations would not be of interest. But one can see that the range of correlations based on class performance is higher than of those based on individual data. The range of the point biserial correlations is from .2 to .6 with a median value of .4 while the range of class-correlations occupies almost the whole possible range of positive correlations, it runs from .2 to .9 with a median value of .7. The reason for this difference is that individual performances are scored dichotomously (0 or 1) while class averages for each item can carry a variety of fraction values from 0 to 1.

It should be added that in the present table the individual point-bis correlations are computed without correction formula for auto correlation and thus their values are slightly inflated[8] while the class correlations are computed with regard to the sum of other items only. Thus the difference between the actual correlation coefficients is larger than those presented here.

It can be seen that in using class data we obtain a coefficient which has more refined discriminational value than that obtained by using individual data.

Standard deviation of class averages

Using class data also enables us to examine some characteristics of items which are not described at all in conventional item analysis.

I refer here to the frequency distribution of class-averages which is presented on the right part of table 1. It can be seen e.g. that on item no. 1 two classes obtained 50-59 percent correct responses and 13 classes obtained 60-69 percent responses. On the same item 21 classes obtained 100 percent of correct responses, i.e. in 21 classes all students responded correctly to this item. The distribution of class averages is characterized by a single parameter i.e. the standard deviation of class averages. Unlike the standard deviation based on individual data, the standard deviation of class averages (see col. 3) is not a function of the item difficulty and thus it conveys information not contained in the item difficulty index itself. Theoretically it is possible that one item having a given difficulty level has widely different averages in a variety of classes while another item with the same difficulty level has quite similar averages in most classes. Thus e.g. in the data presented here both items 37 and items 41 have a difficulty level of 45 and a slightly different standard deviation.

The differences in the size of standard deviation in the present sample are moderate. However, considerable differences have been detected by the author in other tests analyses.

Table 1. — ITEM ANALYSIS OF 4TH GRADE ARITHMETIC TEST

Topic	Item no.	Item Statistics: Ind. Difficulty	Class SD	Ind. corr.	Class corr.	F ratio	Intracl. corr.	Number of Classes obtaining a Given Percent of Correct Responses: Percent of correct Responses 0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	100	All
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
Integer	1	.865	0.118	310	0.589	3.244	.073	-	-	-	-	-	2	13	13	36	22	21	107
Integer	2	.771	0.153	339	0.547	3.877	.092	-	-	-	1	5	10	20	23	25	17	6	107
Int. Division	3	.286	0.234	390	0.590	10.127	.242	31	19	16	11	1	17	5	6	1	-	-	107
Fraction	4	.426	0.251	411	0.611	9.089	.221	14	10	6	17	17	14	12	6	8	2	1	107
Estimation	5	.459	0.228	455	0.710	6.834	.170	6	14	8	18	10	18	18	9	4	2	-	107
Fraction	6	.442	0.295	494	0.705	15.156	.331	15	18	13	8	7	10	10	11	10	3	2	107
Geometry	7	.348	0.236	372	0.491	8.737	.213	17	17	20	13	12	7	10	6	4	1	-	107
Problem	8	.788	0.161	540	0.845	4.586	.112	-	-	1	-	6	8	16	18	29	23	6	107
Problem	9	.812	0.157	523	0.803	4.858	.119	-	-	-	3	-	9	13	19	29	23	11	107
Theory	10	.510	0.223	398	0.673	6.178	.154	2	8	12	10	18	21	12	11	9	3	1	107
Geometry	11	.260	0.198	243	0.282	6.349	.158	20	30	19	13	10	7	5	2	-	-	1	107
Error	12	.018	0.030	-136	-0.320	1.338	.012	105	2	-	-	-	-	-	-	-	-		107
Int. Division	13	.673	0.183	400	0.632	4.623	.113	-	1	3	5	9	19	19	23	19	7	2	107
Fraction	14	.254	0.201	343	0.556	7.792	.192	29	26	13	16	9	6	4	2	2	-	-	107
Estimation	15	.221	0.162	234	0.426	4.541	.110	25	28	34	7	6	4	1	1	-	1	-	107
Fraction	16	.708	0.208	532	0.801	6.483	.161	1	2	3	3	8	16	10	19	31	13	1	107
Geometry	17	.840	0.145	433	0.693	4.563	.111	-	-	1	-	1	4	13	16	28	33	11	107
Problem	18	.547	0.196	522	0.800	4.306	.112	1	3	5	16	21	22	14	15	6	3	1	107
Problem	19	.583	0.195	543	0.837	4.525	.110	1	3	5	7	23	20	20	11	13	3	1	107
Int. Division	20	.563	0.226	535	0.791	6.997	.174	3	8	7	6	14	20	17	13	16	3	-	107
Estimation	21	.782	0.146	404	0.650	3.402	.078	-	-	3	-	4	3	14	32	34	16	1	107
Measurement	22	.480	0.226	527	0.778	6.544	.163	6	6	12	18	19	17	11	9	5	2	2	107
Fraction	23	.629	0.228	575	0.863	7.426	.184	2	2	6	13	8	13	19	16	20	8	-	107
Problem	24	.612	0.181	447	0.740	3.682	.096	1	1	2	8	12	27	21	19	12	2	2	107
Int. Division	25	.615	0.207	523	0.803	5.641	.140	2	3	6	6	15	10	23	23	17	2	-	107
Integer	26	.595	0.216	522	0.775	6.163	.163	3	3	5	10	12	17	22	16	16	1	2	107
Fraction	27	.588	0.259	651	0.910	9.442	.228	5	2	8	16	11	13	11	13	18	8	2	107
Estimation	28	.402	0.170	340	0.550	3.472	.080	2	7	24	25	18	17	9	3	1	1	-	107
Theory	29	.677	0.165	493	0.808	3.502	.081	1	-	3	4	6	18	27	23	24	1	-	107
Problem	30	.742	0.178	558	0.810	4.858	.119		1	1	3	6	16	14	20	28	16	2	107
Problem	31	.487	0.186	509	0.780	4.140	.099	4	4	8	20	22	20	18	8	3	-	-	107
Theory	32	.705	0.161	439	0.729	3.501	.081	1	-	-	4	8	15	21	27	26	4	1	107
Theory	33	.416	0.132	239	0.390	1.996	.034	2	5	11	32	23	28	6	-	-	-	-	107
Theory	34	.681	0.164	386	0.698	3.295	.074	1	-	2	1	10	17	24	28	21	2	1	107
Theory	35	.315	0.181	244	0.488	3.303	.075	8	14	29	31	10	4	7	2	-	-	2	107
Theory	36	.680	0.171	416	0.670	3.598	.083	1	-	2	3	5	20	25	23	21	5	2	107
Measurement	37	.455	0.219	576	0.880	5.882	.146	5	9	16	17	18	12	17	8	3	-	2	107
Integer	38	.714	0.171	502	0.719	4.087	.098	1	1	-	5	6	7	27	25	28	6	1	107
Integer	39	.552	0.229	557	0.760	6.926	.172	4	5	7	12	14	18	18	15	9	4	1	107
Int. Division	40	.655	0.230	629	0.866	7.631	.188	2	3	7	5	11	5	23	20	21	10	-	107
Fraction	41	.451	0.284	512	0.763	13.217	.300	14	13	15	10	9	7	9	17	10	2	1	107
Estimation	42	.777	0.177	499	0.620	5.030	.124	1	-	1	2	6	11	9	20	40	14	3	107

