AN EVALUATION OF ITEM SELECTION METHODS
BY A CRITERION OF INTERNAL CONSISTENCY

FRANK BLASCIK

Submitted in partial fulfillment of the requirements for the degree of Master of Arts, in the Faculty of Psychology,
McGill University

MONTREAL

April, 1949

AN EVALUATION OF ITEM SELECTION METHODS
BY A CRITERION OF INTERNAL CONSISTENCY

Frank Blascik

McGill University

Various methods have been employed in the evaluation of item selection methods. Typical among these studies are those made by Barthelmess, Lentz and Long and Sandiford. Barthelmess (1), has used an intercorrelational technique in evaluating various methods of item selection. Validity values were computed for each of the hundred elements of the McCall Multi-Mental Scale, Elementary Form 1, by each of the methods of item selection being studied, namely, Eta, Long, McCall, Vincent, Corrected Vincent and Bi-serial r. These methods were then evaluated on the basis of intercorrelation of each method with all the other methods. Results showed that the Eta (Correlation Ratio) and the Bi-serial r methods ranked first and second respectively. Using a criterion X, composed of a series of tests (Stanford Achievement Test, Stanford Revision of the Binet-Simon Intelligence Test A, Thorndike-McCall Reading Scale, Woody-McCall Mixed Fundamentals in Arithmetic, Morrison-McCall Spelling Scale), Barthelmess judged the validity methods under study according to their success in selecting the ten best, twenty best, thirty best, forty best, fifty best items. On the basis of correlations with criterion X of the ten best items, as chosen by each method, the Long, McCall, Eta, and Bi-serial r methods ranked first, second, third and fourth respectively.

Lentz (7), in his study, experimentally evaluated four methods, namely Upper and Lower Thirds, Vincent Overlapping, McCall and the Summation of Agreements method. By each of these methods, the 100 best items were selected from the 150 items of a test of conservatism. Each set of 100 items was then tested for reliability; the method which selected a set of items showing the greatest increase in the reliability of the test was rated as the best method of item selection. Using the total number of conservative reactions as the criterion score, and odd versus even Spearman-Brown technique, the reliability coefficients which were obtained ranked the Upper and Lower Thirds method first, and the Summation of Agreements and the Vincent Overlapping methods second and third respectively.

Long and Sandiford (9), evaluated thirteen item selection methods by correlating sets of 50 and 100 items, as selected by each of these methods, with the total score on the 482 items of the test. Using this total score as the criterion, the McCall, McCall-Long-Bliss and Clark methods ranked as the first, second and third best methods respectively. Cf those methods employed in the Long and Sandiford study which are used in this study, Upper and Lower Thirds, Upper and Lower 27% and Bi-serial r rank as the fourth, sixth and ninth best methods respectively, of selecting the best 100 items from the spelling test of 482 items.

The purpose of the present study is to determine the relative efficacy of various methods of item selection when that efficacy is determined by the consistency of response found among the items selected. In formal terms, the item selection problem may be stated as follows: within any group of N items, a sub-group of k items may be selected in ${}_nC_k$ ways; within ${}_nC_k$ possible combinations of k items, one combination of maximal internal consistency exists; how may this combination of k items be identified?

The selection of the most consistent combination of k items from any set of N items is a matter of great arithmetical labour. All methods of item selection now in use provide approximate solutions to the problem. Existing methods may be classified as combinatorial - e.g. Horst's method of successive residuals, and Toops' L method; or non-combinatorial - e.g. Bi-serial r, Upper and Lower Thirds, etc. It is clear that the combinatorial methods provide a more effective solution to the problem than the non-combinatorial methods when used in the practical situation. The combinatorial methods, however, are in most cases prohibitive in practice due to the arithmetical labour involved. The majority of tests, therefore, have utilized some form of non-combinatorial technique. Hitherto, difficulties have been experienced in appraising the relative efficiency of the various techniques used due to the lack of an adequate statistic descriptive of the consistency present in any response pattern.

A statistic termed "a coefficient of consistency" has recently been developed by Dr. G. A. Ferguson, of McGill University, and appears appropriate for this purpose. In this study, it is proposed to use this statistic as the criterion in evaluating a number of the item selection methods commonly in use.

An answer pattern was prepared from the scores obtained by 108 pupils on a Moray House Test. This was done by preparing a table containing N columns (N - the number of persons in the sample) and k rows (k = the number of items in the test). The test used was such that the score of 1 was the sign for a pass, and the score of 0 for a failure on a particular item. The appropriate symbol, 1 or 0, was entered in all cells of the table for all persons and all items.

