AN ITEM ANALYSIS OF AN OBJECTIVE TEST IN BIOLOGY

by

DELBERT ALLEN NEWBERRY

B. S., Fort Hays Kansas State College, 1939

A THESIS

submitted in partial fulfillment of the

requirements for the degree of

MASTER OF SCIENCE

Department of Education and Psychology

KANSAS STATE COLLEGE
OF AGRICULTURE AND APPLIED SCIENCE

1947

Docu-
ment
LD
2668
.T4
1947
N4
c 2

TABLE OF CONTENTS

INTRODUCTION

Testing at the beginning of a course is important in helping the teacher to adjust his teaching to the needs and educational level of his students. Achievement may be measured either with reference to an arbitrary standard of what a student should be like at the end of a course or with reference to what he was like at the beginning of the course and his progress since that time. The use of a standardized test at the end of a course is helpful in estimating the extent to which the objectives of the course have been achieved. In cases where a student requests advanced credit in a course, a standardized test is the best instrument for comparing his achievement with that of students who have completed the course.

The USAFI Tests of General Educational Development are an outstanding example of standardized tests for use at both the high school and college levels. Their major purposes (26) are to provide a basis for vocational and educational guidance for veterans, to assist schools in the placement of returning veterans and to help schools determine the amount of academic credit to be granted for educational experiences in military service. Tyler (26) lists three types of opportunities for educational experiences in military service. They are military training, the off-duty educational program and informal experiences. After World War I, many educational institutions granted blanket credit for military service with unsatisfactory results in many cases.

To avoid similar results after World War II, a special committee called together by the American Council on Education decided that a uniform system of testing to demonstrate competence should be developed to aid the schools in handling requests for advanced credit. Bradley (3) found a correlation of .66 between grade-point averages in social studies and scores on the GED test in Interpretation of Reading Materials in Social Studies, but found no significant relationship between test scores and number of hours of college credit completed in the field. Because they are tests of general educational development rather than achievement tests for specific courses they do not fill the need for standardised tests adapted for use with particular college courses.

The development of college level achievement tests has been encouraged by the Botanical Society of America through the work of its committee on the teaching of botany in American colleges and universities (5, 14). They emphasized the point that a valid achievement test should measure more than the students' memory for facts. The ability to apply the facts learned is also essential. The objectives of the course should be clarified and the test constructed to measure the extent to which the student responds in the desired way in view of all the objectives of the course.

At Kansas State College, Dr. M. J. Harbaugh, Professor of Zoology, constructed a 100 item objective test in biology which he desires to standardize. This test was given to students enrolled in the course, Biology in Relation to Man, in September, 1946. At the conclusion of the two semester course, the same

test was administered again to the same students.

Before the validation and standardization of the test could be completed, an item analysis was necessary to determine wherein it could be revised and improved. Such an item analysis is the problem of this thesis.

In the development of this problem the following aspects were studied:

1. An item analysis was made to determine the validity and difficulty of each item. The relationship between item difficulty and item validity was investigated.

2. The validity of the test was determined by correlation of the total test scores with grades for the two semesters in the course, Biology in Relation to Man.

3. The reliability of the test was determined by the Kuder-Richardson formula.

4. The 60 items with validity coefficients of .20 or higher were selected and reliability and validity coefficients were computed for this 60 item test.

5. An analysis was made of the choices of answers to all questions as a basis for the revision of the items to secure greater validity.

6. Recommendations were made for the revision of the test.

REVIEW OF LITERATURE

Item analysis involves the two general problems of item validity and item difficulty.

A study of item validity deals with the diagnostic value of each item for predicting a criterion. Guilford (10) has expressed the purpose thus "to be diagnostic of any trait, an item must enable us to distinguish between individuals who have more or less of that trait." If the criterion scores of individuals who pass an item are not significantly different from the criterion scores of those who fail to pass that item, the item does not contribute to the measurement of the trait of which the criterion is the standard.

In determining item validity, just as in determining the validity of a total test, the choice of a criterion is of primary importance. The two types of criteria with which test items may be correlated are an independent criterion such as is used in validating a total test and the criterion of internal consistency. Swineford (23) has shown that statistical methods of item-criterion correlation are equally applicable to independent and internal criteria so the criterion chosen should be the one which most nearly represents the trait to be measured.

Internal consistency, or the correlation of the item with the total test score, is the most widely used criterion for the validation of test items. Owens (20) has criticized the method on the basis that it may result in the narrowing of the test so that the validity of the total test may be decreased. Swineford

(23) pointed out that in some cases the total test score is the best measure available of the trait to be tested and therefore an internal criterion may be superior to an independent criterion. She also stated that in any case where a test was known to be valid, item validity could be satisfactorily determined by the correlation of the item with the total test score. Guilford (10) recognized both methods as acceptable although he cautioned against too great narrowing of the test by the method of internal consistency and warned that an independent criterion must be chosen with care because of the difficulty of finding an adequate criterion.

Internal consistency was chosen as the criterion for use in the item analysis of the biology test.

There are numerous statistical methods for determining the correlation of an item with a criterion. Certain standard technics such as the biserial r, the tetrachoric r and the phi coefficient are recognized and described in the statistical literature. Other methods, described in the professional journals, have been developed as practical short cuts to obtain approximately the same result with a simplification of the method of computation.

The biserial coefficient of correlation is generally recognized as one of the most accurate methods of determining item validity because each criterion score is given due weight without change of value by grouping into categories. The formula for the biserial r as given by Guilford (10) is:

$$r_{bi} = \frac{M_p - M_q}{\sigma_t} \times \frac{pq}{y}$$

where M_p = the mean criterion score of the group passing the item

M_q = the mean criterion score of the group failing the item

p = the proportion of cases in the higher group

q = the proportion of cases in the lower group

y = the ordinate of the normal distribution curve with surface equal to 1.00, at the point of division between the segments containing p and q proportions of cases

σ_t = the standard deviation of the total sample in the continuously measured variable (criterion).

This formula is based on the principle that if there is no difference between the mean criterion score of the group giving the correct response to an item and the mean criterion score of the group giving a wrong response to the item, there is no correlation between the item and the criterion. The larger the difference between the means the higher is the correlation. The principal objection to this method of item analysis is the time consumed in sorting and computing mean criterion scores separately for those passing and for those failing each item since the number and particular individuals who pass differ from item to item.

The tetrachoric correlation coefficient is frequently used in item analysis. It assumes that both variables are continuous and normally distributed but are reduced artificially to two categories each. Guilford (10) listed as its disadvantages the fact that it is extremely difficult to compute by formula and that it is less reliable than the Pearson r because of the coarse

grouping into only two categories. For this reason it is useful only with large samples. Computing diagrams may be used to effect considerable saving in time when a large number of tetrachoric r's are to be computed. Guilford (10) recommended the Thurstone computing diagrams by Chesire, Saffir and Thurstone (6). Other computing diagrams have been published more recently by Hayes (12).

The phi coefficient is another statistical technic which is used in item analysis. When one or both of the traits are really dichotomous it is the most suitable method of correlation according to Guilford (10). He found the chief objection to it based on the fact that it is not always equivalent to the Pearson r and can not be interpreted in the same way. When both variables are continuous but one is dichotomously scored, the phi coefficient is smaller than the Pearson r. The phi coefficient also varies in size according to the percentage of the cases included in the upper and lower criterion groups. These disadvantages are unimportant if only the relative validities of the items of a test are needed to evaluate the items.

Jurgensen (15) has developed tables for determining phi coefficients accurate to three places and identical to those obtained by formula if sub-groups are equal in number. He pointed out that by the use of these tables item validities could be determined more accurately and quickly than by many methods designed to reduce computation time which also sacrificed some efficiency and accuracy.

Turnbull (25) presented a normalized graphic method of item analysis which included not only the correlation of the item with the criterion but also an analysis of choice of responses. By this method the students were divided into sixths according to the criterion ratings and the percentage of students in each sixth choosing each response was plotted on the graph. The line for the correct response was expected to show a considerable upward slope from the lowest sixth to the highest sixth as increasing percentages of the better students selected it while the lines for the incorrect responses were expected to slope downward from the poorest to the best students. An item was considered in need of revision if the lines for the responses did not slope as expected. He also stated that any response which did not attract a considerable percentage of the students should be made more plausible.

By using arbitrary distributions of the criterion scores Adkins and Toops (1) derived several modifications of the Pearson correlation coefficient of a dichotomous variable with a multiple-categoried variable. Their formulas simplified computation and effected a considerable saving in time without sacrificing accuracy or requiring correction for coarse grouping.

