PROGRESSIVE VALIDATION OF TEST ITEMS ON THE BASIS OF INTERNAL CONSISTENCY

by

EARL TODD GOODFELLOW

B. S., Kansas State College
of Agriculture and Applied Science, 1928

A THESIS

submitted in partial fulfillment of the

requirements for the degree of

MASTER OF SCIENCE

Department of Education

KANSAS STATE COLLEGE
OF AGRICULTURE AND APPLIED SCIENCE

1938

TABLE OF CONTENTS

INTRODUCTION

Extensive use of objective tests for measuring academic achievement has emphasized the need for an effective technique of identifying the best items for inclusion in a test. In the field of intelligence and aptitude tests this need has been met more fully than in the field of achievement testing. Fortunately, makers of achievement tests are now showing keen interest in the possibility of improving their tests in this manner.

The troublesome problem of how to identify the strongest test items has been the subject of several experiments in the past six years. Two types of methods have been advanced, subjective judgment on one hand, and statistical evaluation on the other. The statistical method, that is, correlation of the various items with some objective criterion, is the method to which most thought has been given. If the criterion selected is the score of the total test, then the correlation of each item with the total test will give indices of the validity of the various test items on the basis of internal consistency, that is, the degree of consistency with which each test item measures whatever the whole test measures.

To find the extent to which this internal consistency

may be improved by a series of successive validations of test items is the purpose of this experiment.

REVIEW OF LITERATURE

A review of the literature concerning test construction shows evidence that more attention is being given each year to the problem of item selection. The writer found several methods in use for finding the validity of individual test items. The methods found were:

1. Biserial r (5).
2. Upper Minus Lower Third (3).
3. Vincent's Overlapping (3).
4. McCall's Method (5).
5. Clark's Method (5).
6. Holzinger's Method (5).
7. The Summation of Agreement (3).
8. Difference Between Means (5).
9. Critical Ratio (5).
10. McCall, Long Bliss Method (4).
11. Long's Weighted Overlapping (4).
12. Boos' Method (4).
13. Toops' Adaptation of Pearson's Product-Moment Correlation Coefficient (6).

The common basis of all methods was the comparison of performance of strong students with that of weak students.

In 1932 Lentz, Hirshstein and Finch (3) made a study comparing the relative merits of four of the methods listed above: (a) The Upper Minus Lower Third Method; (b) Vincent's Overlapping Method; (c) McCall's Method; (d) the Summation of Agreement Method. A test of 150 statements of opinion was given to 211 college students who were asked to express their agreement or disagreement with the statements. The 100 best items were selected by each of the four methods. Each of the sets of 100 items was then tested for reliability, and the method giving the 100 items of highest reliability was judged to be the best method. The results of their experiment show that the best of these four methods was that of the Upper Minus Lower Third.

Lentz and Handy (2) published a later report on a similar study again comparing the Upper Minus Lower Third Method with other methods. They made use of the upper and lower three-tenths rather than the upper and lower one-third. Again this method as here adapted was found to yield the best results.

Henry (1) made a slightly different study in which he compared the difficulty and the validity of achievement

test items. Henry made use of the Biserial r Method, the Upper Minus Lower Third Method, Vincent's Overlapping Method, and Clark's Method. His results show that the use of Biserial r gave the best results with the Upper Minus Lower Third Method next best. Henry reports further, "That except for the extreme items, either too easy or too difficult, the difficulty of an item has little to do with its validity."

Swineford (5) has made a more extensive study by comparing eight selected methods for determining test item validity. Some writers use score on the whole test as the criterion with which to correlate items. Swineford points out that this procedure is permissible only if the test as a whole is known to be valid. The following methods were compared: (a) Biserial r; (b) McCall's Method; (c) Holzinger's Method; (d) Upper Minus Lower Third Method; (e) Vincent's Overlapping Method; (f) Clark's Method; (g) Difference Between Means Method; (h) Critical Ratio Method. Also, another characteristic of a test item known as balance was introduced into the comparison. An item is said to be in perfect balance if it is answered correctly by 50 per cent of the subjects. As the difficulty increases or decreases from this point, the balance decreases. Swineford found when working with 148 subjects

selected from 700 subjects on a test of 284 items that there was a significant correlation between balance, or difficulty and some of the methods. This is contrary to the findings of Henry (1). Swineford's conclusions recommended the use of Holzinger's Method and also the Difference Between Means Method. Biserial r gave results which agreed closely with the Difference Between Means Method, but it was rejected because of the time required in its computation.

Long and Sandiford (4) made use of 13 methods in still another comparison of methods for determining item validity. Like Swineford (5) they found that difficulty does enter into the validity of the items. They state that methods which tend to assign validity values uninfluenced by item difficulty seemed to be inferior to methods which tend to assign higher validity values as items approach 50 per cent difficulty. Their conclusion was that the method of Biserial r gave the best results, also that the method developed by Clark was superior to the other short methods.

