CHAPTER VII

ADMINISTRATION OF THE PILOT TEST FOR ITEM ANALYSIS

In the previous chapter, we saw, how the pilot form of the test was prepared. As is already shown in the table No. 10 , there were, in all, 143 items in the whole test. The next step was to select a sample for pilot testing. The pilot testing was done during the last week of July, 1958, and first week of August, 1958.

SAMPLE FOR PILOT TESTING

How to select the most representative sample? How large should be the sample for pilot testing? These are the two questions, an investigator should answer before finally selecting the sample for pilot testing.

The answer to the first question depends upon the purpose of the test. Our purpose is lucidly defined in chapter No. V. The purpose is to measure aptitude for teaching in secondary schools possessed by prospective teachers. Though, the test may be used in any state in India, it is primarily designed to administer to the prospective teachers in Bombay State (old). So the pilot test should be administered to a sample which is representative of the prospective teachers in the state. In the training colleges, pupil-teachers come from almost all parts of the state. It is highly justifiable, therefore, to select the sample from among the trainees in the colleges. This question was discussed with the experts also. They perfectly agreed that the sample from the trainees would be the most representative for our purpose.

As to the second question, the answer depends upon 'why we collect these data'. Naturally, the data were to be collected for item analysis. We should, then, think of the methods of item analysis we are to follow. We were to follow the most efficient method for estimating 'r'$_b$ for an item. This is Flanagan's abac. This instrument is designed for use when the middle 46 per cent of examinees on total score have been eliminated and each tail contains 27 per cent. It is recommended that the Flanagan 'r' be used when 100 cases remain in each tail, which means examining a sample of approximately 370. Obviously, then, we selected a sample of 371 prospective teachers from the five different institutions as shown below:

(1) Faculty of Education and Psychology, Baroda	109
(2) Secondary Teachers' Training College,Bombay	97
(3) Tilak College of Education, Poona	117
(4) B.T.Students of S.N.D.T.College,Baroda	21
(5) Fresh (untrained) teachers in Secondary Schools in Baroda.	27
Total:	371

INSTRUCTIONS TO THE TESTEES

After the experience of the administration of the tryout form of the test, the investigator prepared a final list of the instructions for testees. The instructions were as follows:

(1) This test-booklet contains five sub-tests. Please answer them in serial order.

(2) Each sub-test contains a number of items. Please omit no item; deal with each item as it comes. Do not leave any item with the intention of returning to it later.

(3) In each sub-test, instructions together with illustrations are given. Please read them very carefully and then answer the items. A separate answer book is provided.

(4) Please write answers in the answer book at the appropriate places. Nothing should be written in the test-booklet.

(5) Please work as rapidly as you can. When you complete your work, please return the test-booklet alongwith the answer book.

(6) Research studies are useful only when reliable and accurate data are collected. This is

possible if the answers are given with sincerity and honesty.

(7) The data will be kept strictly confidential and will be used for research only.

Over and above, these instructions, the instructions as to how to answer items in each sub-test together with illustrations were given at the top of each sub-test.

Though the written instructions are very comprehensive and self-explanatory, the following oral instructions also were given:

(1) If you have any difficulty, please ask the test administrator but do not discuss anything with your neighbours.

(2) You will be given enough time to answer all the items in the test. Do not make unnecessary haste to finish the test. Answer each item after due thinking.

(3) Please see that no item is left out. You have to answer all the items. At the end we shall check whether you have answered all the items.

(4) Your sincere and honest effort will help us a lot in our endeavour.

TIME LIMIT

As this is exclusively a power test, there should not be any time limit for taking the test and a testee should be given as much time as he/she needs. But the test administrator should know the maximum time the slowest testee takes in taking the test. As is mentioned in the previous chapter, the average time needed to answer each item is 3/4th of a minute. There were 143 items in the pilot form of the test. The average time limit, therefore, was tentatively fixed to be 108 minutes. The testees were not told of this time limit. But at the end of the testing programme, it was found that all the testees could finish the test in 110 minutes at all the centres.

The test was administered to all the testees at all centres under the direct supervision of the investigator. For this he went to different centres. He got full co-operation of the teachers at different centers. They helped actively the investigator in carrying out the testing programme. The testees also took much interest in taking the test and co-operated with the investigator in the work beyond his expectations. The testing work, thus, became much scientific, useful and smooth.

SCORING

The scoring also needs tact. It should be done very judiciously and accurately. Arthur E. Traxler[1] says, "If

[1] Traxler, A.E., A Chapter on, 'Administering and Scoring the Objective Test' in 'Educational Measurement', edited by Lindquist, E.F., Op.Cit., p.332.

scoring is to be done well, individuals need to be carefully selected, trained, supervised and checked upon systematically." This question did not arise here, as the scoring work was done throughout, by the investigator himself.

The same correct responses to different items as were selected for scoring the tryout sample, were retained here as no change was affected in the body of any item in the pilot form of the test.

After a rigorous process of item validation 40 items were rejected. So in the pilot test there were 143 items. The method of scoring the items in each sub-test was exactly the same as that followed in scoring the tryout sample. The total score obtainable in pilot testing, then, was 151.

A brief discussion of correction for chance success and how it was done in the present work is necessary here.