They presented several formulas for different numbers of categories and distributions of the criterion but recommended two as the most convenient to use. One was an approximately normal distribution with the total number of cases a multiple of 16, divided into five categories with relative frequencies of 1, 4,

6, 4 and 1. The other was a rectangular distribution with the total number of cases a multiple of five, divided into five equal categories. In either distribution the criterion scores were coded symetrically about zero as -2, -1, 0, 1, and 2, with the coded scores assigned to the five categories in order of excellence.

The formula for use with a rectangular distribution of five equal categories assigned coded scores as above is:

$$r_{xy} = \frac{2a + b - d - 2e}{\sqrt{2\ R\ W}}$$

where r_{xy} = the coefficient of correlation between the item and the criterion

a = the number of right answers in the highest fifth

b = the number of right answers in the second fifth

c = the number of right answers in the third fifth

d = the number of right answers in the fourth fifth

e = the number of right answers in the lowest fifth

R = the total number of right answers

W = the total number of wrong answers

This method was selected for use in the item analysis of the test in biology on the basis of its freedom from certain objectionable features of the other methods.

Adkins and Toops (1) also derived a similar formula for evaluating item alternatives by determining the correlation of each wrong response with the criterion. They showed that all

wrong responses should have low negative validity coefficients and that if any wrong response showed an appreciable positive validity coefficient it should be revised or eliminated. Any response which is closely enough related to the right answer to attract a significantly higher proportion of the better students than of the poorer students materially reduces the diagnostic value of the item and the validity of the correct response.

Determining the difficulty of test items as a part of an item analysis is necessary for the best arrangement of items within the test and as a basis for adjusting the difficulty level to the group to be tested.

Arrangement of test items in order of difficulty, easiest items first, is accepted as a standard procedure in test construction. This allows the students to start with a feeling of confidence and provides that if any questions are omitted for lack of time they will be the most difficult ones which the students would have been least likely to answer correctly. Henry (13) suggested that although a few very easy items at the beginning of a test had little validity value they had another value in encouraging the students. Guilford (10) also approved the use of one or two very easy "shock absorbers" at the beginning of a test even though they contributed nothing to measurement because all the students could pass them.

The difficulty of test items may be graduated more or less steeply from beginning to end or they may be made up mainly of items of average difficulty. Guilford (10) stated that the

maximum discrimination among testees is obtained by items that about half the individuals can pass. Lentz, Hirshtein and Finch (17) agreed that this rule applied to many types of tests and was a good preliminary method of evaluating test items but stated that the rule did not hold for many tests of skill and knowledge. Symonds (24) quoted Thorndyke's statement that "items at any level of difficulty are a valid measure of an individual's ability at all levels in tasks homogeneous in construction and type of material." After further study of the problem Symonds (24) concluded that the best test for measuring a homogeneous group is one in which 50 per cent of all items can be passed by individuals with median scores but that such a test would not measure adequately the upper and lower extremes of a heterogeneous group. He stated that the best test for a heterogeneous group is one constructed with items ranging evenly in difficulty from those passed by 50 per cent of the students in the lowest section of the group to those passed by 50 per cent of the individuals in the highest section of the group.

Henry (13) divided a test into easy, medium and hard items according to the number passing each item and compared the item validities for the three groups of questions. He found no reliable relationship between the difficulty of a test item and its validity.

In studying the relative merits of different methods of evaluating test items Swineford (23) found a correlation of -.0055 between the balance of right and wrong answers and item

validities determined by the biserial r when applied to the same data.

Brogden (4) warned against the danger of decreasing the validity of a test by too great narrowing of the range of difficulty even though this narrowing of range of difficulty might increase the reliability of the test.

The different statements regarding level of difficulty may be summarized in the statement that items which are so easy that they are failed only by chance and items which are so difficult that they are passed only by chance contribute nothing to measurement; items of medium difficulty are preferred by most writers, but there must be sufficient range of difficulty to test both extremes of the group.

DESCRIPTION OF PROCEDURE

Material

The 100 item multiple-choice test in biology was given to Kansas State College students who enrolled in the course, Biology in Relation to Man, in the fall of 1946 and was given again in the spring of 1947 at the conclusion of the two semester course.

The test was dichotomously scored, in that there was one response recognized as right by authoritative judgment of the faculty of the department. Any other response or omission of the item was scored as "not right". The test was presented in a mimeographed booklet. Separate answer sheets were used to facilitate scoring and to permit the re-use of the test booklets. The number of alternative responses to each item varied: 63 items had five alternative answers; 35 questions had four responses from which to choose; and two items had only three optional responses.

There were 508 students enrolled in the course for at least one semester; 69 were eliminated from the study because they either were not enrolled the second semester or, as seniors, did not take the objective test at the end of the course. The number was further reduced by elimination of 30 students who either were not enrolled the first semester or who missed the preliminary test. Because the Toops-Adkins method required that the number

of students be a multiple of five and it was desirable to avoid further arbitrary reductions of the number, one student who had grades for both semesters but had missed the first test was included in the study. This made the number of students, upon whose records computations were based, exactly 400, except in the correlation of the preliminary test scores with grades for which data regarding only 399 students were available.

Reliability

The reliability of a test is defined by Lindquist (13) as its self-consistency. A perfectly reliable test would be one free from errors of measurement so that successive measurements of the same individuals or phenomena would yield exactly the same values. Although no perfectly reliable test of psychological functions exists, it is essential to know how reliable a test is, because, as Bingham (2) has pointed out, no test can have a validity coefficient greater than its reliability coefficient. He has also shown that no test can have a greater validity coefficient than the reliability coefficient of the criterion with which it is correlated although this is often difficult or virtually impossible to determine.

There are three traditional methods of determining the reliability of any test according to Lindquist (13) and Guilford (10). These are the test-retest method, the alternate-forms method and the split-half method. More recently the Kuder-Richardson

formula has been added as a method of determining test reliability and special studies of the split-half method have been made by Guttman (11), Cronbach (7, 8) and others in an attempt to call attention to its weaknesses and to refine the method.

Guilford (10) states that the principal objection to the test-retest method is found in the individual differences in learning during the time between the two tests and in individual differences in learning from the practice effect of the first test. The use of this method to determine the reliability of the biology test would have been of no value because in the time interval between the test and retest all the students had two semesters of study of biology which resulted in changes in individual scores and relative standings.

The chief weakness of the alternate-forms method according to Guilford (10) is the fact that many tests do not have alternate equivalent forms and if there are alternate forms, individual differences in learning from the practice effect of the first test may create differences not due to errors of measurement. This method could not be used with the biology test because there was no alternate form.

Criticism of the split-half method is based on the fact that there are many different possible splits, no one of which can be said to be the only correct one on which to base an estimate of the reliability of the test. Cronbach (7) reported making thirty random splits and fourteen parallel splits of a 38 item silent reading test without securing any two identical reli-

ability coefficients when carried to three decimal places. His purpose was to study the split-half method rather than to study the particular test used and therefore he attempted to make every conceivable split of the data. He did not find parallel splits superior to random splits in a homogeneous test. He concluded that in determining the reliability coefficient of any test by the split-half method at least two splits should be made and that the means and standard deviations of each half should be reported in addition to the coefficient.

The split-half method assumes that the test is divided into two equivalent halves and determines the correlation between the two halves. Since this method yields the reliability coefficient for a test just half the length of the original test the Spearman-Brown formula is used as a correction to determine the reliability coefficient of the full length test. This formula is described by Guilford (10) who calls attention to the fact that its use requires that the two halves must have equal standard deviations.

Guttman (11) has contributed a formula for the split-half reliability coefficient which is not dependent upon equal standard deviations of the two halves. He also recognized the variability of the reliability coefficient and proposed that reliability be described in terms of "lower and upper bounds" rather than as a precise coefficient.

Cronbach (8) has divided the concept of reliability into four different definitions which he classified as the hypothetical

self-correlation, the coefficient of equivalence, the coefficient of stability and the coefficient of equivalence and stability. He stated that the test-retest method yielded the coefficient of stability; the alternate-forms method produced the coefficient of equivalence and stability; the split-half method and the Kuder-Richardson formula yielded coefficients of equivalence and hypothetical self-correlation was obtained by the Guttman formula. He stated that there was no single best estimate of the reliability of a test; that all four were valuable in studying a test but that they were not interchangeable and knowledge of the method used was essential to the interpretation of the reliability coefficient.