From this review of the literature it can be seen that the various writers have recommended the following methods as giving the best results:

1. Biserial r.
2. Upper Minus Lower Third.
3. Holzinger's Method.
4. Difference Between Means.
5. Clark's Method.

The five methods recommended may be expressed in the following formulas:

1. Biserial r: $V = \frac{(M_p - M_q)\,pq}{\sigma_y Z}$

V is the validity of a single test item.
M_p is the mean criterion score of those persons who answer the item correctly.
M_q is the mean criterion score of those persons who answer the item incorrectly.
N is the number of persons taking the test.
p is the per cent answering correctly.
σ_y is the standard deviation of the distribution of the criterion scores.
Z is the ordinate which divides the unit area under the normal curve into two parts equal to p and q respectively.

2. Upper Minus Lower Third: $V = R_U - R_L$

R_U is the number of persons answering an item correctly in the upper one-third of the total group.
R_L is the number of persons in the lower one-third answering the item correctly.

3. Holzinger's Method: $V = \frac{(R_U + W_L) - (W_U + R_L)}{N/2}$

R_U is the number of right answers in the upper one-fourth of the total group.

W_L is the number of wrong answers in the lower one-fourth of the total group.
W_u is the number wrong in the upper one-fourth of the total group.
R_L is the number of right answers in the lower one-fourth of the total group.
N is the total number of cases.

4. Difference Between Means: $V = M_R - M_W$

M_R is the mean criterion score of those persons answering the item correctly.
M_W is the mean criterion score of those persons answering the item incorrectly.

5. Clark's Method: $V = \frac{P - q}{1 - q}$

P is the proportion of wrong answers among the W lowest scoring individuals of the entire group.
W is the number of wrong answers.
q is the per cent of wrong answers.

There are certain limitations affecting several of these methods. The method using Biserial r is recognized, in general, as being the best of the methods, but because of the great amount of time required in its computation, it is desirable to find some shorter method which will give results comparable to it. The same limitation also applies to Clark's Method.

The Upper Minus Lower Third Method is not wholly desirable as it limits the discriminating power of an item to the upper and lower one-third of the total group. No consideration is given to the variations among the middle

one-third of the group. Lentz and Handy (2) mentioned that the optimal upper and lower fraction is still unknown.

A similar criticism has been given in regard to Holzinger's Method. Swineford (5) stated that it does not consider the middle 50 per cent of the subjects, and, further, that it penalizes items whose difficulty lies outside the range (25 to 75 per cent).

The Difference Between Means Method seems to be inadequate since it is concerned only with the means of the right and wrong answers and does not consider the discriminating power to a fine enough degree.

By using arbitrary forms of distribution of the criterion variable, Toops (6) derived a simple formula from Pearson's Product-Moment formula for the correlation coefficient. When a rectangular criterion distribution is divided into five equal categories, the simplified formula may be written:

$$r_{XY} \quad \frac{\Sigma Y_R}{\sqrt{2RW}}$$

X is an item scored dichotomously and administered with no time limit.
Y is the criterion score.
R is the number of persons answering an item correctly.
W is the number of persons answering an item incorrectly.
ΣY_R is the sum of the criterion scores of the persons answering an item correctly.

If the five equal categories* are given a coded value of 2, 1, 0, -1, -2, in the descending order, and a, b, c, d, and e represent the number of right answers in each respective group, then ΣY_R becomes 2a + b - d - 2e. The formula now becomes:

$$r_{XY} = \frac{2a + b - d - 2e}{\sqrt{2N^2}}$$

and is then in a usable form.

MATERIALS

The materials used in this experiment were two multiple-response tests in plane geometry. Test I contained 46 items and covered the work of the first semester. Test II contained 44 items over the work of the second semester. The tests were prepared by the writer and were based on ten years of experience in teaching plane geometry. A great amount of time and care was spent in the preparation of the tests. Many items were rejected subjectively before the tests were given to the experimental group since the number of items was limited because of the time element. The writer was dependent upon fellow teachers in neighboring schools to secure sufficient subjects for the experiment and the time which they could give was limited.

---

*The five equal categories are obtained by dividing the ranked scores of the individuals tested into five equal groups in the order of excellence.

The tests were constructed to measure, primarily, the ability of the pupil to apply the facts and principles of plane geometry. The tests were meant to be answered easily and quickly by pupils who had a mastery of the fundamentals. Those pupils who did not have the fundamentals well mastered were expected to have considerable difficulty in answering the tests. Separate answer sheets were used in order to facilitate the scoring and tabulation.

It had been hoped that all schools contributing to the study would administer both tests. However, some schools felt that they could not give so much time. Consequently, the subjects of the experiment are not identical for both tests. For that reason the tests could not be combined into one test of 90 items but had to be treated as two separate tests of approximately 45 items each.

METHOD OF PROCEDURE

The purpose of this experiment was to study the effect of progressive validation on the test items. To accomplish that purpose Toops' formula for evaluating test items was used. An explanation of the procedure followed is necessary.

Toops' formula is based upon the correlation of a dichotomous variable with a multiple-categoried variable.

In order to secure the required data for substitution in the formula it was necessary to score each of the 300 answer sheets, giving each correct response a score of one and each incorrect response a score of zero. The subjects were then ranked into five equal groups on the basis of the magnitude of their total scores. The response to all items was then tabulated and the number of right answers to each item was recorded for every group. The total number of right answers and the total number of wrong answers for each item were determined from the data above, and Toops' formula was applied.

