6-1-1954

Item Analysis of the Ability Problems Questionaire for Freshman at The University of North Dakota

Lawrence La Fave, Jr.

ITEM ANALYSIS OF THE ABILITY PROBLEMS QUESTIONNAIRE FOR FRESHMEN AT THE UNIVERSITY OF NORTH DAKOTA

A Thesis

Submitted to the Graduate Faculty

of the

University of North Dakota

by

Lawrence La Fave, Jr.

in partial fulfillment of the requirements

for the

Degree of Master of Arts

June, 1954

This thesis submitted by Lawrence La Fave, Jr., in partial fulfillment of the requirements for the degree of Master of Arts at the University of North Dakota, is hereby approved by the committee of instruction under whom the work has been done.

Hermann F. Bruegel
Chairman

Geo. O. Brown Jr.

Peter A. Munch

Daryle E. Kuder
Dean of the Graduate School

ACKNOWLEDGMENTS

The author wishes to thank Dr. Hermann F. Buegel for suggesting the topic of this thesis, and for his guidance and criticism.

Lawrence La Fave, Jr.

TABLE OF CONTENTS

CHAPTER		PAGE
	Approval	i
	Acknowledgments	ii
	Table of Contents	iii
	List of Tables	iv
	Illustrations	iv
I.	THE PROBLEM AND HISTORY	1
II.	THE METHODS OF INVESTIGATION	8
III.	RESULTS AND DISCUSSION	16
	The Slope Method	27
	The Octile Discriminability Method	28
	The Probability Method	31
	Comparison of the Three Methods	35
IV.	SUMMARY, CONCLUSIONS AND SUGGESTIONS	46
	Summary	46
	Conclusions	46
	Suggestions	47
	BIBLIOGRAPHY	49
	APPENDIX	51

LIST OF TABLES

TABLE		PAGE
I.	Symbols and Abbreviations	2
II.	Number of Students Receiving Each Raw Score on the APQ and Corresponding Group	17
III.	Essential Statistics	19
IV.	Number, IN_{gh}, and Percentage, $I\%_{gh}$, of Each Octile Answering Each Item Incorrectly. . . .	20
V.	Item Analysis Information for Slope Method and Octile Discriminability Method	30
VI.	Item Analysis Information for Probability Method	33

ILLUSTRATIONS

FIGURE		PAGE
I.	Graphs of the Percentage of Failures in Each Octile: a) for Item 15; b) for Item 16	29

CHAPTER I

THE PROBLEM AND HISTORY

The Ability Problems Questionnaire (APQ) was designed to discriminate between those who could and those who could not be expected to do college level work. The purpose of this test was to produce a wider separation of individuals scoring below the mean, especially those in the lowest quartile of the group.

This test may be described as a group, objective, paper and pencil, maximum performance test of ability (5,13-18).[1] Since the APQ[2] is presumably intended for prediction to work at the college level, some criterion, external to the test itself, should be established for the purpose of validating the test empirically. The criterion used here is the grades that these students later received.

This was one of two tests administered to matriculating freshmen at the University of North Dakota for the first time in the fall of 1952. At that time, the APQ was administered to 623 students. They were allowed twenty minutes to complete the test. In view of the performance of freshmen on other tests it was believed that most of the students could finish the APQ within such a time limit. Less than forty percent of them, however, did finish the test.

---

[1] Number of references and pages are given.

[2] Abbreviations and symbols are listed in Table I.

TABLE I

SYMBOLS AND ABBREVIATIONS

APQ means the Ability Problems Questionnaire.

UND means the University of North Dakota.

NS means not significant at the .05 level of confidence.

N denotes total number of subjects.

M equals the sample mean.

M_g equals the average test score for those answering item g correctly.

SD equals the sample standard deviation.

SE equals the standard error of measurement.

h is a subscript designating octiles.

g is a subscript designating items.

s_g denotes the standard deviation of item g.

K represents the number of items in the APQ.

$r_{bis.}$ is the sample biserial correlation coefficient.

p_g is the proportion answering the item correctly.

q_g is the proportion answering the item incorrectly.

CN_g is the number of people that answered item g correctly.

IN_{gh} equals the number of subjects in octile h that answered item g incorrectly.

$I\%_{gh}$ equals the percentage of subjects in octile h that answered item g incorrectly.

P is the probability.

TABLE I (Continued)

z equals the height of the ordinate in the normal curve dividing p from q.

r_{xg} is the point-biserial correlation of item with total test score.

r_{xgsg} is the reliability index.

A test designed to produce a wider separation of individuals scoring below the mean than those scoring above would, if it succeeded to any degree in its intent, have negative skewness.

Anderson (1,14) states that he found a positive skewness, .90. His graph (1,17) clearly indicates that this figure is erroneous and that his data yield a moderate negative skewness.

Anderson (1,18) suggests that lengthening the time limit, and the writer's item analysis information, if used correctly, "...may give the curve a more positive skewness and with it more discrimination between good and poor students." On the contrary, if the test was designed to discriminate between those capable of doing college level work and those incapable of doing so, and if the majority of subjects who took the test were capable of doing college level work, as Anderson's data (1,33-52) seem to indicate, then a curve of negative skewness would far better discriminate between good and poor students (7, 8, 10). In this way, large differences would occur between the scores of the poor students. Hence, that minority incapable of doing college level work would be far more readily detected than if the test had the high positive skewness that Anderson apparently desires.

Anderson concerned himself with all the scores of those

matriculating freshmen who took the test in the fall of 1952. The writer, however, could do an item analysis only of the scores of those who completed the test. It would not be surprising, therefore, if the skewness found by Anderson were not in agreement with that found by the author.

The scores on which the present investigation is based showed a skewness of -.38. Neither the skewness found by Anderson nor that found in this investigation are statistically significant. Consequently, the curves of the distributions of both studies roughly resemble the normal probability curve. Anderson believes that this is not the most desirable type of distribution for discriminatory purposes (1,1).

Anderson (1,24) determined the validity of the APQ. He found the quintiserial correlations between APQ scores and grade point averages for each of two semesters. He found grade point averages to correlate .53 with APQ scores for the first semester and .30 for the second.

Since these correlations are not very high, it seems likely that some of the test items are faulty. Item analysis might be employed in an attempt to find out just what items are failing to enhance the predictive value of the test. Deletion of such useless items could cause the APQ to become a more valid instrument for predicting ability to perform college level work at UND. In this way, the error of meas-

urement would probably be decreased and the ability to predict, on the basis of empirical evidence, enhanced (10).

Thus, arose the problem this thesis seeks to solve: to identify items in the APQ that empirically reveal themselves of dubious value in their present form for purposes of prediction at UND.

In recent years, item analysis has attracted the interest and labor of a multitude of people who have attached the problem in an almost unlimited number of ways. Their methods have ranged from the simple, but crude, one of deleting items on which the lower half in test performance has performed as well as the upper half, to the sophisticated and laborious process which has been advocated by Gulliksen (10,363-395).