Richardson and Kuder (21) derived new methods of estimating test reliability coefficients based upon rational equivalence to eliminate the problems of obtaining comparable halves and of determining which of several equally acceptable methods of dividing the test into halves should be used. Cronbach (7, 8) and Guttman (11) agree that the Kuder-Richardson formula is a conservative estimate of reliability and while it may underestimate the reliability of a test it will not overestimate it. Several variations of the Kuder-Richardson formula have been devised to give a shorter approximation of the reliability coefficient. The basic Kuder-Richardson formula as described by Guilford (10) was chosen as the method for estimating the reliability of the biology test. This formula is:

$$r = \frac{n}{n-1} \times \frac{\sigma_t^2 - \Sigma pq}{\sigma_t^2}$$

where r = reliability coefficient for the whole test

n = number of items in the test

σ_t = standard deviation of the total test scores

p = proportion of the group passing an item

q = proportion failing to pass the item

The reliability coefficient of the 100 question biology test as determined by the Kuder-Richardson formula, described above, was .832.

Validity

Validity has been defined by Lindquist (18) as the accuracy with which a test measures that which it is intended to measure. It is expressed as the coefficient of correlation between the total test score and a criterion and may be used to predict criterion scores for other persons from the same population whose test scores are known but whose criterion ratings are not known. Guilford (10) has pointed out that regardless of what a test is intended to measure, it is a valid test for any sphere of behavior in which it makes prediction of behavior possible. Therefore, no test may be said to have a single validity coefficient as any statement of validity depends upon the criterion used to determine the predictive value of the test.

Grades in the course, Biology in Relation to Man, were used as the criterion for determining the validity of the biology

test. Values of 5, 4, 3, 2 and 1 were assigned to the letter grades of A, B, C, D and F respectively for each semester. Grades for each student for the two semesters were combined, giving a range in composite grades from 10 for the students with two A's to 3 for the students with one D and one F.

The Pearson product-moment formula was applied to determine the coefficient of correlation of scores on the test given at the conclusion of the course with the composite grades for the two semesters. A validity coefficient of .624 was obtained, which has a predictive value of 21.9 per cent better than chance according to tables supplied by Bingham (2). The students' scores on this objective test were not used in determining the letter grades for the course so self-correlation did not increase the validity coefficient.

Guilford (10), Bingham (2) and others emphasize the importance of securing adequate criteria. They regard this as one of the most difficult aspects in the validition of tests. Bingham (2) indicates that failure to secure a high coefficient of validity for a test is due not only to the lack of validity and reliability of the test itself, but also to the lack of either perfect reliability or validity of the criterion. In determining the validity of the biology test, semester grades supplied a coarse grouping of criterion scores which could have been improved if the scores on the tests which made up the letter grades for the course had been available to permit more precise grouping. Since both the biology test and grades in the course

were intended to measure basic knowledge in biology the correlation coefficient of .624 of grades for the course and test scores at the end of the course suggests that the test is a fairly valid measure of basic knowledge in biology.

The possibility of predicting grades in the course by scores on the test given before starting the course was investigated by computing the Pearson product-moment coefficient of correlation of composite grades for the two semesters and test scores at the beginning of the course. A coefficient of .38 was obtained which has a predictive value of 5.25 per cent better than chance according to tables supplied by Bingham (2). A survey of the distribution of scores suggested that the limited value of the test for predicting grades when given before the course did not reflect a weakness of the test so much as it indicated that basic knowledge of biology prior to the course is not essential to success in the course. The number of students who had low grades on the preliminary test but received high grades in the course suggests that good students could succeed in the course without previous training in biology. This conclusion is supported by the fact that no prerequisites are required for the course.

Comparison of Item Validities

The validity of each item as expressed by its correlation with the total test score was computed by the Toops-Adkins

method. These validity coefficients are shown in Table 1. They ranged from -.039 to .504. Five of the validity coefficients were above .400; 12 were between .300 and .400; and 43 were between .200 and .300.

Key to Table 1

a - number of right responses chosen by students in highest fifth

b - number of right responses chosen by students in second fifth

c - number of right responses chosen by students in third fifth

d - number of right responses chosen by students in fourth fifth

e - number of right responses chosen by students in lowest fifth

R - total right responses

W - total wrong responses

2RW - product of right and wrong answers multiplied by 2

$\sqrt{2RW}$ - denominator of Toops-Adkins formula

r - correlation coefficient of item and criterion

$2s+b-d-2e$ - numerator of Toops-Adkins formula

Table 1. Validity of test items by Toops-Adkins method of correlation of item and total test score.

Item	a	b	c	d	e	R	W	2RW	$\sqrt{2RW}$	2a+b -d-2e	r
1	72	72	77	72	75	368	32	23552	153	-6	-.039
2	51	53	42	39	44	229	171	78318	280	28	.100
3	78	78	79	78	78	391	9	7038	84	0	.000
4	70	67	62	63	50	312	88	54912	234	44	.188
5	79	78	79	78	78	392	8	6272	79	2	.025
6	63	54	52	38	32	239	161	76958	277	78	.282*
7	49	41	33	44	34	195	205	79950	282	38	.135
8	64	64	63	52	44	287	113	64862	254	52	.205*
9	46	35	35	20	16	152	248	75392	274	75	.274*
10	79	77	75	73	61	365	35	25550	160	40	.250*
11	19	19	24	18	16	96	304	58368	241	7	.029
12	63	53	38	39	33	226	174	78648	280	74	.264*
13	64	48	47	25	17	191	209	79838	282	117	.415*
14	76	71	63	60	60	330	70	46200	215	43	.200*
15	49	40	21	30	12	152	248	75392	274	84	.307*
16	73	63	63	47	44	290	110	63800	252	74	.292*
17	43	47	41	46	22	187	213	79662	282	57	.202*
18	71	70	50	51	31	273	127	69342	263	99	.376*
19	30	24	10	17	19	100	300	60000	245	29	.118
20	55	41	38	34	20	188	212	79712	282	77	.273*
21	40	20	22	20	8	110	290	63800	252	64	.254*
22	47	37	20	25	10	139	261	72558	269	86	.320*
23	6	6	1	0	4	17	383	13022	114	10	.088
24	31	21	27	21	8	108	292	63072	251	46	.183
25	47	28	18	13	8	114	286	65208	255	93	.365*
26	75	73	63	59	60	330	70	46200	215	44	.205*
27	75	62	40	33	19	229	171	78318	280	141	.504*
28	53	42	31	41	37	204	196	79968	283	33	.117
29	77	74	72	70	60	353	47	33182	182	38	.209*
30	44	18	16	10	17	105	295	61950	249	62	.249*
31	34	30	26	22	18	130	270	70200	265	40	.151
32	42	33	20	21	12	128	272	69632	264	72	.273*
33	19	8	14	5	8	54	346	37368	193	25	.129
34	6	4	1	1	2	14	386	10808	104	11	.106
35	59	35	20	16	13	143	257	73502	271	111	.410*
36	79	76	68	70	56	347	53	36888	192	50	.258*
37	13	14	14	9	9	59	341	40238	201	13	.065
38	65	59	58	46	46	274	126	69048	263	51	.194
39	77	76	74	71	65	363	37	26862	164	29	.177
40	48	36	24	27	22	157	243	76302	276	61	.221*
41	78	74	70	72	58	352	48	33792	184	42	.228*
42	60	47	48	37	38	230	170	78200	280	54	.193
43	80	78	79	77	73	387	13	10062	100	15	.150
44	75	75	65	68	65	348	52	36192	190	27	.142

Table 1. (cont.).

Item	a	b	c	d	e	R	W	2RW	$\sqrt{2RW}$	2a+b-d-2e	r
45	78	75	69	60	39	321	79	50580	225	93	.413*
46	61	50	58	52	38	259	141	73038	270	44	.163
47	74	59	59	57	55	304	96	58368	242	40	.165
48	76	69	74	62	55	336	64	43008	207	49	.237*
49	53	31	27	14	5	130	270	70200	265	113	.426*
50	72	64	60	45	36	277	123	68142	261	91	.349*
51	80	79	77	79	79	394	6	4728	69	2	.028
52	79	80	79	78	76	392	8	6272	79	8	.101*
53	77	74	71	71	60	353	47	33182	182	37	.203*
54	73	67	60	53	53	306	94	57528	240	54	.225*
55	68	53	54	49	31	255	145	73950	272	78	.286*
56	60	41	35	38	18	192	208	79872	283	87	.307*
57	74	74	66	62	55	331	69	45678	214	50	.233*
58	78	79	73	75	68	373	27	20142	142	24	.169
59	80	78	78	79	73	388	12	9312	96	13	.135
60	80	79	79	78	73	389	11	8558	92	15	.163*
61	80	78	74	76	65	373	27	20142	142	32	.225*
62	78	72	74	62	55	341	59	40238	201	56	.278*
63	76	73	75	62	51	337	63	42462	206	61	.296*
64	57	47	45	47	46	242	158	76472	276	22	.078*
65	80	69	66	71	49	335	65	43550	209	60	.287*
66	78	76	70	68	54	346	54	37368	193	56	.290*
67	54	40	38	24	20	176	224	78848	281	84	.299*
68	78	58	55	45	40	272	128	69632	264	91	.344*
69	75	70	54	54	37	288	112	64512	254	88	.346*
70	55	48	44	45	34	226	174	78648	280	45	.160*
71	62	37	30	27	16	172	228	78432	280	102	.365*
72	60	53	40	42	39	234	166	77688	279	53	.190
73	75	65	55	60	45	300	100	60000	245	65	.265*
74	57	54	60	59	51	281	119	66878	259	7	.027
75	30	35	25	22	20	132	268	70752	266	33	.124
76	80	79	77	77	70	383	17	13022	114	22	.192*
77	33	20	18	13	10	94	306	57528	240	53	.221*
78	34	32	26	26	13	131	269	70478	266	48	.180*
79	74	66	66	63	39	308	92	56672	238	73	.306*
80	63	50	47	43	33	236	164	77408	278	67	.241*
81	45	39	30	33	19	166	234	76588	279	58	.208*
82	79	80	79	78	73	389	11	8558	92	14	.152*
83	68	60	57	47	36	268	132	70752	266	77	.288*
84	72	67	60	47	50	298	102	60792	247	62	.251*
85	79	75	75	67	65	361	39	28158	168	36	.214*
86	47	44	38	32	22	183	217	79422	282	62	.219*
87	71	47	42	47	26	233	167	77822	279	90	.322*
88	77	63	61	60	48	309	91	56238	237	61	.257*
89	59	50	54	55	38	256	144	73728	272	37	.136
90	58	45	40	32	27	202	198	79992	283	75	.266*