The above procedure was employed three times for each test: (a) it was used on the original tests; (b) it was used on the best 36 items of each test; (c) it was used on the best 30 items of each test.

In order to note the effect of each step in the validation on the reliability of the tests, the coefficients of reliability were found for the original tests, the two sets of the best 36 items, and the two sets of the best 30 items. The coefficients of reliability were obtained by correlating the odd items with the even items. A representative sampling of the entire experimental group was secured by selecting every third individual of the ranked order.

Tables 1 to 6 contain the data of the three steps in the validating process. An explanation of the notations used in all six tables appears on the back of page 13 opposite table 1.

Explanation of Tables

Table 1. Coefficient of validity of the original 46 items of Test I.

Item	a	b	c	d	e	R	W	RW	$\sqrt{RW}$	2a + b - d - 2e	r_1
1	59	60	56	56	53	286	14	8008	90	16	.18
2	56	54	47	44	31	232	68	31552	178	60	.34
3	55	51	41	44	29	218	82	35752	189	53	.29
4	60	59	50	49	42	260	40	20800	144	46	.32
5	56	55	50	45	38	244	56	27328	165	46	.28
6	28	22	17	16	11	94	206	38728	197	40	.20
7	59	52	52	39	22	224	76	34048	185	87	.47
8	60	56	53	42	31	242	58	28072	168	72	.43
9	60	58	55	55	40	268	32	17152	131	43	.33
10	54	43	38	35	38	208	92	36272	196	40	.20
11	57	55	53	50	48	263	37	19462	141	23	.16
12	60	56	55	51	21	241	59	28438	168	83	.49
13	59	51	50	49	26	235	65	30550	175	68	.39
14	59	56	57	56	49	277	23	12742	113	20	.18
15	56	45	41	30	12	184	116	42688	207	103	.50
16	57	55	45	44	31	232	68	31552	178	63	.35
17	60	60	59	55	42	276	24	13248	115	41	.36
18	59	58	55	48	40	258	42	21672	147	50	.34
19	43	30	34	21	21	149	151	44998	212	53	.25
20	59	55	52	50	42	258	42	21672	147	39	.27
21	50	33	26	20	15	144	156	44928	212	53	.39
22	55	49	45	38	28	215	85	36550	191	65	.34
23	46	31	18	22	20	137	163	44662	211	61	.29
24	54	47	39	24	12	176	124	43648	209	107	.51
25	36	35	35	38	32	176	124	43648	209	5	.02
26	60	59	57	50	39	265	35	18550	136	51	.38
27	56	52	49	29	26	212	88	37312	193	83	.43
28	55	53	46	35	29	218	82	35752	189	70	.37
29	52	54	45	32	23	206	94	38728	197	80	.41
30	55	47	38	31	16	187	113	42262	206	94	.46

(Continued on next page)

Table 1 (Cont.)

31	56	57	54	47	29	245	55	26950	164	68	.41
32	47	29	37	25	29	166	134	44488	211	42	.20
33	28	18	17	22	25	109	192	41472	204	6	.03
34	50	41	28	11	6	136	164	44608	211	116	.31
35	60	56	53	40	36	253	47	25702	154	56	.36
36	49	56	47	42	30	224	76	34048	185	92	.28
37	56	46	40	41	30	213	87	37062	193	57	.30
38	34	23	22	19	15	113	187	42262	206	42	.20
39	42	37	20	17	13	129	171	44116	210	78	.37
40	53	43	38	18	11	163	137	44662	211	109	.52
41	36	22	19	24	14	115	185	42550	206	42	.20
42	18	6	1	2	0	27	273	14742	121	40	.32
43	25	8	4	3	4	44	256	28528	150	47	.31
44	50	34	26	19	11	140	160	44800	212	93	.44
45	20	11	6	7	5	49	251	24598	157	34	.22
46	48	32	33	27	24	164	136	44606	211	55	.25

Table 2. Coefficient of validity of the original 44 items of Test II.