When a test of k items is administered to a sample of n persons, a relation is established between every individual in the sample and every item on the test. This yields therefore, a relational field which may be expressed in matrix form by assigning a column to each of the n individuals in the sample, and a row to each of the k items of the test. The matrix thus contains nk elements which may be written in the forms $>$, where the individual passes the item, and $<$, where the individual fails the item. If now, a 1 for a pass and a 0 for a failure is assigned, these relations are replaced by 1 and 0, and the resulting matrix may be spoken of as an answer pattern matrix.

If item 1 is answered correctly p_1 times, item 2 is answered correctly p_2 times, and so on, these items will be placed in order of difficulty when their p's are in order of magnitude. This table of values of p for the items of a test, placed in order of magnitude, is the answer pattern of that test. In practice, when rows of an answer pattern are ordered according to difficulty such that $p_1 > p_2 > p_3 > \ldots\ldots\ldots p_n$, and when the columns of the answer pattern are ordered in terms of the total score on the test obtained by the testees, the result is seldom a unique answer pattern, (showing no inconsistencies), since the pattern depends not only on the difficulty of the items, but also on the character of the group of testees (14). Inconsistencies in the answer pattern are numerous. The consistency criterion used here is a statistic descriptive of the observed consistency in an answer pattern. This statistic serves as an index of the extent to which the criteria of order are satisfied.

The consistency criterion may be defined as: the number of consistent comparisons minus the number of consistent comparisons expected by chance, divided by the total number of possible consistent comparisons minus the number of consistent comparisons expected by chance.

A consistent comparison is such that the following possibilities are found in the comparison of responses of any individual on two items, i and j, of difficulties p_i and p_j, where $p_i > p_j$. Any individual may:

a) Pass both i and j.

b) Fail both i and j.

c) Pass i and fail j.

If a person fails i and passes j, the comparison of responses of the individual on these two items is spoken of as an inconsistent comparison.

The formula for the consistency criterion is written as:

$$C = \frac{S_t^2 - \Sigma p_i + \Sigma p_i^2}{2\Sigma p_i \cdot c_{(i-1)} - (\Sigma p_i)^2 + \Sigma p_i^2}$$

where: S_t^2 is the variance of the obtained test scores.

p_i is the proportion of persons passing the item.

c represents cardinal numbers from 0 to k - 1, these numbers being assigned to the items according to the proportion of persons passing the items; 0 being assigned to that item passed by the greatest number of persons.

The methods of item selection applied to the answer pattern matrix yielded by the results of the Moray House Test were:

a) Bi-serial r

b) Upper and Lower 27%

c) Upper and Lower Thirds

d) Difficulty Value

e) Consistency Index

Combinations of k items from the N items of the test were selected on the basis of these five methods. A consistency coefficient was then computed for each set of k items, the methods being evaluated therefore, in terms of these coefficients.

In this study, the 100 items of a Moray House Test were ranked according to their validity as determined by the five methods and then marked off into groups of 25 items, thus yielding the first, second, third and fourth best groups of items. These groups are shown in Tables IV, V, VI, and VII of the Appendix.

By combining the first two groups, a group composed of the best fifty items as selected by each of the methods was obtained. Similarly, a combination of the first three groups yielded a group composed of the best 75 items as selected by each of the methods.

The consistency coefficients were computed for each of these groups composed of the best 25, 50 and 75 items as selected by each of the five methods of item selection under study. The methods are, therefore, evaluated in terms of these coefficients, the method yielding the sub-group of items with the largest consistency coefficient being the most effective method of item selection among the methods studied. Table II indicates the efficiency of each of these methods in selecting sub-groups containing different numbers of items.

A reliability criterion was also used in the evaluation of these methods. Richardson-Kuder (13) coefficients were computed for the best 25, 50 and 75 items as selected by each of the five methods. The results, shown in Table III, indicate which method, according to the reliability criterion, is most efficient.

In Tables IV, V, VI and VII of the Appendix, the Best, Second Best, Third Best and Fourth Best groups of 25 items respectively, have been ranked according to validity values obtained by the different methods. Under Difficulty Value, the items are ranked from least to greatest difficulty. The correlation coefficients between Difficulty Value and the other methods of Item Selection under study are shown in Table I. The method of Upper and Lower Thirds shows the highest correlation with Difficulty Value. There is an appreciable difference in the correlation coefficients between the Upper and Lower 27% method and Difficulty Value and the Upper and Lower Thirds method and Difficulty Value. This reflects the influence of scores crowding around the middle 50% of the difficulty range. There is a reference made to these scores of 50% difficulty in Appendix B.