Table 1. (concl.).

Item	a	b	c	d	e	R	W	2RW	$\sqrt{2RW}$	2a+b-d-2e	r
91	67	59	63	60	45	294	106	62328	250	43	.172
92	73	59	63	55	40	290	110	63800	253	70	.276*
93	75	71	67	67	54	334	66	44088	210	46	.219*
94	73	72	71	64	58	338	62	41912	205	38	.185
95	18	17	24	16	20	95	305	57950	241	- 3	-.012
96	72	72	70	59	47	320	80	51200	226	63	.279*
97	12	14	15	15	13	69	331	45678	214	- 3	-.014
98	51	27	24	17	17	136	264	71808	268	78	.291*
99	39	17	15	6	7	84	316	53088	231	75	.325*
100	75	65	71	55	44	308	92	56672	238	70	.294*

*one of the 60 questions with highest validities.

Order of Difficulty

The arrangement of items in rank order of difficulty, easiest items first, is shown in Table 2 with the validity coefficient for each item.

Because some difference of opinion was found among other writers regarding the relationship of item validity and item difficulty, their relationship in the biology test was studied. After the test items were ranked in order of difficulty they were also ranked in order of validity. The Spearman formula, as given by Kelley (16), for rank order correlation was applied and a rho coefficient of .08 was obtained. A correction is necessary to make the Spearman rho strictly comparable to the Pearson r. This correction was made according to a table supplied by Guilford (10). The corrected coefficient was .084. The standard error of rho computed according to the formula given by Guilford (10) was .105. At the one per cent level of confidence the limits of the true rho are -.192 and .352. This coefficient of correlation suggests that there is no significant rectilinear relationship between item validity and item difficulty in the biology test.

The possibility of a curvilinear relationship was investigated by the critical ratio technic. The mean item validity of the group of questions composed of the 25 easiest items and the 25 most difficult items was .189 as compared with the mean item validity of .245 of the 50 questions of medium difficulty. The standard error of the difference between the means was .0195

which yielded a critical ratio of 2.87. This indicated that the difference in validity in favor of items of medium difficulty was significant at the 0.5 per cent level of confidence.

Table 2. Order of difficulty of test items.

Rank of item	Number of item on test	Number of right responses	Validity of item
1	51	394	.028
2.5	5	392	.025
2.5	52	392	.101
4	3	391	.000
5.5	80	389	.163
5.5	82	389	.152
7	59	388	.155
8	43	387	.150
9	76	383	.192
10.5	58	375	.169
10.5	61	375	.225
12	1	368	-.039
13	10	365	.250
14	39	365	.177
15	85	361	.214
16.5	29	353	.209
16.5	53	353	.203
18	41	352	.228
19	44	348	.142
20	36	347	.258
21	86	346	.290
22	62	341	.278
23	94	338	.185
24	63	337	.296
25	48	336	.237
26	65	335	.287
27	93	334	.219
28	57	331	.233
29.5	14	330	.200
29.5	26	330	.205
31	45	321	.413
32	96	320	.279
33	4	312	.188
34	88	309	.257
35	100	308	.294
36	54	306	.225
37	47	304	.165

Table 2. (cont.).

Rank of item	Number of item on test	Number of right responses	Validity of item
38	73	300	.265
39	79	299	.306
40	84	288	.251
41.5	16	290	.292
41.5	92	290	.276
43	69	288	.346
44	8	287	.205
45	74	281	.027
46	50	277	.349
47	38	274	.194
48	18	273	.376
49	68	272	.344
50	83	268	.286
51	46	259	.163
52	89	256	.136
53	55	255	.286
54	64	243	.078
55	6	239	.282
56	80	237	.241
57	72	234	.190
58	87	233	.322
59	42	230	.193
60.5	2	229	.100
60.5	27	229	.504
62.5	12	226	.264
62.5	70	226	.160
64	28	204	.117
65	90	202	.266
66	7	195	.135
67	91	194	.172
68	56	192	.307
69	13	191	.415
70	20	188	.273
71	17	187	.202
72	86	185	.209
73	67	176	.299
74	71	172	.363
75	81	166	.208
76	40	157	.221
77.5	9	152	.274
77.5	15	152	.307
79	35	145	.410
80	22	139	.320
81	98	136	.291
82	75	132	.124

Table 2. (concl.).

Rank of item	Number of item on test	Number of right responses	Validity of item
83	78	131	.130
84.5	51	130	.151
84.5	49	130	.426
86	32	128	.273
87	28	114	.365
88	21	110	.254
89	24	108	.193
90	30	105	.249
91	19	100	.113
92	11	97	.029
93	95	95	-.012
94	77	94	.221
95	99	84	.325
96	97	69	-.014
97	37	59	.066
98	33	54	.130
99	23	17	.088
100	34	14	.106

Choice of Responses

The choice of optional responses was analyzed for the purpose of discovering possible revisions to increase item validity. This analysis is shown in Table 3. The Toops-Adkins method (1) was applied to determine the validity of each response.

Alternatives which were chosen by none or almost none of the students should, as Turnbull (25) recommended, be made more plausible because, if they do not attract any one, they contribute nothing to the test. A good example of this is found in Item 40, where 230 students chose response "a"; 157 chose response "b", the right answer; ten students omitted the item; and

only three were attracted to any of the other three wrong responses. Even though five optional responses were offered, it was in effect a two response item. Where possible the questions with less than five options, such as 36 and 44, should be lengthened to make the test uniform, but offering additional options would not serve the desired purpose unless they could be made plausible enough to attract some of the students.

The principle of making unused responses more plausible might also be applied to items which had low validity because nearly all the students chose the right answer.

According to Adkins and Toops (1), the right answer should have a positive validity coefficient as high as possible, and the wrong responses should have low or negative validity coefficients. Any wrong response which has an appreciable positive validity coefficient because it attracts a larger proportion of the better students than of the poorer students should be revised. Examples of such responses are found in 17 c, 25 c, 95 c, and 97 d. Such items are not valid if a wrong response is similar enough to the right answer that it attracts greater numbers of the better students who have some knowledge of the subject, than of the poorer students who divide their choices more evenly among the other wrong answers.

The possibility of changing the wording of the item should be considered in cases where an unusually large number of all students omitted the item. An example is Item 11 which was omitted by more than half of the students. Changing the wording

might also improve items which seemed too difficult because so few answered them correctly. Ruch (22) has recommended the use of "simple everyday words in preference to more technical or literary synonyms." An example of this type may be found in Item 34 which could probably be improved by substituting "changes" for the key word "transforms".

Faint or illegible mimeographing of the right answer may have been a contributing factor in the low validity of Item 95.

The tendency of students not to read everything is well illustrated by the responses to Item 1. This item with the correct answer was given as an illustration on the instruction sheet, but in spite of this it ranked twelfth in difficulty and was missed by more of the good students than of the poorer students.

Ruch (22) recommended placing the correct response in each position an approximately equal number of times. In the biology test the correct response appeared as "a" 17 times, as "b" 21 times, as "c" 24 times, as "d" 24 times and as "e" 14 times. As an optional response "e" was offered with only 65 items that response was the correct one in a fair proportion of items, but the 17 times the correct answer was placed in position "a" was less than should have occured in a chance distribution.