Item	a	b	c	d	e	R	W	2RW	$\sqrt{2RW}$	2a + b - d - 2e	r_1
1	57	59	56	56	50	278	22	12232	111	17	.15
2	60	59	58	56	49	282	18	10152	101	24	.24
3	57	57	49	47	37	247	53	26182	162	50	.31
4	60	57	41	41	26	225	75	33750	184	84	.46
5	56	56	52	44	37	245	55	26950	164	50	.30
6	59	60	59	55	50	281	19	10678	103	25	.24
7	57	50	52	49	25	233	67	31222	177	65	.37
8	57	60	58	52	46	273	27	14742	121	30	.25
9	56	50	47	42	37	232	68	30552	175	46	.26
10	57	45	40	47	47	245	55	26950	164	18	.11
11	57	51	48	57	25	216	84	36288	190	82	.43
12	57	50	44	43	49	243	57	27702	166	23	.14
13	48	34	23	19	5	129	171	44118	210	101	.48
14	55	54	55	52	44	258	42	21672	147	24	.16
15	58	57	53	53	41	262	38	19912	141	38	.26
16	60	56	60	56	47	279	21	11718	108	28	.24
17	57	49	51	41	22	220	80	35200	188	78	.41
18	60	56	51	56	31	236	64	30208	174	74	.43
19	51	42	34	20	17	164	136	44608	211	90	.43
20	58	54	51	42	26	253	67	31222	177	72	.41
21	44	27	15	23	17	126	174	43848	209	56	.28
22	43	21	21	18	17	120	180	43200	208	55	.26
23	56	52	37	34	19	198	102	40392	201	92	.46
24	44	16	15	7	3	85	215	36550	191	91	.48
25	58	53	50	37	24	220	80	35200	188	80	.43
26	58	54	53	49	36	250	50	25000	158	49	.31
27	51	39	27	25	11	153	147	44982	212	94	.44
28	54	51	49	30	33	217	83	36022	190	63	.33
29	50	45	38	37	27	197	103	40582	204	54	.22
30	58	56	45	42	28	229	71	32518	180	74	.41

(Continued on next page)

Table 2 (Cont.)

31	45	40	36	30	20	169	131	44278	210	56	.27
32	49	35	23	11	12	130	170	44200	210	98	.47
33	43	34	24	24	18	143	157	44902	212	60	.28
34	48	41	39	29	27	184	116	42688	211	54	.26
35	43	18	12	6	6	85	215	36850	191	86	.45
36	54	46	42	33	22	197	103	40582	201	77	.38
37	25	18	10	1	7	61	239	29158	171	53	.31
38	51	43	34	31	15	174	126	43848	209	64	.40
39	41	26	21	17	17	122	178	43432	208	57	.27
40	48	51	38	29	22	187	113	48262	206	75	.36
41	32	17	11	4	4	68	232	31552	178	69	.39
42	29	17	9	9	8	72	228	32832	181	50	.28
43	43	22	10	12	9	96	204	39168	198	78	.39
44	40	21	10	6	2	79	221	34918	187	91	.49

Table 3. Coefficient of validity of the best 36 items of Test I.

Item	a	b	c	d	e	$\sqrt{2nw}$	$2a+b-d-3e$	r_2
1								
2	54	56	47	44	31	178	58	.33
3	55	49	42	42	30	189	57	.30
4	60	60	48	49	45	144	43	.31
5	55	57	49	46	37	165	47	.28
6								
7	58	53	52	41	19	185	92	.50
8	60	56	53	42	31	168	72	.43
9	59	59	54	35	41	131	40	.31
10								
11								
12	59	58	54	49	21	168	65	.51
13	56	52	50	40	26	175	67	.38
14								
15	54	48	40	30	13	207	102	.49
16	56	57	47	40	32	178	65	.37
17	60	60	58	55	43	115	39	.34
18	60	57	55	47	39	147	52	.35
19	47	29	30	23	20	212	60	.28
20	59	55	53	49	42	147	40	.27
21	52	31	26	20	15	212	85	.40
22	54	52	41	41	27	191	65	.34
23	47	30	21	20	19	211	66	.31
24	54	47	40	21	14	209	106	.51
25								
26	60	60	56	51	38	136	53	.39
27	35	34	48	30	25	193	84	.44
28	57	52	45	35	29	189	73	.39
29	52	55	42	33	24	197	78	.40
30	53	48	38	31	17	206	89	.43

(Continued on next page)

Table 5 (Cont.)

31	58	55	55	43	29	164	65	.40
32								
33								
34	51	37	30	15	5	211	116	.55
35	60	56	53	50	34	154	58	.38
36	50	53	49	43	29	185	52	.28
37	55	49	42	39	28	193	64	.33
38								
39	44	35	22	14	14	210	81	.39
40	54	42	40	16	11	211	112	.53
41								
42	18	6	1	2	0	121	40	.33
43	28	5	4	2	5	150	49	.33
44	51	36	25	17	11	212	99	.42
45	22	10	6	4	5	157	38	.24
46	50	32	32	26	24	211	58	.27

Table 4. Coefficient of validity of the best 36 items of Test II.

Item	a	b	c	d	e	$\sqrt{2R'}$	$\frac{2a+b}{-d-2e}$	r_2
1								
2								
3	58	56	50	48	35	162	54	.33
4	60	54	48	39	24	184	87	.47
5	57	55	53	43	37	164	52	.32
6								
7	57	52	50	49	25	177	67	.38
8	57	59	59	51	47	121	28	.23
9	57	48	47	45	35	175	47	.27
10								
11	57	53	50	32	24	190	87	.46
12								
13	48	35	25	17	6	210	102	.49
14								
15	58	57	54	52	41	141	39	.28
16								
17	57	49	51	39	24	188	76	.40
18	59	58	49	38	32	174	74	.43
19	50	45	31	22	16	211	91	.43
20	59	52	53	41	28	177	73	.41
21	42	28	16	22	13	209	54	.26
22	43	24	19	16	18	208	58	.28
23	56	52	40	29	21	201	93	.46
24	46	14	16	6	5	191	94	.49
25	55	55	48	39	23	188	80	.43
26	58	54	52	50	36	158	48	.30
27	50	41	28	22	12	212	95	.45
28	54	50	48	33	32	190	61	.32
29								
30	58	56	43	43	29	180	71	.39

(Continued on next page)

Table 4 (Cont.)