The result has been an increasing awareness of what good test items are. Cronbach (5,78-79) remarks on this subject:

> Good test items measure what the testor wants to measure. This is the principal factor in logical validity of a test. With care the test maker can 'purify' his test considerably. One important method of removing test items loaded with irrelevant factors is the 'internal-consistency' test. If an item measures what the remainder of the test does, it should have a high correlation with the total test....Good items are unambiguous (with the exception of some personality tests where ambiguity is deliberate)....if choices are offered in an ability test, competent judges should agree that there is only one acceptable answer....Catch questions are undesirable....Good items should have difficulty appropriate to the group tested. The best tests for measuring all levels within a group are those in which the average item difficulty is

near 50 percent....Items which practically no one passes are of little value, since they do not tell much about individual differences. Items which everyone passes give no information about differences....

CHAPTER II

THE METHODS OF INVESTIGATION

After the APQ had been administered the writer was given 517 of the papers, about 83 percent of the total number. Only 208 of these 517 subjects, less than 40 percent, found time to attempt all the items. The reason is that it was felt that most of the students could complete the test in twenty minutes and, consequently, this amount of time was allowed.

Two of the 208 who completed the test filled in more than K (67) answers. These two scores were discarded. Since this investigation was concerned only with the scores of those who finished the test, for purposes here it may still be considered a power test.

In doing an item analysis a power test must be used. Test construction specialists (2, 8, 10, 14, 16) believe that the measurement of reliability for a test administered once can be achieved relatively satisfactorily on a power test but is considered totally inadequate on a speed test. Mollenkopf (14,312) says:

> In the setting of time limits for try-out forms, whenever it is highly desirable to secure useful information about the characteristics of every item, the test worker should allow adequate time for at least half of the group (preferably more) to attempt every item in the test.

Gulliksen (10,367) remarks on the same theme:

> For a speed test, 'proportion of correct responses' does not represent a characteristic

of the item; hence this type of analysis (item analysis) is inappropriate insofar as a test is speeded.

Because less than half of the subjects managed to complete the APQ, a methodological problem arose that had to be dealt with before item analysis could be begun. The items unattempted by those who did not complete the test were most often the items appearing late in the test. Since nothing was subtracted for guessing, the more items a subject attempted the higher his score was likely to be. Unattempted items were thus treated the same way as items answered incorrectly. Consequently, an item analysis of the APQ, using all the scores, could only result in making the items that appeared late in the test seem to be more difficult for the subjects than they actually were.

Even if only the number of subjects who attempted each item were considered, the methodological problem would not be eliminated. This is true because a different number of subjects would attempt each item with the result that the data would not be statistically comparable.

The items were analyzed by three different methods. Only one of these was to be considered seriously; the other two were much less refined techniques and were employed merely for comparative purpose in order to test their utility.

The procedure that follows is common to all three

methods. The methods peculiar to one technique will be discussed later.

A frequency distribution was set up. From this was found the mean, median, standard deviation, standard error of measurement, skewness, split-half reliability and corrected reliability of the APQ.

The scores were then divided into octiles, A, B, C, D, E, F, G, H. Group A represents the subjects who performed best on the APQ, Group B next best, and so forth.

Eight sub-groups were used because such a number is small enough to make legitimate statistical comparisons possible. Also, eight sub-groups is sufficiently large so that curves whose slopes describe item difficulty can be drawn.

In trying to set up these octiles, however, another methodological problem arose. If the distribution of scores had been rectangular, rather than normal, there would have been no problem. This is true because 12½ percent of the scores could be placed in each octile in a rectangular distribution and yet the distance along the base line for each group, or the z score discrepancy between each adjacent group, would be equal.

However, if the same percentage of scores is placed in each octile in a normal distribution, then the middle groups will occupy a lesser distance along the base line than the

outer groups. Under such circumstances the differences between means of adjacent groups would vary considerably with the result that statistical comparison of such groups would be unwarranted.

On the other hand, if the distance along the base line of each group is made equal, then Groups A and H will be represented by so few subjects that legitimate comparisons of these groups with others cannot be made.

By a process of trial and error, a method was arrived at that seems to be somewhat more appropriate compromise for the data than the two methods just discussed. He began the conventional way by letting the mean equal the point separating the two middle groups, D and E. He then let Group D include all those z scores between .00 and .37. Groups B, C, F and G were also given a distance along the base line of .37SD. Group A was given no upper limit and Group H no lower one.

This appeared a fair compromise between both problems mentioned above. No less than eighteen scores, or nine percent of the scores, fell in any group with the result that the number in each was sufficiently large to compare the octiles statistically. Also, the differences between the means of all adjacent groups, except the difference between the means of Groups G and H, were equal to about three and one-half. The negative skewness of the distrib-

ution caused the mean of Group H to be much lower than that of Group G.

The next step was to construct a total sheet consisting of 8 times 67 or 536 cells. Each cell showed the number of a particular group that answered a particular item incorrectly. The totals for each of the 67 columns were also found. Each of these values represented the total number of subjects answering the particular item incorrectly.

From this data was constructed a similar total sheet of 536 cells. Each of these cells represented the proportion of a given group answering a given item incorrectly. The totals for each of the 67 columns then found represented the proportion answering each item incorrectly, q_g.

The procedure presented thus far is common to all three methods of item analysis that were employed. The next step was to find data that was not contributory to all three methods.

Sixty-seven graphs were drawn, one for each item. Each graph showed the proportion of each group answering the item incorrectly. The A through H groups were plotted from left to right, respectively, along the X axis. The proportion answering the item incorrectly was plotted along the Y axis.

The two axes were made equal in length. This was done in order that the model item would have a slope of 1.00.

The slope of each item was found by drawing a straight line on each graph so as to represent an average of the eight points that had been plotted. In this manner, the tangent of each line represented the slope of the proportion of each group answering the item in question incorrectly. Consequently, the greater the algebraic value of the slope for an item, the better the item. This is true since the good items on a test are those on which the better groups answer less often incorrectly than do the poorer groups.

According to this method of item analysis, the items to be deleted would be those of either negative or very low positive slopes.

There is another method of item analysis in which these 67 graphs are also used. In this procedure a comparison was made between the performance of each group on the item with the group just superior to it in test performance.

Another method of item analysis to be discussed follows Gulliksen (10,363), with some variations.

The steps in this method are:

1. To find what proportion of the subjects answered each item correctly, p_g.

2. The standard deviation of each item is computed by the formula: $s_g = (p_g - p_g^2)^{\frac{1}{2}}$.

3. Gulliksen's method does not require the sub-grouping

of subjects. However, since the writer had already done so in order to use two other methods of item analysis too, and because he felt much time could be saved, he set out to discover some formula that would permit him to perform the next step, that of obtaining the point biserial correlation of each item with the total test score, by employing his eight subgroup means. The formula he derived is:

$$M_g = \frac{M_A CN_{A_g} + M_B CN_{B_g} \cdots M_H CN_{H_g}}{CN_g}$$

4. These results were then substituted in the following formula for determining the reliability index for each item:

$$r_{M_g s_g} = P_g \frac{M_g - M}{SD}$$

5. The point biserial correlation of each item with the total test score was then found simply by dividing each reliability index by the standard deviation for the item.

6. The final step involved finding the standard error of the biserial correlation for each item. The work here follows Garrett (7,347-353), rather than Gulliksen.

$$SE_{r_{bis.}} = \frac{\frac{(pq)^{\frac{1}{2}}}{z} - r^2_{bis.}}{N^{\frac{1}{2}}}$$

From these values the significance of the point biserial correlation for each item was found. All items, whose correlations could have occurred by chance at least one time out of one hundred, were deleted.