CORRECTION FOR CHANCE SUCCESS

The present test consists of five sub-tests. Most of the items in sub-tests 2 to 5 are of the multiple-choice type. In the sub-test, 1, the items in word analogy test are of the multiple-choice type.

When the items are of the multiple-choice type, some examinees may guess blindly among the choices presented and mark the correct answer by chance alone. To cope with this troublesome problem, a correction for chance has been proposed.

There is, however, rather sharp disagreement among test technicians regarding its appropriateness. We studied this. And we came to our own judgment. This is discussed at a later stage in this chapter.

> The principal assumption involved in making use of the conventional correction for chance success is that examinees who do not possess enough knowledge to permit them to select the correct answer will guess blindly among all the choices(correct or incorrect) in the item. The extent to which this assumption is satisfied in actual practice varies with the nature of the items and of the groups tested. To the extent that examinees choose incorrect answers on the basis of misinformation rather than on the basis of blind guessing, the procedure overcorrects for chance success; to the extent that they can eliminate from consideration incorrect choices on the basis of partial information that is correct but not adequate to permit identification of the correct choice, the procedure undercorrects for chance success.[1]

This excellent statement by Davis is reproduced here to make clear where the danger lies while applying correction for chance success and to show that the lack of analytic precision in the conventional correction for chance success is a cause of some disagreement among test constructors about whether to make use of it in computing item analysis data.

Test technicians who oppose 'applying correction for chance success' argue that blind guessing can be reduced to negligible proportions by:

[1] Davis, F.B., A Chapter on 'Item Selection Techniques' in 'Educational Measurement', edited by Lindquist, E.F., Op.Cit., p.268.

(1) warning testees against it,

(2) selecting such distractors as are equally attractive and cannot be eliminated on first sight,

(3) allowing enough time to testees to think over all the items critically.

How far these steps will reduce blind guessing, cannot be said. But it is certain that they help in reducing it to a certain extent.

Those who advocate the correction for chance success are willing to undertake the slight additional labour involved in it even when the amount of blind guessing is of small proportions because they feel that testees who ignore directions should realise that they may suffer a penalty for doing so and thus it will discourage the practice of blind guessing.

Ultimately, the present investigator accepts the opinion of Dr. Micheels and Dr. Karnes.[1] They say,

> There does not seem to be much evidence to indicate that the informal type of test is significantly improved by correcting for guessing. From a statistical standpoint, the procedure can be logically justified, but in terms of the meaning of a single test score some doubts may be raised. The authors are of the opinion that little is to be gained in using correction formulas.

[1] Micheels, W.J. and Karnes,M.R., 'Measuring Educational Achievement', McGraw - Hill Book Company, Inc.,1950, p.148.

The investigator also accepts the suggestion of Davis.[1] He says,

> It is especially important to make use of a correction for chance success in obtaining individual raw scores that are to be used for internal consistency item analysis purposes.

There are two stages at which correction formula may be employed. The test constructor may choose one of the two. He can employ the correction formula while scoring each testee. This formula is[2]:

$$S = R - \frac{KW}{n - K}$$

Where S = Score,

R = the number of right responses,

W = the number of wrong responses,

n = the number of suggested responses for a single item,

K = the number of responses to be selected or marked for each item.

As, K is always equal to 1 for the present test, the formula reduces to:

$$S = R - \frac{W}{n - 1}$$

[1] Davis, F. B., Op.Cit., p. 277.

[2] Lindquist, E. F., Op.Cit., p. 365.

Scoring by this formula involves the assumption that every right response is the result of a guess.

Or he can use the following formula[1] for computing 'Percent of Correct Responses'.

$$P_t = 100 \times \frac{R_t - \frac{W_t}{k_i - 1}}{N_t - NR_t}$$

Where P_t = the percent of correct responses in the entire sample adjusted for chance success and for omissions caused by not reaching the item in the time limit,

R_t = the number of examinees in the entire sample who answer the item correctly,

W_t = the number of examinees in the entire sample who answer the item incorrectly,

k_i = the number of choices in the item,

N_t = the number of examinees in the entire sample,

NR_t = the number of examinees in the entire sample who do not reach the item in the time limit.

If every examinee has reached every item in the time limit, NR_t becomes zero and computation of the adjusted percents is simplified accordingly.

[1] Lindquist,E.F.,Op.Cit.,p.280.

During the testing programme here, the testees were advised hot to guess and were given sufficient time to answer all the items. While constructing the test, the investigator attempted to make each incorrect response as plausible as possible to the testee who does not possess the desired knowledge or ability.

He is quite justified, then, in accepting the suggestion of Dr. Micheels and Dr. Karnes. In the present test, no correction for chance success is, therefore, applied while scoring. Only for internal consistency item analysis purposes, following Davis, correction for chance success is applied.

ITEM ANALYSIS

The effectiveness of a test depends upon the characteristics of the items which comprise it. In both its reliability and its validity a test score is the resultant of the validities, reliabilities and inter-correlations of its component items. In order to produce the most effective and useful test, therefore, we must study each one of the pool of items from which the test is to be assembled. The choice of items for the final form of a test is based in part on the detailed specifications for the content of the final test which were prepared as a part of the process of planning the test. It is based in part on certain statistical characteristics of each item. There are two statistical aspects of the individual item with which we are concerned. The first is the difficulty level of the item for the group under study. The second is the degree to which the item differentiates those who are high from those who are low on some standard. This standard may sometimes be performance on the complete pool of items, in which case we are concerned with the internal consistency of the items. The standard may sometimes be an external criterion of job performance, in which case we are concerned with the validity of each individual item.