Interclass correlation

A third type of information which may be of interest for the evaluation expert is the intraclass correlation coefficient reported in column 6 of the handout (Haggard, 1958)[3]. This parameter is the function of the ratio obtained by the division of the between classes mean sum of square by the within classes mean square. The intraclass correlation is an index of the degree of similarity of responses given by pupils in any single class. As indicated already the topic Fraction has been taught in some classes and in other classes it has not been taught at all. As a result of this the intraclass correlations of items in this category are relatively high. Thus e.g. the intraclass correlation of item 41 is very high. This item represents a relatively easy question in fraction. Those classes who studied the relevant materiel obtained very high averages on this item while classes which did not study this material obtained averages close to zero. The frequency distribution of class means on this item has two peaks i.e. it has bimodal shape. This is an interesting example of an item whose variance stems mainly from between class differences while on the other hand the variance of item 33 stems mainly from within class differences. The former one is a content oriented item while the later one is more saturated with general ability. The comparison of these two items supports the feeling that items testing achievement have higher intraclass correlation than items measuring ability.

Space does not permit to present other statistics which can be computed on the basis of class data. But the purpose of the present paper is not to convey the findings of a specific study but rather to call attention to the existence of basic data which have been completely neglected in dealing with item statistics and which may yield a wide variety of parameters increasing the amount and the better quality of information obtained from test results. It is often thought that the currently based statistics tell everything about test items, which may be of interest. In fact the practice of item analysis and item selection has not changed over decades and the endeavor of psychometrists is more concentrated on provisions of computational shortcuts for the existing parameters than for seeking additional parameters which may increase our understanding of tests.

Topics for Further Study

The present paper introduces three parameters which have not been used before and these are: 1) item-test correlation based on class data; 2) item standard deviation based on class data, and 3) intraclass correlation coefficient of the item. In addition to these parameters some other topics for research utilising class data are as follows:

1. Inter-item correlation matrix computed on basis of class results. It is expected that these correlation coefficients will have a pattern different than those based on individual data.
2. Factor analytic studies can be performed on both types of correlation matrices.
3. Test-reliability coefficients of various types can be computed on basis of class results, too. Thus it is possible to employ split-half computational procedures as well as the formulae of Kude Richardson generalized by Cronbach[(2)] as the coefficient alpha.
4. Studies can be conducted to identify factors which affect the similarity or the difference between statistics obtained from individual and from class data.
5. Studies can be conducted to find out whether items testing achievement can be differentiated from items testing ability by the size of their intraclass correlation.

It can be envisaged a long series of studies based on the idea mentioned here, which may change the nature of major topics dealt with in test theory. Studies based on class-data may contribute new chapters to standard text books on measurement and evaluation and may sharpen the understanding of differences between tests used for the purpose of individual guidance and selection on one hand, and tests used for the prupose of testing the efficiency of classroom teaching practices on the other hand.

BIBLIOGRAPHY

1. Bock, R. D. "Contributions of Multivariate Experimental Design to Educational Research" In Cattell B. R. (ed) Handbook of Multivariate Experimental Psychology. Rand M. Nally, Chicago 1966.

2. Cronbach, L. J. "Coefficient alpha and the internal structure of tests". Psychometrika 1951, v. 16 p. 297-334.

3. Haggard, E. A. Intraclass Correlation and the analysis of variance New York; Dryden 1958

4. Lewy A. and Chen M. "Scholastic Achievements of Fourth Grade Students in Israel" (in preparation)

5. Lewy A. and Crawford R. "Scoring Test Battery" Educational and Psychological Measurement. v. 26 p. 185-188, 1966

6. Lindquist E. F. Design and Analysis of Experiments in Psychology and Education. Boston: Houghton Mifflin 1963.

7. Lord F. M. "Test Norms and Sampling Theory", Jour. of Exp. Ed., v. 27, p. 247-263.

8. Wolf, R. (To be supplied) Journal of Educational Measurement, 1967 (?)