The relative efficacy of the five methods of Item Selection under study is shown in Table II, in terms of consistency coefficients. The method yielding the highest consistency coefficient when 25 items are selected from the Moray House Test is the Consistency Index method. According to the criterion of consistency, Bi-serial r and Upper and Lower 27% rank as the second and third best methods, respectively, employed in this study.

In terms of the Richardson-Kuder reliability coefficients, shown in Table III, the Upper and Lower Thirds method appears to be the most efficient method of those under study for selecting a sub-group of 25 and 50 items from the test, with the Upper and Lower 27% method a close second. Bi-serial r shows the highest reliability coefficient when 75 items are selected from the test.

A comparison of Tables II and III shows that the criterion of internal consistency is a more sensitive instrument for evaluating the methods of Item Selection under study than is the criterion of reliability.

On the basis of both reliability and consistency coefficients, Bi-serial r appears to be the best overall method for selecting sub-groups of items from this test.

However, the method of Upper and Lower 27% which is the second most efficient overall method as indicated by the consistency criterion, may be the better method to employ in the practical situation, since the amount of labour involved in its calculation is much less than in calculating Bi-serial r values for each of the items.

When the item selection techniques are evaluated on the basis of the consistency or reliability criteria used in this study, there appears to be little difference between the value of those techniques, such as the Upper and Lower 27% method, which discriminate against items of 50% difficulty, and those techniques, such as Bi-serial r, which make no such discrimination.

TABLE I

CORRELATION COEFFICIENTS BETWEEN DIFFICULTY VALUE AND METHODS OF ITEM SELECTION UNDER STUDY

	Bi-serial r	Upper and Lower Thirds	Consistency Index	Upper and Lower 27%
Difficulty Value	.62	.57	.48	.30

TABLE II

CONSISTENCY COEFFICIENTS COMPUTED FOR GROUPS OF 25, 50 & 75 ITEMS SELECTED BY 5 DIFFERENT METHODS OF ITEM SELECTION

Method	Best 25 Items	Best 50 Items	Best 75 Items
Bi-serial r	.500	.461	.443
Upper & Lower 27%	.499	.457	.403
Upper & Lower Thirds	.485	.424	.378
Difficulty Value	.366	.335	.346
Consistency Index	.517	.363	.390

TABLE III

RELIABILITY COEFFICIENTS COMPUTED FOR GROUPS OF 25,50 & 75 ITEMS SELECTED BY 5 DIFFERENT METHODS OF ITEM SELECTION

Method	Best 25 Items	Best 50 Items	Best 75 Items
Bi-serial r	.896	.944	.962
Upper & Lower 27%	.904	.948	.960
Upper & Lower Thirds	.913	.948	.932
Difficulty Value	.878	.934	.956
Consistency Index	.907	.933	.959

APPENDIX

A - Bi-serial r

Bi-serial r is a method of correlation applicable to data having two variables, one being quantitative and continuous, and the other being a dichotomy. For example, this method would be suitable for determining the relationship between two variables such as intelligence and weight; intelligence being the quantitative and continuous variable, and weight being the two category variable.

In this instance the Bi-serial r method has been applied to find the coefficient of correlation between total scores and success or failure in an individual item. Each item then, yields its own Bi-serial r which may be considered as a measure of that item's validity; the higher the coefficient, the better the item for predicting the criterion trait.

Holzinger (6), discusses the two basic assumptions on which this method rests. The first assumption is that the variable for which there is only a dichotomous division is distributed normally. Secondly, it must be assumed that the relationship between the two variables is linear.

The formula used in calculating Bi-serial r is:

where M_2 = mean criterion score of the group solving the item correctly.

M_1 = mean criterion score of the group failing to solve the item correctly.

σ = the Standard Deviation of all criterion scores.

p = proportion of total group solving the item correctly.

q = 1 - p, proportion of total group failing to solve the item correctly.

Z = ordinate of normal curve cutting off p proportion of cases; obtained from tables (5).