CHAPTER III

RESULTS AND DISCUSSION

After the 206 scores were chosen, a frequency distribution was set up. These results appear in Table II. Statistics, such as the mean, median, standard deviation, standard error of measurement, skewness, split-half reliability and Spearman-Brown corrected reliability of the APQ, were also found. These results are shown in Table III.

The scores were then divided into octiles or eight groups. The score limits of the octiles, and frequencies of each score, are shown in Table II. The octile means and the number of subjects in each octile are summarized in Table III.

Next, a total sheet was constructed that showed the number, IN_{gh}, and percentage $I\%_{gh}$, of incorrect responses of each octile on each item. The total number of subjects that answered each item incorrectly, IN_g, and the percent of subjects that answered each item incorrectly, $I\%_g$, were also computed. These results are found in Table IV.

Table IV is read in the following manner: Under the column headed "Octile A" and the row labeled "Item 1," and across from "IN_{gh}", appears the number "2". This means that two of the subjects in Octile A answered Item 1 incorrectly. Since 21 subjects, as shown in Table III, were placed in Octile A about 10 percent of these subjects answered Item 1

TABLE II

NUMBER OF STUDENTS RECEIVING EACH RAW SCORE ON THE APQ AND CORRESPONDING GROUP

Score	Frequency	Group	Octile
65	1	A	Highest
64	2	A	
63	1	A	
61	5	A	
60	5	A	
59	7	A	
58	13	B	2
57	13	B	
56	9	B	
55	2	C	3
54	11	C	
53	8	C	
52	7	C	
51	9	D	4
50	13	D	
49	10	D	
48	10	E	5
47	7	E	
46	6	E	
45	5	E	
44	7	F	6
43	6	F	
42	5	F	
41	8	G	7
40	5	G	
39	3	G	
38	2	G	

TABLE II (Continued)

Score	Frequency	Group	Octile
37	4	H	Lowest
36	6	H	
35	3	H	
34	1	H	
31	2	H	
30	5	H	
29	2	H	
28	1	H	
17	1	H	
16	1	H	
	N = 206		

TABLE III

ESSENTIAL STATISTICS

Number in sub groups		Mean of sub groups	
N_A equals	21	M_A equals	60.67
N_B	35	M_B	57.11
N_C	28	M_C	53.29
N_D	32	M_D	49.97
N_E	28	M_E	46.79
N_F	18	M_F	43.11
N_G	18	M_G	40.06
N_H	26	M_H	32.08

Number	206
Mean	48.57
Standard deviation	9.51
Standard error	1.27
Skewness	-.38
Reliability (split-half)	.964
Corrected reliability	.982
Median	49.77

TABLE IV

NUMBER, IN_{gh}, AND PERCENTAGE, $I\%_{gh}$, OF EACH OCTILE ANSWERING EACH ITEM INCORRECTLY

Items		OCTILES A	B	C	D	E	F	G	H	Totals IN_g	$I\%_g$
1.	IN_{gh}	2	1	6	2	0	2	1	4	18	
	$I\%_{gh}$	10	3	21	6	0	11	6	15		9
2.	IN_{gh}	2	7	3	5	6	6	8	10	47	
	$I\%_{gh}$	10	20	11	16	21	33	44	38		23
3.	IN_{gh}	0	2	4	4	9	3	11	17	50	
	$I\%_{gh}$	0	6	14	12	32	17	61	65		24
4.	IN_{gh}	0	2	2	5	9	6	7	17	48	
	$I\%_{gh}$	0	6	7	16	32	33	39	65		23
5.	IN_{gh}	1	4	3	4	16	9	9	18	64	
	$I\%_{gh}$	5	11	11	12	57	50	50	69		31
6.	IN_{gh}	4	13	12	21	16	10	12	22	110	
	$I\%_{gh}$	19	37	43	66	57	56	67	85		53
7.	IN_{gh}	0	3	6	9	7	6	7	13	51	
	$I\%_{gh}$	0	9	21	28	25	33	39	50		25
8.	IN_{gh}	2	11	9	14	18	11	13	20	98	
	$I\%_{gh}$	10	31	32	44	64	61	72	77		48
9.	IN_{gh}	0	7	12	20	21	13	16	23	112	
	$I\%_{gh}$	0	20	43	62	75	72	89	88		54

TABLE IV (Continued)

Items		OCTILES A	B	C	D	E	F	G	H	Totals IN_g	$I\%_g$
10.	IN_{gh}	12	25	23	30	26	17	17	21	171	
	$I\%_{gh}$	57	71	82	94	93	94	94	81		83
11.	IN_{gh}	0	0	0	1	0	0	0	5	6	
	$I\%_{gh}$	0	0	0	3	0	0	0	19		3
12.	IN_{gh}	0	0	2	2	2	5	6	7	24	
	$I\%_{gh}$	0	0	7	6	7	28	33	27		12
13.	IN_{gh}	0	1	2	0	4	3	2	7	19	
	$I\%_{gh}$	0	3	7	0	14	17	11	27		9
14.	IN_{gh}	0	2	2	4	5	2	5	14	34	
	$I\%_{gh}$	0	6	7	12	18	11	28	54		17
15.	IN_{gh}	1	2	2	3	1	2	4	7	22	
	$I\%_{gh}$	5	6	7	9	4	11	22	27		11
16.	IN_{gh}	0	2	2	6	10	10	14	19	63	
	$I\%_{gh}$	0	6	7	19	36	56	78	73		31
17.	IN_{gh}	2	6	5	11	13	7	12	16	72	
	$I\%_{gh}$	10	17	18	34	46	39	67	62		35
18.	IN_{gh}	5	3	4	7	6	6	6	9	46	
	$I\%_{gh}$	24	9	14	22	21	33	33	35		22
19.	IN_{gh}	5	1	6	11	4	8	9	14	58	
	$I\%_{gh}$	24	3	21	34	14	44	50	54		28

TABLE IV (Continued)

Items		OCTILES A	B	C	D	E	F	G	H	Totals IN_g	$I\%_g$
20.	IN_{gh}	0	0	0	0	1	0	2	2	5	
	$I\%_{gh}$	0	0	0	0	4	0	11	8		2
21.	IN_{gh}	2	7	12	13	8	10	12	20	84	
	$I\%_{gh}$	10	20	43	41	29	56	67	77		41
22.	IN_{gh}	2	6	6	15	14	7	11	17	78	
	$I\%_{gh}$	10	17	21	47	50	39	61	65		38
23.	IN_{gh}	3	10	10	10	17	6	10	17	83	
	$I\%_{gh}$	14	29	36	31	61	33	56	65		40
24.	IN_{gh}	1	7	6	10	12	9	16	20	81	
	$I\%_{gh}$	5	20	21	31	43	50	89	77		39
25.	IN_{gh}	0	3	1	10	8	6	10	16	54	
	$I\%_{gh}$	0	9	4	31	29	33	56	62		26
26.	IN_{gh}	12	19	18	17	14	14	11	11	116	
	$I\%_{gh}$	57	54	64	53	50	78	61	42		56
27.	IN_{gh}	4	10	13	16	18	11	13	21	106	
	$I\%_{gh}$	19	29	46	50	64	61	72	81		51
28.	IN_{gh}	0	2	1	2	3	5	6	13	32	
	$I\%_{gh}$	0	6	4	6	11	28	33	50		16
29.	IN_{gh}	11	22	19	27	22	18	15	22	156	
	$I\%_{gh}$	52	63	68	84	79	100	83	85		76
30.	IN_{gh}	0	3	1	4	6	2	2	4	22	
	$I\%_{gh}$	0	9	4	12	21	11	11	15		11