The major objective of an item analysis, then, is to obtain objective information concerning the items we wrote for the test. This information is valuable for several reasons. It provides the opportunity to check up on the test writer's subjective judgment in selecting the items to compose the test. The test-writer also learns how testees react to items of each sub-test. In multiple-choice tests he learns which distractors are not functioning, as shown by their relative unpopularity. He learns where and how items need to be rewritten.

The most common use of the item analysis data is, therefore, in the selection of best items to compose the final test form.

Guilford[1] writes,

> There is no point in item analysis of a test that is designed as a speed test. In fact, only power tests, or

[1] Guilford, J.P., 'Psychometric Methods', McGraw Hill Book Company, Inc., New York, Toronto, London, 1954, p.418.

> those close to power tests, should be
> so treated.

He again writes, "It is more important to analyse aptitude tests than achievement tests."

Gulliksen[1] very clearly marks out the important difference between item selection procedures for aptitude tests and those for achievement tests. He says,

> In the construction of aptitude tests the item statistics may be allowed to control the rejection and selection of items more fully than in the construction of achievement tests. The judgement of the subject matter expert must always play an important part in the selection and rejection of items for an achievement test.

The present one is an aptitude test. It needs, therefore, a very scrupulous item analysis.

Some generalizations can be made with respect to the dependability and the comparative values of the different methods of item analysis. First, it can be said that indices of difficulty are much more stable than indices of item validity.

From the reports of Gibbons[2] and Carter[3] one may conclude that indices of difficulty are highly consistent from

[1] Gulliksen, Harold, 'Theory of Mental Tests', John Wiley & Sons, Inc., 1950, p. 365.

[2] Gibbons, C.C.,The Prediction value of the most valid items of an examination, Journal of Educational Psychology, 1940, pp. 616-621.

[3] Carter,H.D.,How reliable are the common measures of difficulty and validity of objective test items? Journal of Educational Psychology, 1942, pp. 31-39.

sample to sample even with N as low as 50. Indices of item validity that they reported, however, tend to be much less consistent from sample to sample.

The type of test and of tested population would undoubtedly have bearings on the stability of item indices.

PREPARING ITEM ANALYSIS DATA

After the pilot testing was over, the next step that the investigator followed was that of scoring the tests. He personally examined all the 371 answer-sheets and scored all the items in all the sub-tests. The scoring method has already been described in earlier pages. For the purpose of collecting data for item analysis, a slightly different procedure for scoring items in test II was followed from what has been described at an earlier stage. The items that were scored '5' or '4' were considered as correct ones and those that were scored '1' or '2' were considered as incorrect ones. Thus the 'right-wrong' procedure was followed in scoring these items, the highest obtainable score remaining the same, viz. 35. In scoring items 6, 7, 8, 9 in sub-test V, exactly the same change was made.

The two groups - 'high scoring' and 'low-scoring' were, then, formed as follows:

(1) All the testees were assigned ranks. The testee getting highest score was ranked first while that getting the lowest one was ranked last.

(2) All the answer-sheets were arranged in order of ranks, the first number being at the top and the last one at the bottom.

(3) From the pile of the answer-sheets, he took first 100 answer-sheets. This formed the 'upper group' or "high scoring group". This contained 27 per cent cases of the whole group.

(4) Then he took the bottom 100 answer-sheets. This formed the 'lower group' or the 'low scoring group'. This also contained 27 per cent cases of the whole group.

(5) The middle 46 per cent of the testees were, thus, discarded.

After the formation of the two groups, the number of correct responses to an item in each group was found out. The numbers of correct responses in each group for all the 143 items in the total test were found out and tabulated. These numbers, naturally, showed the percentage of correct responses for each item for both the groups, as each group was formed of 100 testees.

Then the correction for chance success was applied. For this, the following formula for computing per cent of correct responses was used. (We have discussed this formula at length in earlier pages).

$$P_t = 100 \times \frac{R_t - \frac{W_t}{ki - 1}}{N_t - NR_t}$$

These corrected percentages of correct responses in both 'upper' and 'lower' groups are shown in table No. 11

The data for item analysis were thus prepared.

ITEM DIFFICULTY

The difficulty on an item may be determined in several ways: (1) by the judgment of competent people who rank the items in order of difficulty, (2) by how quickly the item can be solved, and (3) by the number of examinees in the group who get the item right.

The first two procedures are usually the first step, especially when the items are for use in special aptitude tests, in performance tests, and in areas where qualitative distinctions and opinions must serve as criteria. But the number right, or the proportion of the group which can solve an item correctly, is the "standard" method for determining difficulty in objective tests.

Guilford[1] suggests, "When the item is scored either 0 or 1, the simplest index of its difficulty is its mean item score p." Davis[2] also says,

> Many ways of expressing the difficulty level of an item have been proposed. The most obvious of these is the per cent of the tryout group that marks it correctly.

[1] Guilford, J.P., Op. Cit., p. 418.

[2] Lindquist, E. F., Op. Cit., p. 267.

The investigator, following these suggestions, calculat-ed the indices of item difficulty for items in this test, in terms of percentage of individuals, in the groups selected, who could answer the items correctly. The smaller the percentage succeeding on the item, the more difficult the item, and vice-versa.