31	44	38	35	31	21	210	53	.25
32	50	33	22	13	10	210	102	.49
33	44	36	25	24	17	212	63	.30
34	48	41	39	31	25	211	56	.27
35	42	21	10	8	4	191	89	.47
36	55	44	42	35	21	201	77	.38
37	24	17	11	2	7	171	49	.29
38	52	44	34	26	18	209	86	.41
39	41	25	19	21	16	208	54	.26
40	50	48	40	28	21	206	78	.38
41	33	16	11	4	4	178	70	.39
42	29	18	9	8	8	181	52	.29
43	40	26	8	13	9	198	75	.38
44	41	80	11	6	1	187	94	.50

Table 5. Coefficient of validity of the best 30 items of Test I.

Item	a	b	c	d	e	$\sqrt{2BW}$	$2a+b-d-2e$	r_3
1								
2	58	54	48	45	29	178	63	.35
3	55	49	43	41	30	189	58	.31
4	60	59	50	49	42	144	46	.32
5								
6								
7	59	53	51	39	22	185	88	.48
8	60	56	51	45	30	168	71	.42
9	60	59	53	56	40	131	43	.33
10								
11								
12	60	56	55	48	22	186	84	.50
13	59	52	48	43	28	175	66	.38
14								
15	56	46	45	26	13	207	106	.51
16	56	55	48	43	30	178	64	.36
17	60	60	59	55	42	115	41	.36
18	59	58	53	48	39	147	51	.35
19								
20								
21	50	34	26	22	12	212	88	.42
22	55	50	43	38	29	191	64	.34
23	48	29	19	23	18	211	66	.31
24	56	46	40	25	11	209	113	.54
25								
26	60	60	56	52	37	136	54	.39
27	56	53	49	26	28	193	83	.43
28	56	52	47	34	29	189	72	.38
29	54	52	45	32	25	197	82	.42
30	54	48	37	32	16	206	92	.45

(Continued on next page)

Table 5 (Cont.)

31	58	58	54	46	29	164	70	.43
32								
33								
34	52	37	31	10	6	211	119	.56
35	60	56	53	48	36	154	56	.36
36								
37	54	50	43	38	28	193	64	.33
38								
39	42	38	21	16	12	210	82	.39
40	53	44	37	18	11	211	110	.52
41								
42	18	7	0	1	1	121	40	.33
43	26	7	4	4	3	150	49	.33
44	52	34	25	19	10	218	99	.42
45								
46								

Table 6. Coefficient of validity of the best 30 items of Test II.

Item	a	b	c	d	e	$\sqrt{2RW}$	$\frac{2a+b}{-d-2e}$	r_3
1								
2								
3	56	56	50	48	35	162	54	.33
4	60	56	48	39	22	184	93	.50
5	57	55	54	45	34	164	56	.34
6								
7	56	54	48	49	26	177	65	.37
8								
9								
10								
11	57	54	48	34	23	190	86	.46
12								
13	49	54	25	19	4	210	105	.50
14								
15	58	57	54	53	40	141	40	.28
16								
17	57	49	50	45	19	188	80	.43
18	60	56	51	37	32	174	75	.43
19	51	42	53	22	16	211	90	.43
20	58	54	50	44	27	177	72	.41
21								
22	41	25	19	20	15	208	57	.27
23	57	51	39	33	18	201	96	.48
24	44	16	17	6	2	191	94	.49
25	55	55	45	41	24	188	76	.44
26	58	54	53	48	37	158	48	.30
27	52	41	26	23	11	212	100	.47
28	54	51	50	28	34	190	63	.33
29								
30	59	55	43	45	27	180	74	.41

(Continued on next page)

Table 6 (Cont.)

31								
32	49	35	23	12	11	210	99	.47
33	46	30	26	24	17	212	64	.30
34								
35	45	19	10	5	6	191	92	.48
36	54	46	43	30	24	201	76	.38
37	23	17	12	8	7	171	47	.27
38	51	44	35	28	16	209	86	.41
39								
40	50	48	40	26	25	206	76	.37
41	33	18	9	4	4	178	72	.40
42	30	16	11	5	10	181	51	.28
43	41	25	9	12	9	198	77	.39
44	44	17	10	6	2	187	95	.51

RESULTS

An examination of tables 1 to 6 will show that any change in the validity of a test item must be due to a change in the numerator of the formula; that is, in the value of $2a + b - d - 2e$. The total number of right and wrong answers to a given item could not change since the tests were given only once, hence the denominator of the formula remains constant. The individuals whose scores placed them in a certain criterion group during one validation were not necessarily in the same group during another validation. The shifting of individuals from one group to another because the weakest items are eliminated will cause a change in the numerator, and thus affect the coefficient in the various steps of validation.