Ruch (22) also recommended that the same response should not appear in the same position more than two or three successive times. This was violated only once in the biology test when "c" was the correct response to Items 46, 47, 48, and 49.

Key to Table 5

Horizontal Headings

I - Item number in test

Ch. - Optional choices

A - Choices by highest fifth of students

B - Choices by second fifth of students

C - Choices by third fifth of students

D - Choices by fourth fifth of students

E - Choices by lowest fifth of students

r - Validity of response by Toops-Adkins method

Vertical Code

a - Optional answer "a"

b - Optional answer "b"

c - Optional answer "c"

d - Optional answer "d"

e - Optional answer "e"

o - Omission of choice of answers

\- - Optional answer not offered

Table 3. Analysis of choice of responses by the Toops-Adkins method.

I	Ch.	A	B	C	D	E	r	I	Ch.	A	B	C	D	E	r
1	a	3	1	2	4	0	.034	7	a	0	0	0	1	0	-.036
	b	2	7	1	3	3	.018		b	0	1	2	2	2	-.087
	c	2	0	0	1	1	.018		c	31	37	38	40	35	-.039
	*d	72	72	77	72	75	-.039		*d	49	41	33	44	34	.135
	-	-	-	-	-	-	--		-	-	-	-	-	-	--
	o	1	0	0	0	1	.000		o	0	1	2	4	9	-.189
2	a	19	23	26	26	23	-.043	8	a	13	9	9	16	12	-.085
	b	4	0	2	8	5	-.083		b	2	3	5	6	6	-.085
	c	2	1	3	5	2	-.040		*c	64	64	63	52	44	.205
	*d	51	53	42	39	44	.100		d	1	1	1	2	2	-.041
	-	-	-	-	-	-	-		e	0	1	0	2	2	-.079
	o	4	3	7	2	6	.015		o	0	2	2	2	14	-.228
3	a	0	0	0	0	0	.000	9	a	1	0	1	0	1	.000
	b	1	0	1	2	0	.000		b	27	31	30	36	25	.014
	*c	78	78	79	78	78	.000		c	0	1	2	1	6	-.137
	d	1	1	0	0	0	.075		*d	46	35	35	20	16	.274
	-	-	-	-	-	-	-		e	4	4	2	3	3	.027
	o	0	1	0	0	2	-.020		o	2	9	10	20	31	-.316
4	a	0	0	1	0	1	-.050	10	a	0	0	0	1	2	-.079
	b	1	2	2	2	9	-.144		b	0	0	0	0	2	-.071
	*c	70	67	62	63	50	.188		c	0	1	3	2	4	-.102
	d	8	11	13	13	13	-.063		*d	79	77	75	73	61	.250
	-	-	-	-	-	-	-		e	0	0	1	1	1	-.061
	o	1	0	2	2	7	-.145		o	1	2	1	3	10	-.102
5	a	0	0	0	0	0	.000	11	a	7	16	7	9	16	-.057
	*b	79	78	79	78	78	.025		b	1	0	3	3	2	-.059
	c	1	0	1	0	0	.050		*c	19	19	24	18	16	.029
	d	0	1	0	0	2	-.061		d	7	4	7	4	3	.058
	-	-	-	-	-	-	-		e	5	3	2	0	1	.119
	o	0	1	0	2	0	-.020		o	40	37	37	46	42	-.046
6	a	2	4	2	7	7	-.100	12	a	6	6	16	9	8	-.039
	b	2	3	4	2	1	.031		*b	63	53	38	39	33	.264
	c	11	9	12	20	22	-.150		c	2	1	0	5	1	-.024
	*d	63	54	52	38	32	.282		d	0	1	1	0	4	-.101
	-	-	-	-	-	-	-		e	2	5	7	4	0	.042
	o	2	10	10	13	18	-.182		o	7	14	18	23	34	-.254

Table 3. (cont.).

I	Ch.	A	B	C	D	E	r	I	Ch.	A	B	C	D	E	r
13	a	2	0	5	3	4	-.067	19	a	41	32	46	38	35	.021
	b	7	16	29	28	30	-.229		b	6	15	11	8	8	.016
	*c	64	48	47	25	17	.415		c	1	1	3	4	13	-.073
	d	2	3	3	11	6	-.117		*d	30	24	10	17	19	.118
	-	-	-	-	-	-	-		e	0	0	6	5	6	-.149
	o	5	13	6	13	23	-.178		o	2	8	4	8	9	-.093
14	a	4	6	11	12	11	-.113	20	*a	55	41	38	34	20	.273
	b	0	1	1	3	3	-.101		b	7	11	13	12	11	-.046
	c	0	0	2	0	1	-.041		c	2	3	3	6	2	-.027
	d	0	0	2	3	2	-.094		d	6	6	7	3	14	-.080
	*e	76	71	63	60	60	.200		e	6	11	11	13	13	-.129
	o	0	2	1	2	3	-.071		o	4	8	8	12	15	-.143
15	a	4	8	13	7	11	-.074	21	a	5	8	6	9	5	-.006
	*b	49	40	21	30	12	.307		b	11	10	11	6	11	.022
	c	14	12	15	15	19	-.059		c	8	9	6	13	7	-.011
	d	5	7	16	10	12	-.091		d	12	13	17	13	19	-.064
	e	6	10	14	14	6	-.021		*e	40	20	22	20	8	.254
	o	1	3	6	4	10	-.142		o	4	20	18	19	30	-.215
16	a	0	0	1	0	0	.000	22	a	16	24	33	25	28	-.098
	b	4	14	12	23	24	-.239		b	0	0	1	1	1	-.061
	c	1	1	0	2	3	-.068		c	0	2	0	2	0	.000
	d	2	1	3	1	3	-.023		d	0	0	1	0	1	-.050
	*e	73	63	63	47	44	.292		*e	47	37	20	25	10	.320
	o	0	1	1	1	6	-.143		o	17	17	25	27	40	-.214
17	a	1	3	4	6	7	-.119	23	*a	6	6	1	0	4	.088
	b	3	5	7	5	5	-.029		b	6	10	9	11	11	-.060
	c	0	0	0	2	2	-.108		c	47	32	42	36	27	.135
	*d	45	47	41	46	22	.202		d	2	1	2	1	2	.000
	e	27	11	11	8	10	.175		-	-	-	-	-	-	-
	o	6	14	17	23	36	-.161		o	19	31	26	32	36	-.128
18	a	1	1	4	10	10	-.193	24	a	9	9	12	15	20	-.134
	*b	71	70	50	51	31	.376		*b	31	21	27	21	8	.183
	c	1	3	9	8	8	-.129		c	7	13	10	11	5	.033
	d	2	2	6	3	2	-.009		d	17	5	12	13	26	-.119
	e	1	3	6	3	7	-.097		-	-	-	-	-	-	-
	o	4	1	5	5	22	-.244		o	16	22	19	20	21	-.033

Table 3. (cont.).

I	Ch.	A	B	C	D	E	r	I	Ch.	A	B	C	D	E	r
25	a	18	32	36	30	30	-.081	31	a	3	6	6	12	5	-.065
	*b	47	28	18	13	8	.365		*b	34	30	26	22	18	.151
	c	7	10	13	19	14	-.112		c	8	12	18	18	24	-.168
	d	4	3	7	8	10	-.111		d	31	20	18	14	9	.210
	e	2	0	0	2	1	.000		e	2	7	5	5	10	-.095
	o	2	7	6	8	7	-.074		o	2	5	7	9	14	-.171
26	a	0	0	0	0	0	.000	32	a	2	8	20	13	15	-.155
	b	3	6	16	20	14	-.179		b	4	5	3	1	7	-.016
	*c	75	73	65	59	60	.205		c	20	14	15	15	7	.117
	d	1	1	0	0	2	-.018		*d	42	33	20	21	12	.273
	e	0	0	0	1	2	-.079		e	9	4	6	14	12	-.059
	o	1	0	1	0	2	-.018		o	3	16	18	16	27	-.212
27	*a	75	62	40	33	19	.504	33	a	13	28	34	37	25	-.085
	b	0	8	15	17	15	-.201		b	3	4	2	1	3	.030
	c	4	8	18	16	17	-.165		c	7	6	13	8	1	.062
	d	0	0	0	4	0	-.071		d	11	5	3	1	4	.134
	e	1	0	2	3	7	-.150		*e	19	8	14	5	8	.129
	o	0	2	4	7	22	-.306		o	22	29	14	28	29	.050
28	e	1	4	4	6	7	-.108	34	a	38	41	40	46	41	-.004
	b	2	6	8	3	6	-.036		b	30	21	31	28	25	.011
	c	6	3	6	6	5	-.007		c	1	4	2	0	5	-.042
	*d	53	42	31	41	37	.117		*d	6	4	1	1	2	.106
	e	12	17	22	13	14	.000		e	0	1	1	0	0	.025
	o	6	8	9	11	11	-.072		o	5	10	5	7	7	-.006
29	a	1	2	0	1	4	-.063	35	a	8	14	26	30	36	-.284
	b	2	4	7	8	16	-.195		b	0	7	2	7	4	-.065
	*c	77	74	72	70	60	.209		c	5	4	7	4	5	.000
	d	0	0	1	0	0	.000		d	1	9	11	3	7	-.040
	-	-	-	-	-	-	-		*e	59	35	20	16	13	.410
	o	0	0	0	1	0	-.036		o	8	11	14	22	25	-.199
30	a	18	39	40	46	36	-.155	36	a	1	6	12	9	23	-.250
	*b	44	18	16	10	17	.249		*b	79	74	68	70	56	.258
	c	5	5	8	11	7	-.062		c	0	0	0	0	1	-.071
	d	1	4	5	5	2	-.026		-	-	-	-	-	-	-
	e	4	1	2	1	1	.071		-	-	-	-	-	-	-
	o	8	13	9	7	17	-.063		o	0	0	0	1	0	-.036

Table 3. (cont.).