As Coombs[1] has pointed out the difficulty of an item varies for different individuals. We do not have accurate information concerning an item's difficulty for an individual. All we know is that if he passes it, the item is less difficult than his ability to cope with it and if he fails it is more difficult than his ability to cope with it.

FACTORS AFFECTING ITEM DIFFICULTY

The following factors are likely to affect the difficulty level of an item.:

(1) The nature of its (item's) content and the type of behaviour it requires of the examinee.

(2) Unusual vocabulary may markedly, albeit unintentionally, influence responses to an item.

(3) Awkward sentence structure and undue formality in the style of the language used, often have unpredicted effects.

[1] Coombs, C. H., The Concepts of reliability and homogeneity, Educ. Psychol., Meassut, 1950, pp.43-56.

(4) A shift from use of third person to first or second may make an item significantly easier.

(5) Even such apparently extraneous factors as the form of the item and the directions to the examinees may affect item difficulty.

It is much more likely that the presence of any of the above factors may raise or lower the difficulty level of items and a false picture of the difficulty level may be obtained. To guard against this situation, the author of this test took utmost care while constructing the test, to see that the effect of these factors was reduced to the minimum.

CALCULATING THE DIFFICULTY LEVELS OF THE TEST ITEMS

The following formula was used to calculate the difficulty value, 'D', of each item.:

$$D = \frac{U + L}{2}$$

Where D = difficulty value of the item.

U = percentage of testees scoring the item correctly in the upper 27% after being corrected for guess work.

L = percentage of testees scoring the item correctly in the lower 27 per cent after being corrected for guess work.

This method of calculating difficulty value of an item involves the elimination of the middle 46 per cent testees. Some doubts may, therefore, be raised, as to the reliability of the difficulty values of items computed by this method.

We would like to quote Davis[1] to answer such doubts. From the results of his investigations, he concludes,

> The writer has computed the reliability coefficient of a group of typical item difficulty indices estimated in this way and has found it to be 0.98 when the sample included 100 examinees in the highest 27 per cent and 100 examinees in the lowest 27 per cent. These data suggested that the loss of reliability incurred by estimating indices from only 54 per cent of the sample tested is not sufficient to be of practical consequence when the two criterion groups employed include at least 100 examinees apiece.

He further adds,[2]

> Experimental evidence has shown that difficulty indices of the sort described are extremely reliable when they are based on samples as large as 400.

The sample for the present item-analysis consistéd of 371 testees. The reliability of the difficulty levels of items computed by this method can, therefore, be granted.

The difficulty values, D, of all the items in the test are shown in the table No. 11 . The lower the value of 'D', the higher the difficulty level.

[1] Lindquist, E. F., Op.Cit., p. 282.

[2] Ibid., p. 282.

In item selection, not only must individual item difficulty be considered, but the item discrimination indices as well. We shall, now, proceed to this phase of item analysis.

ITEM DISCRIMINATION INDICES

The term "item discrimination indices" includes both 'internal consistency item discrimination' and 'item validity' indices.

This discrimination may be in terms of total score on the test, or it may be in terms of some external criterion score of job performance.

The relationships between the total scores derived from a test and item scores are referred to as internal-consistency item discrimination indices. Internal consistency is also known as 'item consistency'.

The relationships between item scores and scores in a criterion variable other than the total score on the test are referred to as "item validity indices".

In some cases one type of analysis will be appropriate, in some cases the other. There may be some cases in which both seem reasonable.

The question arises: which are the situations that call for analysis of the relationship of item to total test score (internal consistency analysis) and the situations that call for analysis of the relationship of item to an external criterion (item validation)?

We shall quote Thorndike[1] to answer this question. He says,

> Where the separate test items represent such a heterogeneous assortment of materials,..........................
> ......there is no point in carrying out an internal consistency item-analysis.

The same author[2] again concludes,

> When the items in a test are completely homogeneous in that each item measures exact -ly the same factors or aspects of individual ability in the same combination...............
> any further analysis of individual items against an external criterion is futile.

In actual practice the situation may not be so clear-cut. Many tests may have a degree of both homogeneity and heterogeneity. So it is safe and advisable to resort to both the types of analysis.

The present aptitude test includes five different sub-tests designed to measure five different factors - hypothetical - of course. So the test as a whole can be said to be heterogeneous, while the sub-tests are homogeneous to a great extent as the items in each sub-test are constructed to measure the same factor. The author of the present test is, therefore,

[1] Thorndike, R. L., 'Personnel Selection' Test and Measurement Techniques, John Wiley & Sons, Inc., New York, 1949, p. 231.

[2] Thorndike, R. L., Ibid., p. 232.

justified in applying both the techniques of item-analysis here.

Guilford[1] also advocates the use of both the techniques. He says,

> When a test is to have maximum validity, each item must correlate as high as possible with the external criterion and as low as possible with other items in the test. This goal calls for item analyses against both the total score criterion and the outside criterion.

We followed a little novel way of item-analysis. We applied both the techniques at two different stages. 'Item validity' technique was applied on the data obtained from tryout testing. The two criterion groups were formed and each item in the test was validated against these two groups. Only the valid items were included in the pilot test form. This has been discussed at length in chapter, VI. Item analysis for internal consistency was done at this stage.