B - Upper and Lower 27% Method

This method is an elaboration of the Upper and Lower method. The validity of the item is taken not by the difference between the percentage proportions of the Upper and Lower groups answering the item correctly, but by the distance, in sigma units, between the ordinates which cut off these respective proportions from the area of the normal probability curve. In calculating the correlation coefficient by the method originated by Dr. T. L. Kelley (7), the following steps are involved:

(1) Find the proportion of the Upper 27% passing the item correctly.

(2) Find the proportion of the Lower 27% passing the item correctly.

(3) From the tables of the normal probability curve (6), find the position, in sigma values, of the ordinate cutting off the proportion of cases found in step (1).

(4) Find the position, in sigma values, of the ordinate cutting off the proportion of cases found in step (2).

(5) Subtract the result yielded in step (4) from the result yielded in step (3); the remainder will represent the validity value of the item as derived by the Kelley method.

The chief disadvantage of this correlation coefficient arises from the fact that it becomes more and more difficult to raise a coefficient by a certain number of hundredths as perfect correlation is approached; a difference of .15 between .80 and .95, for example, represents a far greater disparity in actual relationship than does a difference of .15 between .05 and .20.

In order to obtain an index of discriminating power that is comparable from item to item, and which would consequently be easy to interpret and convenient to use, F. B. Davis (2), suggests the use of the Z statistic developed from the product-moment r, by R. A. Fisher (5). This statistic can be employed as a direct measure of the amount of discriminating power possessed by an individual item and may be considered essentially comparable from item to item because a given increase in the value of Z has a constant meaning at any part of its range of values. Davis, after summarizing other properties of Z, suggests, as mentioned above, that for test-construction work, it would be even more convenient to use a linear function of Z that would eliminate decimals and permit the required range of discrimination indices to be restricted to 0 to 100. The next step was to construct an item-analysis chart in such a way as to yield discrimination indices having these desirable properties.

J. C. Flanagan (4), had constructed a chart which gave the index of discrimination in terms of the degree of relationship shown between the item and the criterion based on Pearson's work (11). The chart derived shows the values of the product-moment correlation coefficient corresponding to given proportions of successes in the Upper and Lower 27% of the criterion group.

Since Flanagan had already obtained the values of the correlation coefficients in a normal bivariate surface corresponding to the desired combinations of proportions of successes in the highest and lowest 27% of the sample that were needed to construct Davis' item-analysis chart, the values in Flanagan's table were simply converted into discrimination indices by means of a table which included equivalent values of product-moment r, Fisher's Z, and Davis' discrimination index. After the coefficients in Flanagan's table had been transformed into discrimination indices, the values in the item-analysis chart were smoothed and checked for accuracy.

In order to obtain the discrimination index and correlation coefficient for each item, the following steps had to be taken:

(1) The proportion of the Upper 27% passing the item correctly was found.

(2) The proportion of the Lower 27% passing the item correctly was found.

(3) With these figures it was possible to enter the item-analysis chart which yielded the discrimination index.

(4) From Table 1 in Davis' "Item-Analysis Data" it was possible to read off the correlation coefficient equivalent to the discrimination index obtained in step (3).

C - Upper and Lower Thirds Method

This method is a derivative of the Upper and Lower Halves method in which the procedure employed in selecting items is in terms of the percentage of higher scoring persons passing the item subtracted from the percentage of lower scoring persons passing the item, the higher scoring and lower scoring pupils having been selected on the basis of their criterion scores being above or below the median score of all pupils (10).

When this method is employed, the fact that the scores crowding around the middle 50% of the range tend to make the difference between the percentage of higher scoring pupils passing the item and the percentage of lower scoring pupils passing the item smaller than it would be if the middle range were cut out, becomes apparent.

This difference increases, on the average, as one passes from Upper and Lower Halves, through Upper and Lower Thirds, Quarters, and Tenths. However, as the number of cases diminishes, the probable error increases. Kelley has developed the proof that the size of the Upper and Lower categories should each be 27% of the total number of persons to produce a maximum ratio between the difference of their means and the probable error of the difference.

In this study, items were selected on the basis of the method of Upper and Lower Thirds, as well as the method of Upper and Lower 27%.

The steps involved in the Upper and Lower Thirds methods are as follows:

(1) The students who attempted the test were arranged in order of the size of respective criterion scores.

(2) The highest one-third and the lowest one-third of the pupils were marked off.

(3) Each item was evaluated by the difference between the number of passes obtained on it by the highest third of the pupils and the number of passes obtained on it by the lowest third of the pupils.

(4) This difference was divided by $\frac{N}{3}$.

(5) The formula employed was $\dfrac{N_1 - N_2}{\frac{N}{3}}$.

where N_1 = number of students in highest third who passed the item.