TABLE IV (Continued)

Items		OCTILES A	B	C	D	E	F	G	H	Totals IN_g	$I\%_g$
31.	IN_{gh}	2	3	4	5	0	5	2	11	32	
	$I\%_{gh}$	10	9	14	16	0	28	11	42		16
32.	IN_{gh}	0	1	4	7	8	3	4	16	43	
	$I\%_{gh}$	0	3	14	22	29	17	22	62		21
33.	IN_{gh}	3	6	5	7	10	8	7	16	62	
	$I\%_{gh}$	14	17	18	22	36	44	39	62		30
34.	IN_{gh}	0	1	2	2	3	2	1	3	14	
	$I\%_{gh}$	0	3	7	6	11	11	6	12		17
35.	IN_{gh}	0	0	5	7	6	5	5	10	38	
	$I\%_{gh}$	0	0	18	22	21	28	28	38		18
36.	IN_{gh}	3	5	5	12	7	4	6	15	57	
	$I\%_{gh}$	14	14	18	38	25	22	33	58		28
37.	IN_{gh}	2	9	13	12	15	11	13	21	96	
	$I\%_{gh}$	10	26	46	38	54	61	72	81		47
38.	IN_{gh}	0	1	2	2	4	2	4	13	28	
	$I\%_{gh}$	0	3	7	6	14	11	22	50		14
39.	IN_{gh}	2	1	5	6	2	1	2	11	30	
	$I\%_{gh}$	10	3	18	19	7	6	11	42		15
40.	IN_{gh}	1	1	1	1	1	1	0	9	15	
	$I\%_{gh}$	5	3	4	3	4	6	0	35		7
41.	IN_{gh}	3	7	7	13	9	14	10	16	79	
	$I\%_{gh}$	14	20	25	41	32	78	56	62		38

TABLE IV (Continued)

Items		OCTILES A	B	C	D	E	F	G	H	Totals IN_g	$I\%_g$
42.	IN_{gh}	1	2	7	9	7	8	4	17	55	
	$I\%_{gh}$	5	6	25	28	25	44	22	65		27
43.	IN_{gh}	1	4	9	6	5	5	5	19	54	
	$I\%_{gh}$	5	11	32	19	18	28	28	73		26
44.	IN_{gh}	1	4	10	11	10	11	8	18	73	
	$I\%_{gh}$	5	11	36	34	36	61	44	69		35
45.	IN_{gh}	1	12	7	15	11	11	10	17	84	
	$I\%_{gh}$	5	34	25	47	39	61	56	65		41
46.	IN_{gh}	2	7	9	11	11	10	8	15	73	
	$I\%_{gh}$	10	20	32	34	39	56	44	58		35
47.	IN_{gh}	8	14	12	18	17	10	9	19	107	
	$I\%_{gh}$	38	40	43	56	61	56	50	73		52
48.	IN_{gh}	1	0	3	3	4	3	5	11	30	
	$I\%_{gh}$	5	0	11	9	14	17	28	42		15
49.	IN_{gh}	0	1	0	1	0	0	0	3	5	
	$I\%_{gh}$	0	3	0	3	0	0	0	12		2
50.	IN_{gh}	1	1	1	1	0	1	1	4	10	
	$I\%_{gh}$	5	3	4	3	0	6	6	15		5
51.	IN_{gh}	1	2	1	1	1	0	1	4	11	
	$I\%_{gh}$	5	6	4	3	4	0	6	15		5
52.	IN_{gh}	0	1	1	1	1	1	2	8	15	
	$I\%_{gh}$	0	3	4	3	4	6	11	31		7

TABLE IV (Continued)

Items		OCTILES A	B	C	D	E	F	G	H	Totals IN_g	$I\%_g$
53.	IN_{gh}	1	1	1	3	0	3	2	7	18	
	$I\%_{gh}$	5	3	4	9	0	17	11	27		9
54.	IN_{gh}	3	10	6	6	7	9	8	11	60	
	$I\%_{gh}$	14	29	21	19	25	50	44	42		29
55.	IN_{gh}	2	2	5	6	4	5	4	6	34	
	$I\%_{gh}$	10	6	18	19	14	28	22	23		17
56.	IN_{gh}	2	1	0	3	5	2	3	8	24	
	$I\%_{gh}$	10	3	0	9	18	11	17	31		12
57.	IN_{gh}	0	2	0	1	6	0	4	4	17	
	$I\%_{gh}$	0	6	0	3	21	0	22	15		8
58.	IN_{gh}	2	1	2	0	4	4	2	12	27	
	$I\%_{gh}$	10	3	7	0	14	22	11	46		13
59.	IN_{gh}	3	6	7	15	15	12	12	20	90	
	$I\%_{gh}$	14	17	25	47	54	67	67	77		44
60.	IN_{gh}	2	4	4	3	6	8	10	15	52	
	$I\%_{gh}$	10	11	14	9	21	44	56	58		25
61.	IN_{gh}	0	5	5	10	12	7	9	19	67	
	$I\%_{gh}$	0	14	18	31	43	39	50	73		33
62.	IN_{gh}	0	4	4	6	2	6	5	15	42	
	$I\%_{gh}$	0	11	14	19	7	33	28	58		20
63.	IN_{gh}	3	6	4	11	10	3	10	16	63	
	$I\%_{gh}$	14	17	14	34	36	17	56	62		61

TABLE IV (Continued)

Items		OCTILES A	B	C	D	E	F	G	H	Totals IN_g	$I\%_g$
64.	IN_{gh}	0	4	5	2	17	9	9	21	67	
	$I\%_{gh}$	0	11	18	6	61	50	50	81		33
65.	IN_{gh}	3	14	14	15	22	11	13	19	111	
	$I\%_{gh}$	14	40	50	47	79	61	72	73		54
66.	IN_{gh}	7	18	16	20	20	12	11	21	125	
	$I\%_{gh}$	33	51	57	62	71	67	61	81		61
67.	IN_{gh}	1	4	6	10	8	9	9	12	59	
	$I\%_{gh}$	5	11	21	31	29	50	50	46		29

incorrectly. This number, 10, appears directly beneath the number, 2, in the same column and next row, and across from $I\%_{gh}$. By the same process, it can be seen that four students in Octile B, or 11 percent, answered Item 5 incorrectly.

The column headed "Totals" concerns itself with the total number of failures on each item. Thus, the table shows that a total of 18 of the 206 subjects, or 9 percent, answered Item 1 incorrectly.

The results thus far mentioned in this chapter are common to all three methods of item analysis in this investigation. These methods, in the chronological order in which they were developed, are: The Slope Method, the Octile Discriminability Method, and the Probability Method. The last is the most refined method employed and is treated most seriously.