CORRELATION INDICES OF ITEM CONSISTENCY

A number of methods have been devised for use in determining the discriminative power of an item. Out of these, four coefficients of correlation are commonly used to indicate the correlation of an item with the total test score. These are the biserial 'r', point-biserial 'r', tetrachoric 'r', and the phi coefficient.

[1] Guilford, J. P., Op.Cit., p.442.

F. Davis[1] suggests,

> To provide an index of discriminating ability that is essentially unaffected by differences in the percent of testees answering correctly items scored "right" or "wrong", the biserial 'r' may be employed when the criterion variable is continuous.

Similarly, Garrett[2] also says, "....but biserial correlation is usually regarded as the standard procedure in item analysis." Biserial 'r' gives the correlation of an item with total score on the test. There is no one type of item analysis data that is best under all circumstances. Never -theless, circumstances that favour the use of biserial 'r' as an index of discriminating power appear to be most numerous. Considering all these points, the investigator used the biserial 'r' to determine the discriminative power of an item here.

The best formula to use for the biserial 'r' in the item-analysis application is:

$$r_b = \frac{M_p - M_t}{\sigma_t} \times \frac{p}{y}$$

Where M_p = mean criterion score of those passing item,

M_t = mean criterion score of all examinees,

σ_t = Standard deviation of all total scores,

[1] Lindquist, E. F., Op.Cit., p. 292.

[2] Garrett, H.E., Statistics in Psychology and Education, Longmans, Green and Co., New York, 1958, p. 365.

p = proportion passing item,

y = ordinate in unit normal distribution corresponding to p.

Computing "r"'s by this formula for all 143 items, would be much laborious and much time consuming. Some easy method was needed. This was that devised by Flanagan. He, working with Kelley, devised an ingenious procedure based on the fact that, since the magnitude of a correlation coefficient is determined by extreme cases to a much greater extent than by cases near the middle of the bivariate surface, an estimate of the coefficient may be obtained with a much greater decrease in labour than in accuracy by utilizing only the data in the tails of the two distributions.

The most satisfactory item validity index based on the upper and lower 27 per cent is the estimate of the coefficient of correlation between item and test obtainable from tables prepared by Flanagan.[1] Flanagan's table makes it extremely simple to compute item validity coefficients from the percentages of success in the upper and lower 27 per cent. By entering the table in the appropriate row and column, the correlations are read directly. We prepared our item-analysis data with a view to use this table for obtaining biserial "r"'s between our test items and test score.

[1] Thorndike, R. L., Op. Cit., p. 348.

We thus found out indices of item discrimination using the table prepared by Flanagan.

It should be noted once again that the 'r'$_{bis}$ read from the table is less accurate than is the usual 'r'$_{bis}$ as it utilises only about 1/2 of the test data - the middle 46 per cent being not used. The loss of accuracy in these validity indices is of little consequence when they are used comparatively and the ease of computation is a practical advantage.

In the table No. 11 , are given the indices of internal consistency for each item alongwith its difficulty value.

TABLE NO. 11

SHOWING INTERNAL CONSISTENCY DATA, 'INTERNAL CONSISTENCY INDICES' AND 'DIFFICULTY VALUES' OF THE ITEMS

Sub-Test No.	Item No.	'U' %	'L' %	'D'	'r'	Remarks
1	2	3	4	5	6	7
IA	1	87	57	72.0	0.368	-
	2	85	57	71.0	0.335	-
	3	70	50	60.0	0.210	-
	4	97	81	89.0	0.395	-
	5	97	87	92.0	0.300	-
	6	59	39	49.0	0.203	-
	7	88	49	68.5	0.455	-
	8	91	72	81.5	0.300	-

Sub-Test No.	Item No.	'U' %	'L' %	'D'	'r'	Remarks
1	2	3	4	5	6	7
IA	9	87	57	72.0	0.368	-
	10	73	63	68.0	0.115	Rejected
IB	11/1	81	46	63.5	0.380	-
	12/2	83	34	58.5	0.505	-
	13/3	97	71	84.0	0.495	-
	14/4	82	90	86.0	-0.150	Rejected
	15/5	96	82	89.0	0.330	-
	16/6	97	64	80.5	0.550	-
	17/7	97	75	86.0	0.458	-
	18/8	71	35	53.0	0.370	-
	19/9	77	31	54.0	0.468	-
	20/10	95	63	79.0	0.490	-
	21/11	76	39	57.5	0.380	-
IC	22/1	67	43	55.0	0.250	-
	23/2	42	30	36.0	0.130	Rejected
	24/3	83	63	73.0	0.255	-
	25/4	43	25	34.0	0.200	-
	26/5	95	63	79.0	0.490	-
	27/6	76	47	61.5	0.247	-
	28/7(i)	97	81	89.0	0.395	-
	29/7(ii)	79	58	68.5	0.245	-
II	1	79	38	58.5	0.430	-
	2	99	79	89.0	0.540	-