I	Ch.	A	B	C	D	E	r	I	Ch.	A	B	C	D	E	r
37	a	3	5	6	7	13	-.139	43	a	0	2	1	2	3	-.076
	b	15	12	16	14	12	.019		*b	80	78	79	77	73	.150
	c	4	7	9	13	15	-.133		c	0	0	0	0	1	-.071
	d	42	34	29	28	14	.227		d	0	0	0	1	2	-.079
	*e	13	14	14	9	9	.065		-	-	-	-	-	-	-
	o	3	8	6	9	19	-.184		o	0	0	0	0	1	-.071
38	a	0	3	2	2	2	-.036	44	*a	75	75	65	68	65	.142
	b	3	1	1	2	3	-.011		b	5	4	15	10	14	-.130
	c	4	3	5	5	6	-.045		c	0	0	0	0	1	-.071
	d	5	11	8	15	17	-.133		-	-	-	-	-	-	-
	*e	65	59	58	46	46	.194		-	-	-	-	-	-	-
	o	3	3	6	10	6	-.097		o	0	1	0	2	0	-.020
39	a	2	0	1	2	1	.000	45	a	0	1	2	6	5	-.143
	b	0	1	0	0	0	.036		b	0	0	0	0	0	.000
	c	0	0	1	7	3	-.141		c	2	1	8	11	26	-.315
	d	0	1	1	0	2	-.054		*d	78	75	69	60	39	.413
	*e	77	76	74	71	65	.177		-	-	-	-	-	-	-
	o	1	2	3	0	9	-.140		o	0	3	1	3	10	-.175
40	a	31	42	55	52	50	-.171	46	a	2	4	8	6	5	-.058
	*b	48	36	24	27	22	.221		b	3	5	3	5	7	-.061
	c	0	0	0	0	0	.000		*c	61	50	58	52	38	.165
	d	0	1	0	0	1	-.025		d	0	1	2	2	5	-.125
	e	0	1	0	0	0	.036		-	-	-	-	-	-	-
	o	1	0	1	1	7	-.148		o	14	20	9	15	25	-.074
41	*a	78	74	70	72	58	.228	47	a	5	17	15	16	15	-.089
	b	1	1	2	2	5	-.098		b	0	0	2	0	2	-.071
	c	0	3	3	2	2	-.033		*c	74	59	59	57	55	.165
	d	0	1	1	4	3	-.119		d	0	1	3	2	2	-.063
	e	0	0	1	0	0	.000		-	-	-	-	-	-	-
	o	1	1	3	0	12	-.184		o	1	3	1	5	6	-.108
42	a	9	12	12	13	18	-.098	48	a	0	2	1	4	6	-.140
	b	0	0	0	2	0	-.050		b	1	2	1	3	2	-.036
	*c	60	47	48	37	38	.195		*c	76	69	74	62	55	.257
	d	1	6	11	5	3	-.021		d	1	2	1	2	4	-.068
	e	6	4	3	11	4	-.021		-	-	-	-	-	-	-
	o	4	11	6	12	17	-.144		o	2	5	3	9	13	-.169

Table 3. (cont.).

I	Ch.	A	B	C	D	E	r	I	Ch.	A	B	C	D	E	r
49	a	8	16	15	13	19	-.088	55	a	10	23	21	21	30	-.153
	b	4	8	11	20	28	-.268		*b	68	53	54	49	31	.286
	*c	53	31	27	14	5	.426		c	0	2	2	2	0	.000
	d	15	21	22	29	17	-.049		d	0	0	0	2	3	-.127
	-	-	-	-	-	-	-		e	0	0	0	0	3	-.122
	o	0	4	4	4	11	-.166		o	2	2	3	6	13	-.185
50	a	3	6	6	5	13	-.121	56	a	5	5	3	6	12	-.099
	*b	72	64	60	45	36	.349		b	13	30	36	33	34	-.165
	c	4	9	10	12	8	-.063		*c	60	41	35	38	18	.307
	d	0	0	2	7	3	-.135		d	0	0	0	0	0	.000
	-	-	-	-	-	-	-		e	0	0	0	0	0	.000
	o	1	1	2	11	20	-.300		o	2	4	6	4	16	-.183
51	a	0	1	2	1	1	-.032	57	*a	74	74	66	62	55	.238
	b	0	0	1	0	0	.000		b	2	2	1	0	3	.000
	c	0	0	0	0	0	.000		c	3	3	10	10	9	-.119
	*d	80	79	77	79	79	.028		d	0	1	0	2	5	-.139
	e	0	0	0	0	0	.000		e	0	0	1	1	1	-.061
	o	0	0	0	0	0	.000		o	1	0	2	5	7	-.159
52	a	0	0	0	1	2	-.079	58	*a	78	79	73	75	68	.169
	b	0	0	1	0	0	.000		b	2	1	5	3	4	-.056
	c	0	0	0	0	0	.000		c	0	0	1	0	3	-.108
	*d	79	80	79	78	76	.101		d	0	0	0	0	0	.000
	e	0	0	0	0	0	.000		e	0	0	1	2	1	-.084
	o	1	0	0	1	2	-.054		o	0	0	0	0	4	-.143
53	a	0	0	0	0	0	.000	59	a	0	1	0	1	0	.000
	b	2	5	4	7	10	-.125		b	0	0	0	0	0	.000
	*c	77	74	71	71	60	.203		c	0	1	0	0	0	.036
	d	1	0	1	0	0	.050		d	0	0	0	0	0	.000
	e	0	1	3	2	6	-.135		*e	80	78	78	79	73	.135
	o	0	0	1	0	4	-.127		o	0	0	2	0	7	-.166
54	*a	73	67	60	53	53	.225	60	a	0	0	0	0	0	.000
	b	1	2	2	3	4	-.073		b	0	0	0	1	0	-.036
	c	3	5	11	10	12	-.134		c	0	0	0	0	0	.000
	d	1	0	4	6	3	-.096		*d	80	79	79	78	73	.163
	e	0	0	0	0	0	.000		e	0	1	1	1	4	-.108
	o	2	6	3	8	8	-.099		o	0	0	0	0	3	-.122

Table 3. (cont.).