The Slope Method

The next step was to draw 67 graphs, one for each item. Each graph showed the proportion of each group answering the item incorrectly. The A through H groups were plotted from left to right, respectively, along the X axis. The proportion answering the item incorrectly was plotted along the Y axis.

The two axes were made equal in length. This was done in order that the model item would have a slope of 1.00.

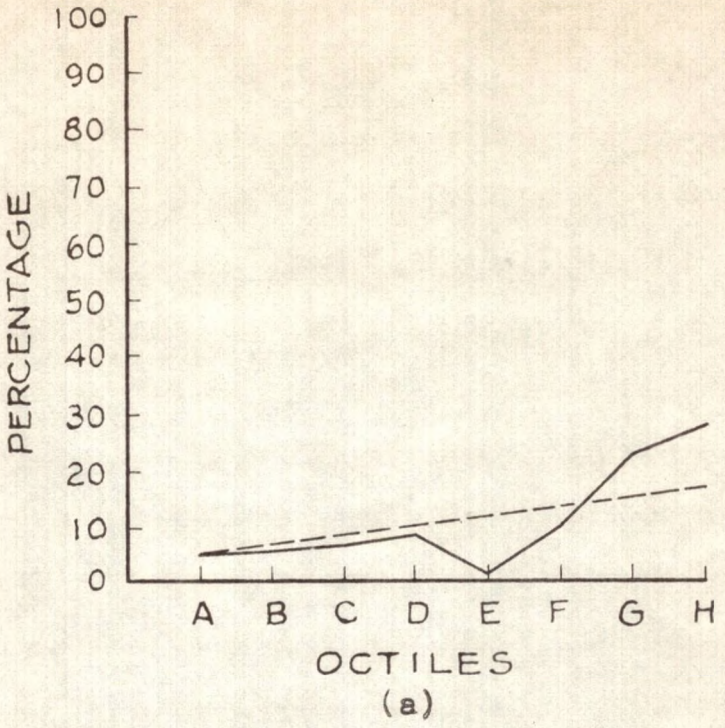
The slope of each item was found by drawing a straight line on each graph so as to represent an average of the eight points that had been plotted. In this manner, the tangent of each line represented the slope of the proportion of each group answering the item in question incorrectly. Consequently, the greater the algebraic value of the slope for an item, the better the item. This is true since the good items on a test are those on which the better groups answer less often incorrectly than do the poorer groups.

According to this method of item analysis, the items to be deleted would be those of either negative or very low positive slopes. These values are presented in column A of Table V. Graphs of two contrasting items are shown in Figure I.

It can be seen from column A of Table V that the poorest fifteen items indicated by this method have slope values of less than .17. These fifteen items are: 1, 11, 13, 15, 18, 26, 34, 40, 49, 50, 51, 53, 55, 56, and 57.

The Octile Discriminability Method

These graphs were also employed in order to obtain data for the Octile Discriminability Method. In this procedure a comparison was made between the performance on each item of each octile with the octile just superior to it in test performance. For example, on Item 1, Group B did better than Group A so a value of minus one was given.



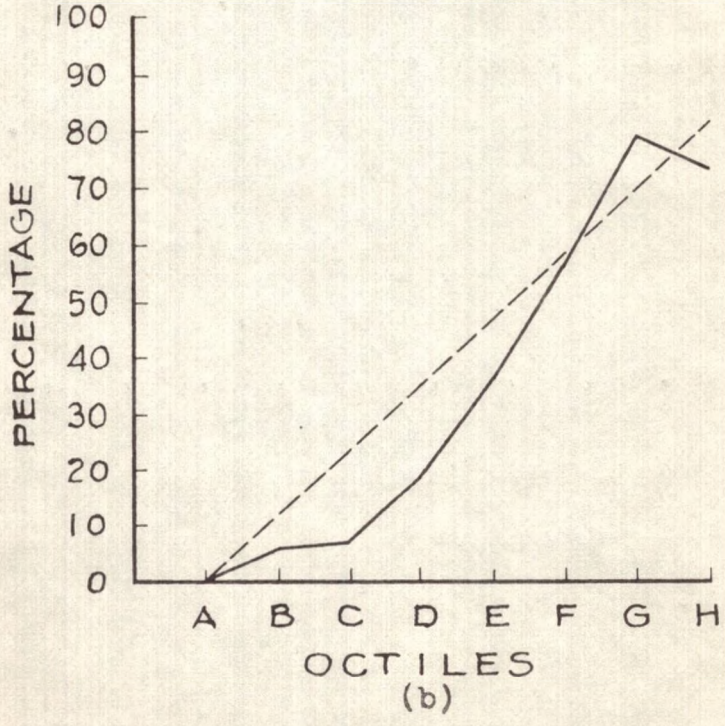


FIGURE I. GRAPHS OF THE PERCENTAGE OF FAILURES IN EACH OCTILE: (a) for Item 15; (b) for Item 16.

TABLE V

ITEM ANALYSIS INFORMATION FOR SLOPE METHOD AND OCTILE DISCRIMINABILITY METHOD

A. Item Discriminability by Slope Based on Octiles.

B. Item Discriminability by Octiles.

Item	A	B	Item	A	B	Item	A	B
1.	.00	-.14	24.	.82	+.71	47.	.23	+.43
2.	.35	+.43	25.	.55	+.71	48.	.32	+.43
3.	.64	+.43	26.	.04	-.43	49.	.06	+.14
4.	.64	+1.00	27.	.55	+.71	50.	.07	.00
5.	.60	+.43	28.	.48	+.71	51.	.05	+.14
6.	.50	+.43	29.	.37	+.43	52.	.18	+.71
7.	.42	+.71	30.	.20	+.29	53.	.15	+.14
8.	.62	+.71	31.	.20	+.14	54.	.30	-.14
9.	.88	+.43	32.	.50	+.71	55.	.16	+.14
10.	.22	+.14	33.	.43	+.71	56.	.13	+.14
11.	.05	+.14	34.	.10	+.29	57.	.15	+.14
12.	.20	+.29	35.	.35	+.43	58.	.30	+.14
13.	.13	+.43	36.	.35	+.29	59.	.70	+.86
14.	.38	+.71	37.	.65	+.71	60.	.42	+.71
15.	.12	+.71	38.	.38	+.43	61.	.68	+.71
16.	.82	+.71	39.	.20	+.14	62.	.42	+.43
17.	.60	+.43	40.	.10	+.14	63.	.37	+.43
18.	.13	+.29	41.	.53	+.43	64.	.70	+.29
19.	.30	+.43	42.	.45	+.43	65.	.63	+.43
20.	.20	.00	43.	.53	+.29	66.	.32	+.43
21.	.37	+.43	44.	.60	+.43	67.	.50	+.29
22.	.33	+.71	45.	.58	+.14			
23.	.32	+.43	46.	.47	+.71			

$r_{AB} = .62$

However, Group C did poorer than Group B so a value of plus one was added to this. Seven such values were added for each item; there were seven values because there were eight groups. The sum of values for Item I was minus one and the discriminability of Item 1 by groups could then be expressed as - 1/7 or -.14.