Sub-Test No.	Item No.	'U' %	'L' %	'D'	'r'	Remarks
1	2	3	4	5	6	7
II	3	93	89	91.0	0.100	Rejected
	4	71	65	68.0	0.070	Rejected
	5	86	70	78.0	0.220	-
	6	79	71	75.0	0.105	Rejected
	7	79	72	75.5	0.100	Rejected
	8	76	63	69.5	0.150	* -
	9	57	36	46.5	0.215	-
	10	78	43	60.5	0.370	-
	11	49	19	34.0	0.335	-
	12	41	33	37.0	0.090	Rejected
	13	81	43	62.0	0.408	-
	14	48	34	41.0	0.150	* -
	15	79	62	70.5	0.205	-
	16	45	24	34.5	0.230	-
	17	75	49	62.0	0.280	-
	18	66	26	46.0	0.410	-
	19	59	32	45.5	0.280	-
	20	65	60	62.5	0.050	Rejected
	21	35	15	25.0	0.265	-
	22	72	51	61.5	0.220	-
	23	62	41	51.5	0.202	-
	24	68	41	54.5	0.280	-
	25	35	18	26.5	0.215	-

Sub-Test No.	Item No.	'U' %	'L' %	'D'	'r'	Remarks
1	2	3	4	5	6	7
II	26	70	50	60.0	0.210	-
	27	65	40	52.5	0.260	-
	28	53	37	45.0	0.170	- *
	29	41	22	31.5	0.220	-
	30	42	28	35.5	0.150	- *
	31	37	29	33.0	0.090	Rejected
	32	36	21	28.5	0.180	- *
	33	60	40	50.0	0.210	-
	34	80	40	60.0	0.420	-
	35	64	36	50.0	0.290	- -
	36	77	59	63.0	0.205	-
	37	42	32	37.0	0.110	Rejected
	38	68	39	53.5	0.300	-
	39	61	35	48.0	0.270	-
	40	80	56	68.0	0.270	-
	41	17	7	12.0	0.215	-
	42	84	55	69.5	0.340	-
	43	20	41	30.5	-0.250	Rejected
	44	81	68	74.5	0.165	- *
III	1	73	63	68.0	0.115	Rejected
	2	74	43	58.5	0.325	-
	3	70	51	60.5	0.200	-
	4	55	29	42.0	0.270	-

Sub-Test No.	Item No.	'U' %	'L' %	'D'	'r'	Remarks
1	2	3	4	5	6	7
III	5	86	73	79.5	0.190	- *
	6	16	7	11.5	0.200	-
	7	33	21	27.0	0.148	- *
	8	12	4	8.0	0.230	-
	9	37	16	26.5	0.270	-
	10	71	42	56.5	0.290	-
	11	24	10	17.0	0.230	-
	12	87	65	76.0	0.295	-
	13	73	43	58.0	0.312	-
	14	56	39	47.5	0.170	- *
	15	73	53	63.0	0.215	-
	16	70	60	65.0	0.110	Rejected
	17	86	63	74.5	0.300	-
	18	48	59	53.5	-0.110	Rejected
	19	91	59	75.0	0.425	-
	20	35	27	31.0	0.075	Rejected
	21	44	54	49.0	-0.05	Rejected
	22	97	81	89.0	0.395	-
	23	78	55	66.5	0.260	-
	24	57	33	45.0	0.250	-
	25	53	37	45.0	0.168	- *
IV	1	55	35	45.0	0.210	-
	2	48	26	37.0	0.240	-

Sub-Test No.	Item No.	'U' %	'L' %	'D'	'r'	Remarks
1	2	3	4	5	6	7
IV	3	42	22	32.0	0.230	-
	4	59	41	50.0	0.182	- *
	5	48	27	37.5	0.225	-
	6	25	18	21.5	0.095	Rejected
	7	28	14	21.0	0.200	-
	8	41	23	32.0	0.205	-
	9	61	22	41.5	0.410	-
	10	68	44	56.0	0.250	-
	11	59	37	48.0	0.223	-
	12	68	64	66.0	0.040	Rejected
	13	45	26	35.5	0.210	-
	14	54	21	37.5	0.305	-
	15	39	15	27.0	0.305	-
	16	18	7	12.5	0.230	-
	17	33	17	25.0	0.210	-
	18	51	30	40.5	0.220	-
	19	39	25	32.0	0.160	- *
	20	46	32	39.0	0.150	- *
	21	67	44	55.5	0.240	-
	22	42	22	32.0	0.230	-
	23	36	26	31.0	0.120	Rejected
	24	69	43	56.0	0.270	-
	25	53	23	38.0	0.322	-

Sub- Test No.	Item No.	'D' %	'L' %	'D'	'r'	Remarks
1	2	3	4	5	6	7
IV	26	54	36	45.0	0.190	- *
	27	50	39	44.5	0.115	Rejected
	28	28	14	21.0	0.200	-
	29	63	43	53.0	0.202	-
	30	79	43	61.0	0.382	-
	31	63	47	55.0	0.168	- *
	32	67	50	58.5	0.180	- *
	33	79	79	79.0	0.000	Rejected
	34	38	18	28.0	0.250	-
	35	22	15	18.5	0.105	Rejected
	36	34	15	24.5	0.255	-
V	1	72	39	55.5	0.340	-
	2	95	79	87.0	0.335	-
	3	73	46	59.5	0.310	-
	4	63	36	49.5	0.280	-
	5	64	40	52.0	0.250	-
	6	67	35	51.0	0.330	-
	7	95	70	82.5	0.430	-
	8	90	71	80.5	0.290	-
	9	59	41	50.0	0.225	-

ANALYSIS OF THE INCORRECT DISTRACTORS

A good test should include only those items which have fairly good discriminating value. Not only an item should have good discriminating value, but the distractors of a multiple -choice type item should also be in a position to discriminate well between individuals. We consider an item good if it is answered correctly by more individuals from the upper group than it is answered correctly by individuals from the lower group and that the difference is statistically significant. Similarly, we say that an incorrect distractor is good, if it is selected by more individuals from the lower group than it is selected by the individuals from the upper group and that the difference is statistically significant.