Tables 7 and 8 show a comparison of the coefficients of validity obtained in each step of the progressive validation. Only two of the original 90 items were revealed as really poor items. For the most part the other items had coefficients which were relatively high and were fairly consistent. This study is primarily concerned with the changes that took place in the validity of the best items retained throughout the experiment.

When the changes in the validity of the best 30 items which form the final refined test are analyzed, it will be noted in table 7 that 17 items increased in validity, three remained the same, and ten decreased in the second validation. The greatest increase was .04, and the greatest decrease was .03. The average increase was .004*.

The first validation was on the basis of 46 items, and the second was on the basis of 36 items. The smallness of the number of items eliminated would preclude any great change.

In comparing the third and first validations it will be seen that 19 items increased in validity, seven remained the same, and only four decreased in validity. The average increase in the coefficient of validity was .01. The six additional items eliminated in the final validation bore out the observation that the number of items eliminated has a direct bearing on the amount of change that may take place.

A similar trend was seen in the coefficients for the items of Test II, table 8. In the second validation, 17 items increased in validity, seven remained the same, and

---

*It was recognized that the average of several correlation coefficients does not give a true measure of central tendency because the magnitude and significance change at different rates; however, the writer could not find a better method of presenting the facts.

Table 7. Progressive validity coefficients of the best 30 items of Test I.

Item	r_1	r_2	r_3
2	.34	.33	.35
3	.29	.30	.31
4	.32	.31	.32
7	.47	.50	.48
8	.43	.43	.42
9	.33	.31	.33
12	.49	.51	.50
13	.39	.38	.38
15	.50	.49	.51
16	.35	.37	.36
17	.36	.34	.36
18	.34	.35	.35
21	.39	.40	.42
22	.34	.34	.34
23	.29	.31	.31
24	.51	.51	.54
26	.38	.39	.39
27	.43	.44	.43
28	.37	.39	.38
29	.41	.40	.42
30	.46	.43	.45
31	.41	.40	.43
34	.51	.55	.56
35	.36	.38	.36
37	.30	.33	.33
39	.37	.39	.39
40	.52	.53	.52
42	.32	.33	.33
43	.31	.33	.33
44	.44	.42	.42
M	.391	.395	.401

Table 8. Progressive validity coefficients of the best 30 items of Test II.

Item	r_1	r_2	r_3
3	.31	.33	.33
4	.46	.47	.50
5	.30	.32	.34
7	.37	.38	.37
11	.43	.46	.46
13	.48	.49	.50
15	.26	.28	.28
17	.41	.40	.43
18	.43	.43	.43
19	.43	.43	.43
20	.41	.41	.41
22	.26	.28	.27
23	.46	.46	.48
24	.48	.49	.49
25	.43	.43	.44
26	.31	.30	.30
27	.44	.45	.47
28	.33	.32	.33
30	.41	.39	.41
32	.47	.49	.47
33	.28	.30	.30
35	.45	.47	.48
36	.38	.38	.38
37	.31	.29	.27
38	.40	.41	.41
40	.36	.38	.37
41	.39	.39	.40
42	.28	.29	.28
43	.39	.38	.39
44	.49	.50	.51
M	.387	.393	.398

six decreased. In the third validation, 18 items increased in validity, ten remained the same, and two decreased in validity. The average increase in the coefficients of validity for the two validations were .006 and .011 respectively.

The slight change that did occur throughout the validating process was in the direction of improvement. The writer was unable to discover any tables or any precedent by which he could determine the statistical significance of these slight differences. A procedure was used which will give at least a rough estimate of the significance of the difference noted. The three validity indices of the individual items of Test I and of Test II were correlated, that is, r_1 was correlated with r_2, r_2 with r_3, and r_1 with with r_3 for the two tests separately. The long formula for determining the probable error of the difference between the means from correlated material was then used, and the critical ratios were obtained. Tables 9 and 10 show the results of the correlations and the critical ratios found.

Lentz, Hirshstein and Finch (3) compared the reliability of a refined test with the reliability of the original test in order to study the effect that eliminating certain items had upon the reliability. A similar comparison of results as listed in tables 11 and 12 shows the

Table 9. Significance of the differences noted between the means of the three sets of validity coefficients of the best 30 items in Test I.

	r_1	r_2	r_3
r*	.97	.98	.98
M**	M_1 .391	M_2 .395	M_3 .401
$M_x - M_y$	.004	.006	.01
P.E.(Diff.)	.002	.00174	.0017
Critical ratio	2.0	3.4	5.9
Probability of a true difference	.91130	.98910	.99996

Table 10. Significance of the differences noted between the means of the three sets of validity coefficients of the best 30 items in Test II.

	r_1	r_2	r_3
r*	.98	.99	.98
M**	M_1 .387	M_2 .393	M_3 .398
$M_x - M_y$	.006	.005	.011
P.E.(Diff.)	.00173	.0014	.0017
Critical ratio	3.5	3.6	6.5
Probability of a true difference	.99090	.99240	.99999

*This correlation is included in this table because it enters into the calculation of the critical ratio by the long formula.

**M_1, M_2, and M_3 are all the means of the correlations of the retained items with the entire tests in successive stages of the refinements.