In this method too, the greater the algebraic value, the better is the item. These values are given in Table V, column B.

Eighteen items were found to have discriminability values of less than .17. These items are: 1, 10, 11, 20, 26, 31, 39, 40, 45, 49, 50, 51, 53, 54, 55, 56, 57, and 58.

Thus, the value of .17 was used as the cut-off point for both methods discussed above. In these two methods, the model item would be found to have a value of 1.00. These two methods, however, are simple and were used merely for comparative purpose in order to test their utility.

The Probability Method

The main method of item analysis used, the Probability Method, follows Gulliksen (10,363), with some variations. All procedure already discussed, up to the time when the 67 graphs were drawn, was employed in this method. The data for this procedure common to all three methods are found in Tables II, III, and IV. The data and discussion of Table V and Figure I made no contribution to the

Probability Method. In completing the discussion of the development of results for the main method, all results referred to are given in Table VI.

The column headed "p_g" in this table refers to the proportion of subjects who answered each item correctly. Since the percentage of subjects who answered each item incorrectly, $I\%_g$, was already known (Table IV), each percentage was divided by one hundred. This process gave the proportion of subjects who answered each item incorrectly, q_g. Thus, p_g could be obtained by the simple formula: $p_g = 1.00 - q_g$.

Next, the standard deviation of each item was computed by the formula: $s_g = (p_g - p_g^2)^{\frac{1}{2}}$. These results are given in Table VI.

The following step was to find the reliability index for each item, $r_{xg}s_g$. (The formula used and the methodological problem involved were discussed in the previous chapter.) The point-biserial correlation of each item with the total test score was then found simply by dividing the reliability index for each item by the standard deviation of each item, s_g. The values of both these statistics are also found in Table VI.

Next, the standard error of the biserial correlation for each item was found. (This formula appears at the end of the previous chapter.) The correlation coefficient of

TABLE VI

ITEM ANALYSIS INFORMATION FOR PROBABILITY METHOD

g	p_g	s_g	r_{xg}	r_{xgsg}	P
1	.91	.28	.039	.011	NS
2	.77	.42	.207	.087	.01
3	.76	.42	.450	.189	.01
4	.77	.42	.433	.182	.01
5	.69	.45	.438	.197	.01
6	.47	.50	.318	.159	.01
7	.75	.44	.293	.129	.01
8	.52	.50	.356	.178	.01
9	.46	.50	.494	.247	.01
10	.17	.37	.168	.062	.05
11	.97	.17	.241	.041	.05
12	.88	.33	.285	.094	.01
13	.91	.28	.254	.071	.01
14	.83	.37	.359	.133	.01
15	.89	.32	.191	.061	.05
16	.69	.45	.531	.239	.01
17	.65	.48	.331	.159	.01
18	.78	.41	.161	.066	.05
19	.72	.45	.280	.126	.01
20	.98	.14	.193	.027	NS
21	.59	.49	.355	.174	.01
22	.62	.49	.322	.158	.01
23	.60	.49	.257	.126	.01
24	.61	.49	.443	.217	.01
25	.74	.44	.409	.180	.01
26	.44	.50	-.040	-.020	NS
27	.49	.50	.342	.171	.01
28	.84	.36	.386	.139	.01
29	.24	.42	.217	.091	.01
30	.89	.31	.113	.035	NS
31	.84	.36	.211	.076	.01
32	.79	.41	.363	.149	.01
33	.70	.46	.293	.135	.01
34	.93	.26	.104	.027	NS
35	.82	.39	.279	.109	.01
36	.72	.45	.240	.108	.01
37	.53	.50	.368	.184	.01
38	.86	.35	.357	.125	.01
39	.85	.36	.200	.072	.01
40	.93	.26	.254	.066	.01

TABLE VI (Continued)

g	p_g	s_g	r_{xg}	r_{xgsg}	P
41	.62	.49	.308	.151	.01
42	.73	.45	.676	.304	.01
43	.74	.44	.343	.151	.01
44	.65	.48	.358	.172	.01
45	.59	.49	.288	.141	.01
46	.65	.48	.265	.127	.01
47	.48	.50	.180	.090	.01
48	.85	.36	.306	.110	.01
49	.98	.14	.150	.021	NS
50	.95	.22	.123	.027	NS
51	.95	.22	.095	.021	NS
52	.93	.26	.277	.072	.01
53	.91	.28	.214	.060	.05
54	.71	.46	.170	.078	.01
55	.83	.37	.130	.048	NS
56	.88	.33	.215	.071	.01
57	.92	.26	.181	.047	.05
58	.87	.33	.312	.103	.01
59	.56	.50	.388	.194	.01
60	.75	.44	.355	.156	.01
61	.67	.47	.398	.187	.01
62	.80	.40	.338	.135	.01
63	.69	.46	.289	.133	.01
64	.67	.47	.477	.224	.01
65	.46	.50	.286	.143	.01
66	.39	.49	.204	.100	.01
67	.71	.46	.276	.127	.01

each item with the total test score was then divided by the standard error of the item. From this, the statistical significance of each item, P, was then determined. The table employed for estimating P was found in Garrett (7,299). These results are also shown in Table VI.

The items with correlations which could have occurred by chance at least one time out of 100, i.e., those correlations which have P values greater than .01, were considered for possible deletion. Items 1, 10, 11, 15, 18, 20, 26, 30, 34, 49, 50, 51, 53, 55 and 57 fell into this category.

Before deciding whether or not to delete them, however, comparisons were also made with the data obtained by the other two methods of item analysis.

Comparison of the Three Methods

Twelve of the fifteen items chosen for deletion by the Probability Method were among the poorest fifteen items as determined by the Slope Method. (Items 10, 20 and 30 were chosen for removal by the main method but not by the Slope Method. Items 13, 40, and 56 were not considered for deletion by the Probability Method but were among the fifteen poorest found by the Slope Method.) Consequently, the Slope Method agreed with the items considered for deletion by the main method in eighty percent of the cases, or seventy-four percent above chance prediction.

Eleven of the fifteen items chosen for deletion by the Probability Method were among the poorest eighteen items as determined by the Octile Discriminability Method. (Items 15, 18, 30 and 34 were selected for removal by the Probability Method, but not by the other. Items 31, 39, 40, 45, 54, 56 and 58 were not considered for deletion by the Probability Method but were among the eighteen poorest items found by the Octile Discriminability Method.) Therefore, the latter method agreed with the items considered for deletion by the Probability Method in seventy-four percent of the cases, or forty-seven percent above chance prediction.

The Pearson product-moment correlation between the Slope Method and the Octile Discriminability Method was found to be .62. While these two methods, especially the latter, are crude, the above discussion should indicate that they do predict the poorest items with fairly substantial accuracy.

What is most important here, however, is what the main method showed. It revealed that fifteen items did not correlate with the APQ to sufficient extent to make the probability that these item-test correlations occurred by chance less than one out of 100.

None of these fifteen items has an algebraic value greater than .22 when the Slope Method of item analysis is used. This is shown in column A of Table III. Since the

perfect item would have a slope of 1.00, one could say that the slopes of none of these items excel twenty-two percent of this ideal. Furthermore, most of the remaining items have substantially greater slopes than the above fifteen.