The analysis of an item will, therefore, remain incomplete if the analysis of the incorrect distractors is not done. To realise the full advantage of item-analysis, the analysis of the responses to different incorrect distractors is, therefore, of much importance.

The data prepared for item analysis were useful here also. The responses given to each item by testees in each of the two groups were studied very minutely and the frequencies of alternatives selected by them for each item were found out and tabulated.

If an incorrect alternative is really discriminating, it is more likely to be selected by the testees in the lower 27 per cent group than by those in the upper one. If it is not discriminating,the result will be just the reverse.

In the present test, 120 items were finally selected from the results of item-analysis. About 68 items in the whole test, are of multiple-choice type. Each item has four distractors. Thus the investigator had to study responses given by 200 testees to these 272 distractors. This analysis, also, was carried out by the investigator. The results of this analysis showed that there were six items which had non-discriminating incorrect distractors.

In the following table, the analysis of the responses to these six items having one or more non-discriminating distractors is given: (Table given on the next page).

In the table No. 12 the numbers underlined show the frequencies of the correct responses whereas those bracketed show the non-discriminating distractors.

Inspection of choice-by-choice data of the kind illustrated above revealed a few incorrect choices that were discriminating in the wrong direction and a few that attracted virtually no testees. The former operate to destroy an item's discriminating power while the latter are non-functioning and may waste space and reading time.

As shown in the table No. 12 only six items out of 120 items finally selected for the final test, possess some non-discriminating distractors. Even these items have only one

TABLE NO. 12

ANALYSIS OF THE NON-DISCRIMINATING DISTRACTORS

Sr. No.	Item No.	Group	Percentage of different distractors selected			
			Distractors			
			(a)	(b)	(c)	(d)
1	Sub-Test I Item No. 6	Upper group Lower group	2 21	59 39	24 28	(15) (12)
2	Sub-Test III Item No. 5	Upper group Lower group	5 11	(8) (7)	86 73	1 9
3	Sub-Test IV Item No.16	Upper group Lower group	35 50	10 13	18 7	(37) (30)
4	Sub-Test IV Item No. 24	Upper group Lower group	(19) (15)	0 20	69 43	12 22
5	Sub-Test V Item No. 3	Upper group Lower group	73 46	5 21	(15) (12)	7 21
6	Sub-Test V Item No. 8	Upper group Lower group	90 71	(5) (4)	0 16	5 9

non-discriminating distractor each. Ordinarily such items as have grossly invalid and non-functioning choices should be deleted or the distractors be changed or a new order be given to the distractor.

The author of the present test has not deleted these items for the following reasons.:

(1) The differences between frequencies in case of all but one non-disćriminating distractors are not significant. In item No. 16 - Sub-test IV, the difference is 7, but this is also not so significant.

(2) All these items have acceptable difficulty values, have high values of internal consistency and are proved valid against the two criterion groups (see table No. 11 on page 192).

(3) Occasionally, an incorrect c̈hoice that is markedly discriminating in the wrong direction represents a concept that cannot be removed without destroying the whole point of the item. Davis[1] writes,

> If revision of an item involves destroying its point, the item either should be discarded or should be used without revision inspite of its probable lack of efficiency.

Some test-constructors advocate the use of multiple-

[1] Lindquist, E. F., Op.Cit., p. 306.

regression techniques in the final selection of the items of a predictor test. But Thorndike[1] says, "Analytical treatment of item validities by the procedures of multiple regression must be ruled out." For this conclusion, he gives much plausible reasons. The present test-constructor,following Thorndike's conclusion did not use multiple regression techniques in finally selecting the items of this predictor test.

Ultimately, it should be remembered that item-analysis techniques cannot alone be relied upon to detect errors and ambiguities; expert criticism and editing are indispensable in test construction. The criticisms of the item by experts were obtained during the test construction process and due changes in the body of the itemswere made. Thus the test author tried to realise the full value of item-analysis techniques by applying all possible known procedures.

Even then, if the investigator gets an opportunity in future to revise the test, he may certainly change these distractors.

New order to some of the non-discriminating distractors was given.

ITEM SELECTION

Having item information, we use it most commonly to guide us in composing the final form of a test. Besides the

[1] Thorndike, R. L., Op.Cit., p. 244.

obvious goals of maximising reliability and validity, there are a number of secondary goals. Among them are a good total-score distribution and a rank ordering of items as to difficulty.

The final selection of the items became very smooth and simple as each item had to pass through many ordeals beforehand. They were:

(1) Experts' criticism.

(2) Item validity.

(3) Difficulty value.

(4) Internal consistency.

1. Experts' criticism: As discussed in the previous chapter, the test items, after they were framed, were circulated among the experts for their criticism as to the content validity of each item. Many items were omitted and many were revised after studying these criticisms. Due care was taken to see that each item was precise and as concise as possible.