N_2 = number of students in lowest third who passed the item.

N = total number of students.

The validity of the item thus obtained is in terms of proportion of successes.

D - Difficulty Value Method

Items Validated in Terms of Difficulty

Every test item which is scored right or wrong performs a dichotomous function; that is, it divides the persons tested into two groups, (1) persons passing the item, and (2) persons failing the item. The level of ability at which a particular item is able to dichotomize the persons tested depends on the difficulty of the item. The difficulty of an item is defined as the proportion of the sample of testees who mark the item correctly.

The procedure used here in establishing the validity of each item in terms of Difficulty Value was simply to find the proportion of persons who passed the item.

E - Consistency Index Method

Different test items vary considerably in the degree of correspondence between the n persons passing them and the n persons who score highest on the test. It is desirable however, to choose items which show a high degree of such correspondence. For this purpose, a method based on Consistency Theory and developed by Dr. G. A. Ferguson, has been devised. The method establishes the validity of each item on the basis of the ability of the item to dichotomize the persons tested. This level of ability depends on the difficulty of the item which is defined by the number of persons passing it. Every item is said to discriminate at the particular level of ability at which it dichotomizes the group tested. If a test is desired which will arrange the persons tested reliably according to their abilities, it is necessary that the n persons passing each item should correspond as closely as possible to the persons making the n highest scores on the whole test.

This method of item selection employs the following steps:

(1) The persons tested are arranged in order of total score from highest to lowest.

(2) For each item, the proportion of persons failing the item who would have passed had the item discriminated perfectly, is found. This proportion for each item is marked W.

(3) The proportion of persons who passed the item, p, is found.

(4) The proportion of persons who failed the item, q, is found.

(5) The validity, r, of the item is obtained from the formula

$$r = 1 - \frac{W}{pq}.$$

In order to clarify step (2), let the following be an answer pattern of a single item in a test administered to a group of 12 persons. The persons are arranged according to their scores on the whole test, p_1 having made a higher score on the whole test than p_2, etc.

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}
1_a	1	1	1	0	1	0	0	1	1	1	0	0
1_b	1	1	1	1	1	1	1	0	0	0	0	0

Item 1_a is passed by 7 of the 12 persons tested. It will be observed that the 7 persons passing this item are not the 7 persons making the highest scores on the whole test. If this had been so, the answer pattern would have been as shown in row 1_b. Three persons pass the item who are not among the seven best pupils as selected by the item, and three persons fail the item who are not among the five lowest scoring pupils. Thus the proportion of persons failing the item who would have passed had the item discriminated perfectly, or W, is equal to $3/12 = .25$.

TABLE IV

BEST 25 ITEMS AS SELECTED BY

5 DIFFERENT METHODS OF ITEM SELECTION

Difficulty Value	Bi-serial r	Consistency Index	Upper and Lower 27%	Upper and Lower Thirds
1	72	18	63	17
2	91	32	17	37
22	68	37	67	76
10	17	17	37	43
23	18	72	14	32
19	37	5	65	33
9	77	43	72	39
47	63	67	18	40
24	65	77	29	63
50	80	36	32	65
4	32	7	50	67
30	67	40	53	7
42	79	22	34	18
48	95	65	56	56
56	29	71	64	71
15	53	41	43	5
29	34	33	7	72
8	69	63	71	8
3	100	68	2	11
17	7	69	73	14
5	36	56	11	15
31	56	80	15	16
37	87	20	16	29
45	41	34	80	68
44	71	91	69	6

TABLE V

SECOND BEST 25 ITEMS AS SELECTED BY 5 DIFFERENT METHODS OF ITEM SELECTION

Difficulty Value	Bi-serial r	Consistency Index	Upper and Lower 27%	Upper and Lower Thirds
12	33	6	4	41
16	40	16	19	53
21	50	19	20	57
32	98	53	57	4
33	15	76	68	19
39	5	10	91	34
7	60	15	40	36
11	76	42	5	66
43	94	4	33	80
63	54	27	61	3
6	8	47	36	10
40	14	64	41	30
58	51	66	66	48
67	57	2	74	50
38	61	3	76	64
85	73	12	77	12
86	16	14	8	20
26	64	21	10	42
65	66	57	90	45
66	74	8	92	27
71	4	97	45	69
89	19	13	30	73
14	20	48	38	13
36	11	1	94	24
57	75	30	42	47