Column B of Table III reveals that only one of the fifteen items has an algebraic value by the Octile Discriminability Method of more than .29. Because the item that discriminates perfectly by groups would have a discriminative value of 1.00, one could also infer that only one of the fifteen items discriminates by octiles with more than twenty-nine percent accuracy. This is Item 15. The discriminability of this item by octiles is .71.

While this figure is very high, it is easily explained. Even though the item-test correlation of item fifteen was not statistically significant at the .01 level of confidence, it was at the .05 level. Also, the Octile Discriminability Method in no way considers the strength of the discrepancies between the performances of the groups. Account is taken only of whether or not each octile performed poorer on the item than the octile just better than it did. Since it happened that the poorer groups quite consistently did just a shade poorer on this item than the better groups, the Octile Discriminability Method is very misleading in this case.

The Slope Method, however, takes account of the amount of the discrepancies between the performances of the groups, and the slope of Item 15 is only .12.

While Item 15 seemed to be the best of those chosen for deletion, Item 1 and Item 26 seemed to be the poorest. The data from all three methods indicate that these two items are the most inadequate for separating the good students from the poor.

Since the other two methods were in good agreement with the Probability Method with respect to the fifteen items considered for deletion by the latter, and because the Probability Method is the most refined of the three, it was decided to delete the fifteen items that this method indicated. These are the items which had item-test correlations that were not statistically significant at the one percent level of confidence, that is, that had P values greater than .01. These fifteen items have already been revealed and can readily be detected by reference to the last column of Table VI.

The writer has stated that the main method of item analysis employed follows Gulliksen (10,363-395), with some modifications. One such modification was that the author did not find validity indices. Conrad (4,41) remarks:

> The item-subtest correlation ($r_{bis.}$) should, whenever possible, be supplemented by correlating each item with a valid external criterion. This is especially desirable if the subtest itself has only a low correlation with the external criterion

> (say less than .45); because in such a case, there is danger of retaining items which, although homogeneous among themselves, are only slightly related to the external criterion which the test aims to measure. Similarly, it is desirable to correlate each item with a valid external criterion when the item-subtest correlations tend to be low (median $r_{bis.}$ below .45); because in this case, there is inadequate assurance from the item-subtest correlations that the items are sufficiently meritorious to be worth retaining. Knowledge of the correlations with the external criterion may also help to improve the homogeneity of the test.

Gulliksen (10,380-381), indicates, however, that item analysis can be done without the use of validity indices:

> In most practical cases it is probable that selecting items to increase the reliability of a test will also incidentally increase test validity....In other words, if no criterion is available it is highly desirable to take steps to increase test reliability....

Gulliksen (10,363-395) computed the validity indices and the point biserial correlation of each item with the criterion score, which in the present study would be the grade point average. Then Gulliksen graphed the validity indices on the Y axis and the reliability indices on the X axis. He chose those items for deletion that fell in or nearest the fourth quadrant.

Instead of deleting items on this basis, it was decided to delete them on the basis of their statistical significance, with consideration also for the slopes of the items and their discriminability by groups.

Many test constructors, including (10, 4, 5), prefer items of fifty percent difficulty, that is, items in which the value of p_g is proximate to that of q_g. One reason they give priority to such items is that more statistical significance can be attached to the correlations of these items with the test itself and the criterion. Gulliksen (10,374-375) writes:

> There have been several empirical studies that show that tests composed of items answered correctly by about fifty percent of the group have a higher validity than tests composed of items that are easier or harder than fifty percent, but otherwise of the same type....it is suggested that the higher validity found for tests composed of items with fifty percent difficulty may be due to and directly measured by the increase in item-criterion correlation.

The APQ was designed to distinguish those minority of subjects incapable of doing college level work. As a result, it is composed of many easy items, that is, items with considerably less than fifty percent difficulty. Consequently, the nature of this test is such that construction of valid and reliable items becomes exceedingly difficult.

Furthermore, Conrad (4) feels that data such as that obtained in this investigation do not call for a full item analysis. He (4,47) states, "Full item analysis should not generally be applied to tests which, by the evidence of a high reliability coefficient (over .90), are already highly homogeneous." (The corrected reliability coeffic-

ient for the APQ was found to be .98. See Table II.) Conrad (4,47) continues, "The size of the sample taking the experimental form of a group-test should be large, never less than 500, and preferably larger." (It will be recalled that in the present study the test scores of 517 subjects were available but only 206 of them were used. Also (4,45), "The PE (probable error) of $r_{bis.}$ rises sharply as p rises above 80 or falls below 20. This is another limiting factor in the case of very easy or very difficult items." (Twenty-five items of the APQ were found to have a p above eighty percent. This is shown in the second column of Table IV.)

Various factors can affect the item-test and/or item-criterion correlations. One such factor is whether or not the test is one of speed or power. The APQ was intended to be a power test. However, the time allotment proved too short for the majority to attempt all the items. Uncompleted tests were discarded. Had they been used in this sample, the reliability of the test could not have been adequately appraised.

About the only adequate method that could be used for testing and reliability of the APQ was the split-half, coefficient of equivalence method. According to Cronbach (5,69), this method is justified only if the halves are equivalent. He (5,70) says, "In a speed test the reliability

coefficient by the split-half method would be falsely high." He also states that for other tests, the coefficient is too low if the halves are not equivalent. Such is the situation here. Therefore, the Spearman-Brown correction formula was used to raise the reliability coefficient to the proper estimate of what it should be. (The type of split-half method employed was the odd-even.)

Why the reliability coefficient is falsely high in a speed test is explained by Conrad (4,21):

> Suppose now, that a given individual answers a certain item of a subtest incorrectly; not only does the individual lose credit on the item which he answered wrong, but he has comparatively less time in which to answer the remaining items of the subtest. Those persons who pass the item, on the other hand, not only obtain credit for this particular item, but also have more time to attempt later items; thus, those passing the item gain an advantage over those who failed. If we multiply this advantage several fold (to take account of the fact that other items besides the particular one under discussion are also answered incorrectly), it is clear that the speed factor tends to <u>increase</u> the value of biserial <u>r</u>. This increase in biserial <u>r</u> is likely to be especially noticeable for the later items of the subtest. Other factors besides item-position which determine the extent of increase in biserial <u>r</u> are a) the degree to which speed determines scores on the subtest; and b) the correlation between speed and 'power' (i.e., ability to answer items at increasingly higher levels of difficulty-assuming that the later items of a subtest are progressively more difficult than the earlier).

Lord (13) mentions several other factors that affect the reliability of a test:

> Brogden's numerical results on test reliability confirmed Gulliksen's conclusions that the reliability of a test increases a) as the average item intercorrelation increases, b) as the dispersion of the item difficulties decreases, and c) as the mean item difficulty approaches fifty percent correct.

Changes in ability level of the group can affect the validity of the test. Gulliksen (10,393) states that M. W. Richardson, "The Relation of Difficulty to the Differential Validity of a Test," Psychometrika, 1, 33-49, has shown that systematic changes can occur in biserial correlation with changes in ability level of the group.