I	Ch.	A	B	C	D	E	r	I	Ch.	A	B	C	D	E	r
61	a	0	0	4	0	6	-.138	67	a	1	7	8	9	11	-.137
	*b	80	78	74	76	65	.225		b	11	17	15	22	23	-.124
	c	0	1	0	0	2	-.020		*c	54	40	38	24	20	.299
	d	0	1	1	3	1	-.058		d	6	10	11	15	15	-.116
	e	0	0	0	0	0	.000		-	-	-	-	-	-	-
	o	0	0	1	1	6	-.165		o	8	6	9	10	11	-.056
62	a	0	0	0	1	0	-.036	68	a	2	10	9	15	14	-.156
	b	0	0	2	0	0	.000		b	0	0	0	0	1	-.071
	c	1	6	2	6	10	-.131		*c	78	68	53	43	40	.344
	d	1	2	2	10	15	-.242		d	0	3	6	1	5	-.075
	*e	78	72	74	62	55	.273		e	0	1	5	5	3	-.096
	o	0	0	0	1	0	-.036		o	0	8	7	16	17	-.228
63	a	0	0	0	4	1	-.095	69	a	0	0	0	0	1	-.071
	b	0	0	0	0	0	.000		b	2	3	8	17	19	-.259
	c	2	4	4	10	11	-.172		c	0	0	2	0	0	.000
	*d	76	73	75	62	51	.296		d	4	5	12	6	11	-.090
	e	1	1	0	1	4	-.081		*e	73	70	54	54	37	.346
	o	1	2	1	3	13	-.203		o	1	2	4	3	12	-.178
64	a	2	7	7	3	3	.015	70	a	9	13	14	12	19	-.190
	*b	57	47	45	47	46	.078		*b	55	48	44	45	34	.160
	c	0	2	3	4	3	-.183		c	2	5	4	7	4	-.046
	d	11	10	7	9	11	.005		d	13	7	13	11	14	-.030
	e	2	5	7	2	3	.008		-	-	-	-	-	-	-
	o	8	9	11	15	14	-.091		o	1	7	5	5	9	-.099
65	a	0	3	7	2	9	-.135	71	a	2	11	5	9	7	-.051
	b	0	1	1	0	0	.025		b	9	12	18	18	16	-.091
	c	0	4	4	4	6	-.102		*c	62	37	30	27	16	.365
	*d	80	69	66	71	49	.287		d	1	6	8	3	7	-.066
	e	0	0	0	0	0	.000		-	-	-	-	-	-	-
	o	0	3	2	3	16	-.239		o	6	14	19	23	34	-.268
66	a	2	3	7	8	19	-.232	72	a	17	24	23	21	25	-.051
	b	0	1	2	1	0	.000		*b	60	53	40	42	39	.190
	c	0	0	1	0	0	.000		c	2	1	5	4	5	-.077
	*d	78	76	70	68	54	.290		d	1	1	6	8	5	-.119
	e	0	0	0	1	0	-.036		-	-	-	-	-	-	-
	o	0	0	0	2	7	-.098		o	0	1	6	5	6	-.157

Table 3. (cont.).

I	Ch.	A	B	C	D	E	r	I	Ch.	A	B	C	D	E	r
73	*a	75	65	55	60	45	.265	79	*a	74	66	66	63	39	.306
	b	0	6	8	8	10	-.144		b	0	3	4	1	8	-.126
	c	0	3	1	1	1	.000		c	0	0	1	5	0	-.072
	d	4	0	11	8	16	-.190		d	0	0	0	0	4	-.071
	-	-	-	-	-	-	-		e	5	8	6	4	14	-.085
	e	1	6	5	3	8	-.083		o	1	3	3	7	15	-.217
74	*a	57	54	60	59	51	.027	80	*a	63	50	47	43	33	.241
	b	12	13	8	11	5	.086		b	2	4	7	9	12	-.158
	c	0	0	0	0	1	-.071		c	1	4	1	1	4	-.033
	d	1	1	2	1	0	.063		d	1	5	0	0	4	-.011
	e	3	5	6	1	5	.000		e	10	15	17	17	16	-.018
	o	7	7	4	8	18	-.130		o	3	2	8	10	11	-.152
75	a	3	5	8	6	4	-.021	81	a	10	3	5	4	5	.063
	b	1	1	2	3	2	-.048		*b	45	39	30	33	19	.208
	c	10	10	7	12	15	-.062		c	4	16	15	9	19	-.112
	*d	30	35	25	22	20	.124		d	0	0	0	1	0	-.036
	e	20	11	26	17	18	-.008		e	0	0	0	2	2	-.108
	o	16	18	12	20	21	-.051		o	21	22	30	31	35	-.157
76	a	0	0	0	0	0	.000	82	a	0	0	0	0	0	.000
	b	0	0	0	1	4	-.071		b	0	0	0	0	0	.000
	c	0	1	3	2	3	-.083		c	1	0	0	0	0	.071
	*d	80	79	77	77	70	.192		d	0	0	0	1	0	-.036
	e	0	0	0	0	2	-.100		*e	79	80	79	78	73	.152
	o	0	0	0	0	1	-.071		o	0	0	1	1	7	-.072
77	*a	33	20	18	13	10	.221	83	a	6	12	13	8	18	-.101
	b	7	11	12	10	11	-.037		*b	68	60	57	47	36	.286
	c	11	10	10	13	11	-.015		c	2	1	0	5	1	-.024
	d	5	11	12	14	12	-.088		d	2	1	4	10	9	-.164
	e	6	4	8	9	1	.035		e	1	1	3	2	2	-.036
	o	13	24	20	21	35	-.120		o	1	5	3	8	15	-.203
78	a	12	10	20	18	23	-.145	84	a	0	1	3	0	1	-.016
	b	9	8	11	5	9	.017		b	3	5	6	11	12	-.146
	*c	34	32	26	26	13	.180		c	0	3	4	4	1	-.031
	d	13	7	6	10	8	.039		*d	72	67	60	47	50	.251
	e	1	1	2	0	1	.016		e	0	0	0	1	1	-.061
	o	12	22	15	21	26	-.112		o	5	4	7	15	15	-.172

Table 3. (cont.).

I	Ch.	A	B	C	D	E	r	I	Ch.	A	B	C	D	E	r
85	a	0	4	1	6	4	-.092	91	*a	67	59	63	60	45	.172
	b	1	0	2	3	8	-.163		b	7	9	9	5	8	-.012
	c	0	1	0	0	1	-.025		c	3	1	4	3	6	-.070
	d	0	0	1	1	2	-.089		d	1	3	1	2	7	-.106
	*e	79	75	75	67	65	.214		-	-	-	-	-	-	-
	o	0	0	1	3	0	-.054		o	2	8	3	10	14	-.158
86	a	0	4	2	2	9	-.150	92	a	3	9	9	8	16	-.139
	b	27	22	18	17	10	.158		b	2	5	4	4	7	-.069
	c	3	4	13	16	25	-.276		*c	73	59	63	55	40	.276
	*d	47	44	38	32	22	.219		d	2	4	3	8	10	-.141
	e	1	6	7	7	9	-.114		-	-	-	-	-	-	-
	o	2	0	2	6	5	-.112		o	0	3	1	5	7	-.067
87	a	0	2	5	0	4	-.065	93	a	0	1	4	3	2	-.068
	b	3	7	5	5	11	-.092		*b	75	71	67	67	54	.219
	*c	71	47	42	47	26	.322		c	1	1	1	8	5	-.135
	d	1	0	0	1	2	-.054		d	3	5	7	0	13	-.104
	e	2	5	6	5	0	.034		e	1	1	0	1	3	-.058
	o	3	19	22	22	37	-.271		o	0	1	1	4	3	-.107
88	a	0	1	0	0	0	.026	94	*a	73	72	71	64	58	.185
	b	1	5	4	2	4	-.027		b	2	3	7	11	15	-.127
	*c	77	63	61	60	48	.257		c	0	1	0	0	1	-.025
	d	0	11	13	17	23	-.251		d	2	0	1	3	2	-.025
	-	-	-	-	-	-	-		-	-	-	-	-	-	-
	o	2	0	2	1	5	-.078		o	3	4	1	2	4	.000
89	a	0	2	2	1	1	-.015	95	a	4	2	0	0	1	.108
	b	18	17	20	15	21	-.017		b	7	6	8	8	12	-.069
	*c	59	50	54	55	38	.136		c	40	43	38	39	29	.067
	d	0	4	0	4	11	-.183		*d	18	17	24	16	20	-.012
	-	-	-	-	-	-	-		-	-	-	-	-	-	-
	o	3	7	4	5	9	-.069		o	11	12	10	17	18	-.089
90	a	1	0	0	6	10	-.211	96	a	6	4	4	5	4	.023
	b	3	11	11	11	8	-.056		*b	72	72	70	59	47	.279
	c	13	17	18	18	17	-.039		c	0	1	1	4	3	-.107
	*d	56	45	40	32	27	.266		d	0	1	2	4	7	-.163
	-	-	-	-	-	-	-		-	-	-	-	-	-	-
	o	5	7	11	13	18	-.166		o	2	2	3	8	19	-.252

Table 3. (concl.).

I	Ch.	A	B	C	D	E	r	I	Ch.	A	B	C	D	E	r
97	*a	12	14	15	15	13	-.014	99	*a	39	17	15	6	7	.325
	b	4	5	5	7	7	-.055		b	2	0	2	3	1	-.013
	c	5	6	6	3	8	-.020		c	23	33	41	45	38	-.131
	d	50	48	44	42	31	.238		d	8	16	15	17	21	-.127
	-	-	-	-	-	-	-		-	-	-	-	-	-	-
	o	8	7	10	13	21	-.139		o	8	9	7	9	13	-.056
98	a	1	3	4	2	2	-.010	100	a	3	7	5	15	17	-.197
	b	5	12	10	9	7	-.023		b	1	4	2	3	2	-.010
	*c	51	27	24	17	17	.291		*c	75	63	71	55	44	.294
	d	0	4	6	9	6	-.124		d	0	2	1	0	6	-.119
	e	2	1	1	0	2	.015		-	-	-	-	-	-	-
	o	21	33	35	43	46	-.215		o	1	4	1	7	10	-.159

*-Right response to the question.