TABLE VI

THIRD BEST 25 ITEMS AS SELECTED BY

5 DIFFERENT METHODS OF ITEM SELECTION

Difficulty Value	Bi-serial r	Consistency Index	Upper and Lower 27%	Upper and Lower Thirds
68	6	78	60	61
25	88	81	97	86
41	22	24	3	91
51	10	39	22	2
13	30	45	96	9
20	2	54	13	21
27	24	61	54	22
64	27	74	81	58
46	42	9	51	25
53	92	99	52	31
76	97	35	86	54
18	96	58	100	23
72	78	83	6	35
49	3	96	24	51
52	35	29	70	52
35	45	50	12	74
59	70	51	31	92
80	86	70	58	38
34	12	73	25	85
73	43	94	75	26
87	25	31	98	44
90	47	44	1	59
92	55	86	27	77
69	1	89	85	96
81	52	59	47	1

TABLE VII

FOURTH BEST 25 ITEMS AS SELECTED BY 5 DIFFERENT METHODS OF ITEM SELECTION

Difficulty Value	Bi-serial r	Consistency Index	Upper and Lower 27%	Upper and Lower Thirds
54	58	25	48	28
61	81	28	79	46
91	21	23	95	55
28	39	52	21	70
82	48	90	59	81
96	83	92	9	60
55	9	100	23	94
70	13	38	35	78
74	38	85	55	82
78	59	55	26	89
77	23	11	28	90
94	31	26	88	97
97	28	60	39	79
88	85	88	62	88
60	90	46	83	49
79	89	95	46	100
84	26	83	78	95
83	82	49	93	98
93	84	79	82	75
99	93	98	87	83
75	44	87	89	62
98	46	62	44	84
100	49	75	49	87
95	62	84	84	93
62	99	93	99	99

BIBLIOGRAPHY

1. BARTHELMESS, H. M.: The Validity of Intelligence Test Elements. New York, Bureau of Publications, Teachers College, Columbia University, 1931. Contributions to Education, No. 505.

2. DAVIS, F. B.: Item Analysis Data. Harvard Education Papers No. 2.

3. FERGUSON, G. A.: The Reliability of Mental Tests, University of London Press Ltd., October 1941.

4. FLANAGAN, J. C.: General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from the Data at the Tails of the Distribution. Jour. Educ. Psychol., XXX (December 1939), pp. 674-680.

5. FISHER, R. A.: Statistical Methods for Research Workers, Oliver and Boyd, 1938.

6. HOLZINGER, A. J.: Statistical Methods for Students in Eduction. Boston, Ginn & Company, 1928.

7. KELLEY, T. L.: Statistical Method, New York, Macmillan, 1924.

8. LAWLEY, D. N.: On the Problems Connected with Item Selection and Test Construction, Proceedings of Royal Society of Edinburgh, 61, Section A, Part III, 1942-3.

9. LENTZ, T. F., HIRSHSTEIN, B., and FINCH, J. H.: Evaluation of Methods of Evaluating Test Items. Jour. Educ. Psychol. XXIII, 5, May, 1932.

10. LONG, J. A. and SANDIFORD, P.: The Validation of Test Items. Bulletin No. 3 of the Dept. of Educ. Research.

11. PEARSON, K.: Tables for Statisticians and Biometricians, London: Biometric Laboratory, University College, 1931, Part II, Tables VIII and IX.

12. RICHARDSON, M. W.: The Relation Between the Difficulty and Differential Validity of a Test. Psychometrika, Vol. I, No. 2, 1936.

13. RICHARDSON, M. W., and KUDER, G. F.: The Calculation of Test Reliability Coefficients Based upon the Method of Rational Equivalence, Jour. Educ. Psychol., 1939, 30, pp. 681-687.

14. SYMONDS, P. M.: Choice of Items for a Test on the Basis of Difficulty. Jour. Educ. Psychol. 1929.

15. THURSTONE, T. G.: The Difficulty of a Test and its Diagnostic Value. Jour. Educ. Psychol., Vol. 23, 1932.

16. WALKER, D. A.: Answer Pattern and Score - Scatter in Tests and Examinations. Brit. Jour. Psychol., Vol. XXII, 1931, pp. 73-86.

________ Answer Pattern and Score - Scatter in Tests and Examinations. ibid., Vol. XXVI, 1936, pp. 301-302.

________ Answer Pattern and Score - Scatter in Tests and Examinations. ibid., Vol. XXX, 1940, pp. 248-260.