2. Item validity: This was the main criterion used in eliminating weak and defective items. As discussed in the previous chapter, each item was validated against the two criterion groups and all the invalid items were totally rejected. As many as 40 items were eliminated as a result of this process. This procedure was adopted at a tryout stage. For 'internal consistency' and 'difficulty value' analysis altogether new data were collected. This process helped the investigator in avoiding any confusion regarding item validity and internal

consistency.

3. <u>Difficulty value:</u> As said above, only the valid items were included in the pilot test form.

Before the items are screened out on the basis of difficulty value, it is necessary to decide as to what range of difficulty values of items should be maintained in a good predictor test. Many authors have given their views on this theme. We shall examine two of them.

Garrett[1] says,

> The larger the variance of the item, the greater the number of separations among individuals the test item is able to make. Other things being equal,items of moderate difficulty - 40 - 50 - 60 per cent passing - are to be preferred to those which are much easier or much harder.

The same author again suggests,

> The normal curve can be taken as a guide in the selection of difficulty indices. Thus, 50 per cent of the items might have difficulty indices between 0.25 and 0.75, 25 per cent indices larger than 0.75, and 25 per cent smaller than 0.25.

While, W. Summer[2] has suggested that the items of different difficulty levels should be selected for inclusion in the following proportion.:

[1] Garrett, H.E., Op.Cit., pp. 364-365.

[2] Summer, W., " Statistics in Education", London, Basil Blackwell & Co., p. 180.

Items of difficulty range from 0-40	20 per cent
Items of difficulty range from 41-60	60 per cent
Items of difficulty range from 61-100	20 per cent

An item passed by 0 per cent or 100 per cent has no differentiating value, and so such an item should be discarded at the first stretch. Such items were, therefore, discarded after the tryout testing and were not included in the pilot test form.

In the following table the frequencies of items of different difficulty values, expected number of items in various difficulty zones as suggested by Thorndike and Summer and the obtained number of items in different zones are shown. The numbers in the bracket indicate the expected frequencies in each zone while numbers underlined indicate the frequencies obtained in the zone.

TABLE NO. 13

DISTRIBUTION OF ITEMS ACCORDING TO 'D' VALUES

Difficulty indices	f	Garrett's Distribution 1. below 25 'D' - 25% 2. between 26 D & 75 D - 50% 3. above 75 D - 25%	W. Summer's Distribution 1. 0- 40 - 20% 2. 41-60 - 60% 3. 61-100 - 20%
1 - 5	0		
6 - 10	1		
11 - 15	3	(30)	(24)
16 - 20	1	10	
21 - 25	5		
26 - 30	6		31
31 - 35	9		
36 - 40	6		
41 - 45	10		(72)
46 - 50	10	(60)	
51 - 55	12	92	47
56 - 60	15		
61 - 65	10		
66-- 70	6		
71 - 75	8		
76 - 80	5		(24)
81 - 85	5	(30)	
86 - 90	7		42
91 - 95	1	18	
96 -100	0		

N = 120

The distribution of the items in the present test, thus, does not agree so closely either with Garrett's distribution or with that of W. Summer's. But a glance at the table will show that the divergence is not so great that if the proportion is not revised the adverse effect will be made on the efficacy of the test. Moreover, these distributions suggested are arbitrary and not final. About 102 items fall between 21 and 80 'D' range. This much range is sufficiently good and acceptable for any good predictor test. Under these conditions not a single item has been dropped on account of its being too easy or too difficult.

4. Internal consistency: Since the total test score is used as the criterion, biserial 'r' (Flanagan 'r') is computed.

It can be seen from table No. 11 , that all but only 4 items show a positive value of 'r'. This is quite natural as the validity of all the items had been tested at the tryout stage and only the valid items were included in the pilot test form.

The question arises as to what should be the minimum internal-consistency discrimination index of an item for its being selected in the final form of the test. According to Thorndike[1], "an item with a validity coefficient (internal - consistency discrimination index) as high as 0.25 usually

[1] Thorndike, R. L., Op. Cit., p. 245.

represents an outstandingly 'valid' item." And according to Garrett[1],

> the size of an acceptable validity index will depend upon the length of the test, the range of the difficulty indices, and the purposes for which the test is designed,.........as a general rule,items with validity indices of 0.20 or more are regarded as satisfactory.

While according to Davis[2] (in the case of a predictor test),

> if item analysis data are available with the total score on all the items, some preference should be given to items that have the lowest internal-consistency discrimination indices.

Keeping all these suggestions in view and that the present test is a predictor one, the investigator decided to retain the items having internal consistency indices 0.15 or more. It will be noted that out of 120 items finally selected, only 16 items have internal consistency indices between 0.15 and 0.19. These items are marked with asterisks. The rest have internal consistency indices 0.2 or more. This is fairly in agreement with Garrett's suggestion. This also shows that the items in each sub-test are to a more or less extent homogeneous.

The items,with internal consistency indices less than 0.15, were rejected altogether. Thus, out of 143 items in the

[1] Garrett, H. E., Op.Cit., pp. 367-368.

[2] Lindquist, E. F., Op.Cit., p. 316.

pilot test, 23 items were rejected on the basis of low internal consistency indices. 120 items were retained. This 120 items formed the final test.

We shall discuss the final form of the test in the next chapter.