Of the various factors that affect the item-test correlations, one seems especially interesting. Gulliksen (10,394-395) remarks on this subject:

> There is one additional factor affecting item-test correlations that does not influence item-criterion correlations. The length of the test of which the item is a part will affect the item-test correlation but cannot influence the item-criterion correlation. For very short (two or three items) tests, the item score will form a considerable fraction of the test score; hence the item-test correlation will at first tend to decrease as items are added to the test. For tests larger than fifty or a hundred items, this effect is negligible; and, as the test length increases, a slight increase in item-test correlation could be expected because of the decrease in the error component of the total test score as test length is increased.

Thus, Gulliksen is apparently aware of a kind of error introduced in item-test correlation by virtue of the fact that the item is part of the test. He also seems aware

that the amount of this error is inversely proportional to the number of items in the test. It appears, however, that no one has ever treated this error statistically when doing an item analysis.

It appears that this error is a constant, rather than random one. Its apparent effect is to increase algebraically all item-test correlations. While the size of this error seems to be inversely proportional to the number of items in the test, it also appears to be directly proportional to the item-test correlation.

It is believed that what should be sought in item analysis data, however, is not the item-test correlations, r_{xg}. In order to properly determine how poor items are, and how they compare with each other in this respect, one must find out what the item-test correlations are when allowance is made for the fact that the items are part of the test.

A suggested formula for reducing the error mentioned above somewhat is:

$$r_{cxg} = r_{xg} - \frac{r_{xg}}{N}$$

The meaning of this formula is simply that the corrected item-test correlation (after the error has been subtracted that occurs due to the fact that the item is part of the test) equals the original item-test correlation, minus the latter, divided by the number of items in the test. This

formula may be somewhat in error and probably makes the problem look more simple than it is.

Before leaving the discussion of various factors that can affect correlations, one additional factor will be mentioned that is pertinent to the problem. This pertains to the validity of the APQ. Anderson (1,24) reported that the quintiserial correlation between the APQ and the criterion, grades of subjects, was .53 for the first semester and .30 for the second. This means that the average quintiserial correlation for the two semesters is .415.

Anderson (1,23) stated that a correlation of .45 is considered a minimum validity coefficient for a test of practical usefulness used singly. Unfortunately, he did not correlate the APQ scores with the grades for the entire school year. Had he done so, he would have probably found a correlation above the average of .415, and probably above the .45 he indicates as a minimum for practical usefulness of a test used singly. The reason is that a test will usually correlate higher with the grades for two semesters than it will with the grades for the average of the two because of the larger numbers involved in the data.

Thus, it cannot be concluded that the APQ is or is not presently practically useful when employed by itself. It will very likely become more valid, however, when the fifteen items suggested for deletion are removed.

CHAPTER IV

SUMMARY, CONCLUSIONS AND SUGGESTIONS

Summary

1. The Ability Problems Questionnaire was analyzed by three different methods of item analysis.

2. The items selected for possible deletion on the basis of the Probability Method were also evaluated by the Slope Method and Octile Discriminability Method.

3. An original formula was presented for finding reliability indices from grouped data.

4. A new formula was presented for accounting for the constant error introduced by the fact that the item is part of the test.

Conclusions

1. It was decided to delete fifteen items from the APQ: items 1, 10, 11, 15, 18, 20, 26, 30, 34, 49, 50, 51, 53, 55 and 57.

2. It was not known whether or not the APQ was practically useful for prediction when used singly. It is quite probable that the test will become such, however, after the fifteen items mentioned above have been deleted.

3. The Probability Method seemed both empirically, and in terms of logical validity, to be the superior method of the three employed. The Slope Method seemed almost as good in both respects. The Octile Discriminability Method,

however, appeared to have poor face validity and was empirically sporatic when compared with the other two.

Suggestions

1. That the APQ be again administered, after the fifteen items suggested for deletion have been removed or improved.

2. That the time limit be extended so that the vast majority of students, if not all, can complete the test. (Deletion of fifteen items, however, will have the same effect as increasing the time limit.)

3. That the test be again item analyzed, after it has been re-administered.

4. That the correction formula for the error introduced by the fact that the item is part of the test be refined and, if worthy, used in later item analysis procedure.

5. That consideration be given to the following paradox: Deletion of poor items makes the curve, for the subjects' scores that are based on only the remaining items, more platykurtic than the original curve that was based on all the items. The standard error of measurement is generally found to be larger for platykurtic curves than mesokurtic ones. The purpose of item analysis is to decrease the standard error of measurement, not increase it. Perhaps the curve for the subjects' scores that are based on only the remaining items is not the one to be considered.

Perhaps the curve should be considered instead that results from administering the shortened test to a similar group of subjects. Regardless of which curve is preferable, clarification seems necessary. Contemporary item analysis literature does not seem to deal with these probblems. It is possible that item analysis progress will be accelerated if workers in the field spend a larger proportion of their time reflecting on such matters.

BIBLIOGRAPHY

1. Anderson, Wayne, "The Reliability, Standard Error of Measurement and Validity of Entrance Tests for Freshmen at the University of North Dakota," Master's Thesis, (Aug., 1953) pp. 13-20.

2. Angoff, W. H., "Test Reliability and Effective Test Length," Psychometrika, Vol. 18, pp. 1-14.

3. Bedell, B. J., "Determination of the Optimum Number of Items to Retain in a Test Measuring a Single Ability," Psychometrika, Vol. 15, pp. 419-430.

4. Conrad, Herbert S., "Characteristics and Uses of Item-Analysis Data," Psychological Monographs, Vol. 62 (8), pp. 1-48.

5. Cronbach, Lee J., Essentials of Psychological Testing, New York: Harper and Brothers, Publishers, 1949.

6. Davis, Frederick B., "Item Analysis in Relation to Educational and Psychological Testing," Psychological Bulletin, Vol. 49 (1952) pp. 97-119.

7. Garrett, Henry E., Statistics in Psychology and Education, New York: Longmans, Green and Co., 1947.

8. Guilford, J. P., Fundamental Statistics in Psychology and Education, New York: McGraw-Hill Book Company,

9. Gulliksen, Harold, "Item Parameters Which Are Invariant with Respect to Group Ability Level," American Psychologist, Vol. 5 (1950) p. 288.

10. Gulliksen, Harold, Theory of Mental Tests, New York: John Wiley & Sons, Inc., 1950.

11. Herfinal, Orris C., "An Application of Chi-Square to the Determination of the Discriminating Power of Test Questions," Journal of Educational Psychology, Vol. 40 (1949) pp. 371-377.

12. Johnson, A. Pemberton, "Notes on a Suggested Index of Item Validity: The U-L Index," Journal of Educational Psychology, Vol. 42 (1951) pp. 499-504.

13. Lord, F. M., "The Relation of the Reliability of Multiple-Choice Tests to the Distribution of Item Difficulties," Psychometrika, Vol. 17 (1952) pp. 181-194.

14. Mollenkopf, W. G., "An Experimental Study of the Effects on Item-Analysis Data of Changing Item Placement and Test Time Limit," Psychometrika, Vol. 15 (1950) p. 312.

15. Schmid, John Jr., "Sequential Analysis of Test Items," Journal of Experimental Education, Vol. 20, (1952) pp. 261-264.

16. Thorndike, R. L., "Reliability," in Lindquist, E. F., Educational Measurement, Washington, D. C., American Council on Education, 1951.

APPENDIX