Revised Test of Sixty Items

For experimental purposes the 60 items having the highest validity coefficients were selected and all answer sheets were scored again on the basis of these items only.

The validity of this 60 question test was determined by computing the Pearson product-moment coefficient of correlation of test scores and grades for the two semesters in Biology in Relation to Man. The obtained validity coefficient of .656 had a predictive value of 24.53 per cent efficiency according to tables supplied by Bingham (2). This validity coefficient was significantly higher than the coefficient of .624 for the entire 100 item test. The standard error of the difference between the two r's was .0095 which yielded a critical ratio of 3.37 indicating that there were only 14 chances in 1000 that the difference was due to sampling error.

The fact that the correlation with an independent criterion was significantly increased by the elimination of items having low validity by the criterion of internal consistency suggests that some of the criticism of the method of internal consistency is not applicable in all cases.

The Pearson product-moment correlation coefficient between the scores on the 100 item test and the 60 item test was .953.

The reliability coefficient of the 60 question test as determined by the Kuder-Richardson formula was .838 as compared with .832 for the 100 question test by the same method. Guil-

ford (10) has pointed out that there is an increase in reliability with an increase in the length of a test. Therefore, the slight increase in reliability in spite of the reduction in length of the test was significant. The Spearman-Brown prophecy formula as given by Guilford (10) indicated that a test of 100 items homogeneous with the 60 items would have a reliability coefficient of .896.

The more reliable a valid test becomes, the higher its validity coefficient may be expected to be if other variables remain the same. Therefore, increasing the length of a test with homogeneous items may increase its validity coefficient. Formulas for estimating the validity coefficient of a test when lengthened are given by Edgerton and Toops (9), Guilford (10), Kelley (16) and Lindquist (19). Edgerton and Toops (9) also furnished tables from which the new validity and reliability coefficients of a test may readily be computed when a test of known validity and reliability is increased by two to 15 times its length. All the formulas are based on the same principle and yielded a validity coefficient of .678 for a 100 item test consisting of items homogeneous with the 60 item test, as compared with the validity coefficient of .656 for t he 60 item test.

Reduction to a 60 item test was not recommended but revision of the optional responses to some items and substitution of new items for some others to maintain the 100 item length was suggested.

SUMMARY AND CONCLUSIONS

A 100 item objective test in biology was taken by Kansas State College students at the beginning of the course, Biology in Relation to Man, and again at the end of the two semester course. An item analysis of the test was made to obtain information for use in the refinement of the test before final validation and standardization.

The reliability coefficient of the test by the Kuder-Richardson formula was found to be .832.

A validity coefficient of .624 for the test was obtained by correlation of the test scores with grades for the two semesters in the course.

The Toops-Adkins method of item analysis was used to determine the validity of each item by its correlation with the total test score. Item validities ranged from -.039 to .504.

The relationship of item validity and item difficulty was investigated. The mean validity of items of medium difficulty was significantly higher than that of the extremely easy or extremely difficult questions.

The sixty items with the highest validities were selected and all answer sheets were rescored on the basis of these items only. The validity coefficient of the 60 question test obtained by correlation with grades was .656 which was significantly higher than the validity coefficient of .624 for the total 100 question test. The reliability coefficient by the Kuder-

Richardson method of the 60 item test was .838 which was slightly higher than the reliability coefficient of .832 for the 100 item test. The Spearman-Brown formula indicated that a 100 item test consisting of items homogeneous with these 60 questions would have a reliability coefficient of .896. The validity coefficient of a 100 item test homogeneous with the 60 item test, according to a formula given by Guilford (10) was estimated at .878.

The choice of responses to all items was analyzed by the Toops-Adkins method as a basis for improvement of items. Revision of test items was recommended by either eliminating or making more plausible the responses which were chosen by few or no students. Revision to reduce the similarity to the right answer or elimination of the response was recommended in cases where a wrong response had a relatively high validity coefficient.

The test as a whole met minimum standards as to reliability and validity but item analysis showed that it could be significantly improved by revision of optional responses to certain items and elimination of other items of low validity.

ACKNOWLEDGMENT

The writer wishes to acknowledge his deep indebtedness to Dr. J. C. Peterson, Professor of Psychology, for his invaluable suggestions regarding references, preparation and presentation of the data and for his help and encouragement regarding all phases of the thesis. Grateful acknowledgement is also made to Dr. M. J. Harbaugh, Professor of Zoology, who constructed and administered the test, for his cooperation in supplying data to be used. The writer is also indebted to Dr. M. Leigh Baker, Head of the Department of Education and Psychology, for the reading and criticism of the tentative copy of the thesis.

LITERATURE CITED

(1) Adkins, Dorothy C. and Herbert A. Toops.
Simplified formulas for item selection and construction.
Psychometrika. 2: 165-171. 1937.

(2) Bingham, Walter Van Dyke.
Aptitudes and aptitude testing. New York. Harper and Brothers. 1937.

(3) Bradley, Mary Edith.
A study of the validity of the Armed Forces Institute tests of General Educational Development in the field of social studies. Ed. and Psychol. Measurement. 6: 265-268. 1946.

(4) Brogden, Hubert E.
Variation in test validity with variation in the distribution of item difficulties, number of items and degree of their inter-correlation. Psychometrika. 11: 197-217. 1946.

(5) Committee on teaching of botany in American colleges and universities. An exploratory study in the teaching of botany in the colleges and universities of the United States. Botanical Society of America. Bull. 119. 46 p. 1938.

(6) Chesire, L., M. Saffir and L. L. Thurstone.
Computing diagrams for the tetrachoric correlation coefficient. Chicago. University of Chicago Bookstore. [59] p. 1933.

(7) Cronbach, Lee J.
A case study of the split-half reliability coefficient. Jour. Ed. Psychol. 37: 473-480. 1946.

(8) Cronbach, Lee J.
Test reliability: its meaning and determination.
Psychometrika. 12: 1-16. 1947.

(9) Edgerton, Harold A. and Herbert A. Toops.
A table for predicting the validity and reliability coefficients of a test when lengthened. Jour. Ed. Res. 18: 225-234. 1928.

(10) Guilford, J. P.
Fundamental Statistics in psychology and education. New York. McGraw-Hill. 333 p. 1942.

(11) Guttman, Louis.
The test-retest reliability of qualitative data.
Psychometrika. 11: 81-95. 1946.

(12) Hayes, Samuel P. Jr.
Diagrams for computing tetrachoric correlation coefficients from percentage differences. Psychometrika. 11: 163-172. 1946.

(13) Henry, Lorne J.
A comparison of the difficulty and validity of achievement test items. Jour. Ed. Psych. 25: 537-541. 1934.

(14) Horton, Clark W.
Achievement tests in relation to teaching objectives in general college botany. Botanical Society of America. Bull. 120. 71 p. 1939.

(15) Jurgansen, C. E.
Table for determining Phi coefficients. Psychometrika. 12: 17-29. 1947.

(16) Kelley, Truman L.
Statistical Method. New York. MacMillan. 390 p. 1923.

(17) Lentz, T. F. Jr., Bertha Hirshtein and F. H. Finch.
Evaluation of the methods of evaluating test items.
Jour. Ed. Psychol. 23: 344-350. 1932.

(18) Lindquist, E. F.
A first course in statistics. New York. Houghton Mifflin. 242 p. 1942.

(19) Lindquist, E. F.
Statistical analysis in educational research. New York. Houghton Mifflin. 266 p. 1940.

(20) Owens, William A.
An empirical study of the relationship between item validity and internal consistency. Ed. and Psychol. Measurement. 7: 281-288. 1947.

(21) Richardson, M. W. and G. F. Kuder.
The calculation of test reliability and coefficients based upon the method of rational equivalence. Jour. Ed. Psychol. 30: 681-687. 1939.

(22) Ruch, Giles M.
The objective or new type examination. New York. Scott Foresman. p. 318-357. 1929.

(23) Swineford, Frances.
Validity of test items. Jour. Ed. Psychol. 27: 68-78. 1936.

(24) Symonds, Percival M.
Choice of items for a test on the basis of difficulty. Jour. Ed. Psychol. 20: 481-493. 1929.

(25) Turnbull, William W.
A normalized graph method of item analysis. Jour. Ed. Psychol. 37: 473-480. 1946.

(26) Tyler, Ralph W.
Sound credit for military experience. Ann. Amer. Acad. of Polit. and Social Sci. 58-64. 1944.