Table 11. Reliability of the original test and of the two sub-tests of Test I when extended to equal length.

Test	Reliability
Original 46 items	.797 ± .025
Best 36 items	.847 ± .019
Best 30 items	.898 ± .013

Table 12. Reliability of the original test and of the two sub-tests of Test II when extended to equal length.

Test	Reliability
Original 44 items	.880 ± .015
Best 36 items	.856 ± .018
Best 30 items	.864 ± .015

effect on the reliability of this experiment as the weakest items were omitted in successive steps. The Spearman-Brown formula (3) for predicting the reliability of lengthened tests was used, and all coefficients based on tests of equal length. The reliability of Test I shows a steady improvement as the weakest items were eliminated, but no improvement was noticed in the reliability of Test II. The first set of refined items of Test II had a slightly lower reliability than the original test. It may be noticed that the reliability of Test II was quite high and little improvement could be expected. Furthermore, the split-halves method of determining the coefficients of reliability of a test and of its sub-tests formed by eliminating certain items may not be a stable measure when there are so few items. For example, the original arrangement of the items gave the two sub-tests of Test II reliability coefficients of .856 and .884 respectively. When the items were rearranged, placing them in the order of difficulty, the coefficients became .927 and .879 respectively.

CONCLUSIONS

The results of this study show that there was a slight improvement in the validity of the test items when a system of progressive validation was used. For a short test which is to be used locally, the method would not be practical. If a test is to be prepared for wide usage and is composed of many more items than are expected to be retained finally, the results indicate that a larger increase in the validity of each retained item would be expected. In that case, progressive validation would be valuable, and a test maker should be expected to improve each item validity as much as possible.

ACKNOWLEDGMENTS

The writer wishes to express his gratitude to Dr. J. C. Peterson for his suggestions and counsel in the preparation of this thesis, to H. C. Moggie for his advice on the statistical methods employed, and to Dr. V. L. Strickland for his instructions on test construction.

For aid in securing an experimental group, the writer is indebted to the officers and mathematics teachers in the high schools of Robinson, Newton, Fairfield, Casey, Altamont, and Toledo, Illinois.

LITERATURE CITED

(1) Henry, L. J.
A comparison of the difficulty and validity of achievement test items. Jour. Educ. Psychol. 25: 537-541. Oct. 1934.

(2) Lentz, T. F. and Handy, Uran.
Item value and test reliability. Jour. Educ. Psychol. 25: 703-708. Dec. 1934.

(3) Lentz, T. F., Hirshstein, Bertha and Finch, J. H.
Evaluation of methods of evaluating test items. Jour. Educ. Psychol. 23: 344-350. May, 1932.

(4) Long, J. A. and Sandiford, P.
The validity of test items. Univ. of Toronto. Dept. of Educ. Res. Bul. 5, 126 p. 1935.

(5) Swineford, Frances.
Validity of test items. Jour. Educ. Psychol. 27: 68-78. Jan. 1936.

(6) Toops, Herbert A.
A note on item selection. Ohio Col. Assoc. Bul. 103: 2207-2212. Sept. 1936.

APPENDIX

Test I. Survey Test in Plane Geometry - First Semester

Rectilinear Figures and Circles

Instructions: After each of the following questions or problems several suggested answers are listed. Choose the best answer to each question; notice the letter before your chosen answer and draw a circle around that same letter on the line with the corresponding number on your answer sheet.

* 1. If two straight lines intersect, the vertical angles formed are: (a) complementary; (b) equal; (c) isosceles; (d) adjacent.

2. Two tangents drawn to a circle from an external point are: (a) parallel; (b) perpendicular; (c) unequal; (d) equal.

3. A line 4" long joins the midpoints of two sides of a triangle. The third side of the triangle is: (a) 2"; (b) 12"; (c) 8"; (d) 16".

4. A major arc of a circle is an arc that is: (a) less than a diameter; (b) equal to a semicircle; (c) less than a semicircle; (d) greater than a semicircle.

** 5. The perpendicular from the vertex of a triangle to the opposite side is called the: (a) altitude; (b) median; (c) perimeter; (d) diagonal.

* 6. Two chords intersect within a circle forming an angle which intercepts an arc of 84°. Its vertical angle intercepts an arc of 32°. The angle will contain: (a) 58°; (b) 26°; (c) 42°; (d) 96°.

7. Figures which have the same size and shape are said to be: (a) complementary; (b) congruent; (c) similar; (d) regular.

8. A diameter perpendicular to a chord will: (a) be twice the chord; (b) be equal to the chord; (c) be tangent to the chord; (d) bisect the chord.

9. One acute angle of a right triangle is 41°, the other acute angle will be: (a) 39°; (b) 45°; (c) 49°; (d) 139°.

*10. A quadrilateral is a parallelogram if: (a) one pair of opposite sides are equal; (b) two consecutive angles are supplementary; (c) the opposite sides are equal; (d) one angle is a right angle.

*11. Supplements of the same angle or equal angles are: (a) equal; (b) right angles; (c) straight angles; (d) unequal.

12. If one of the equal angles of an isosceles triangle is 68°, the third angle will be: (a) 68°; (b) 22°; (c) 44°; (d) 112°.

13. Alternate interior angles of parallel lines are: (a) supplementary; (b) equal; (c) adjacent; (d) right angles.

*14. The length of a circle is called the: (a) radius; (b) arc; (c) circumference; (d) sector.

15. If an angle formed by a tangent and a chord intercepts an arc of 76° on a circle, it will contain: (a) 38°; (b) 76°; (c) 142°; (d) 36°.

16. An angle formed by two chords meeting on the circle is: (a) an acute angle; (b) a central angle; (c) an inscribed angle; (d) a right angle.

17. The hypotenuse of a right triangle is the: (a) side opposite the right angle; (b) the shortest side; (c) equal to the shortest side multiplied by two; (d) equal to the sum of the other two sides.

18. Two angles are supplementary if their sum is: (a) 100°; (b) 90°; (c) 180°; (d) 360°.

**19. A general statement which is accepted as true without proof is: (a) a postulate; (b) a theorem; (c) the hypothesis; (d) an axiom.

**20. A polygon of six sides is called: (a) a quadrilateral; (b))a pentagon; (c) a decagon; (d) a hexagon.

21. The exterior angle at the vertex of a triangle is 114° and one base angle is double the other. The smaller of the two base angles will be: (a) 38°; (b) 33°; (c) 57°; (d) 76°.

22. The locus of points equidistant from a given point is: (a) a circle; (b) a parallel line; (c) the bisector of an angle; (d) the perpendicular bisector.

23. Geometry was probably first compiled in a written form by: (a) Aristotle; (b) Plato; (c) Pythagoras; (d) Euclid.

24. If an exterior angle at the base of an isosceles triangle is 110°, the vertex angle of the triangle will be: (a) 70°; (b) 40°; (c) 110°; (d) 140°.

*25. The part of a theorem which states what is to be proved is called the: (a) hypothesis; (b) demonstration; (c) proposition; (d) conclusion.

26. An obtuse triangle is a triangle in which: (a) all angles are obtuse; (b) one angle is obtuse; (c) all angles are unequal; (d) all angles are acute.

27. The hypotenuse of a 30°- 60° right triangle is 13" long. The shorter side of the triangle will be: (a) 6"; (b) 12"; (c) 9"; (d) 36".

28. Two chords equidistant from the center of a circle are: (a) parallel; (b) drawn through a common point; (c) perpendicular; (d) equal.

29. A straight line intersecting a circle in two points is called a: (a) tangent; (b) chord; (c) secant; (d) radius.

30. If one angle of a parallelogram is 80°, its consecutive angle is: (a) 10°; (b) 20°; (c) 100°; (d) 80°.

31. A regular polygon is a polygon with (a) all sides equal and all angles equal; (b) all sides equal; (c) all angles equal; (d) an even number of sides.

*32. A segment of a circle is the figure formed by: (a) two radii and an arc; (b) two radii and a chord; (c) two chords and an arc; (d) an arc and a chord.

*33. An angle formed by two tangents drawn to a circle from the same point contains 76°. The minor arc intercepted will have: (a) 128°; (b) 152°; (c) 38°; (d) 104°.

34. The bisectors of the angles of a triangle are: (a) concurrent; (b) equal; (c) proportional; (d) perpendicular to the opposite sides.

35. Two circles which have the same center are said to be: (a) tangent internally; (b) tangent externally; (c) concentric; (d) equal.

**36. The sum of the exterior angles of a regular polygon is: (a) 360°; (b) 180°; (c) 540°; (d) 90°.

37. Doubling an arc will also double the: (a) central angle; (b) chord; (c) radius; (d) diameter.

*38. If a point is equidistant from the sides of a triangle, it will be in the intersection of the: (a) altitudes; (b) medians; (c) angle bisectors; (d) perpendicular bisectors of the sides.

39. The sum of the interior angles of a polygon of seven sides is: (a) 900°; (b) 1080°; (c) 1260°; (d) 720°.

40. An altitude of an acute triangle will divide the triangle into two triangles which are: (a) congruent; (b) right; (c) similar; (d) obtuse.

*41. The center of a circle circumscribed about a triangle is the point of intersection of the three:

(a) altitudes; (b) medians; (c) bisectors of the angles; (d) perpendicular bisectors of the sides.

42. An angle inscribed in an arc of 128° contains: (a) 128°; (b) 116°; (c) 64°; (d) 52°.

43. One angle of a triangle is 40°. The bisectors of the other two angles will form an angle of: (a) 140°; (b) 110°; (c) 70°; (d) cannot be determined.

44. If in quadrilateral ABCD AD is the longest and BC is the shortest side, then: (a) angle B equals angle D; (b) angle B is less than angle D; (c) angle B is greater than angle D; (d) angle B and angle D are right angles.

**45. An angle inscribed in a circle and an angle formed by two tangents intercept the same arc. The inscribed angle is 50°. The angle formed by the tangents is: (a) 100°; (b) 80°; (c) 50°; (d) 130°.

**46. The equilateral quadrilateral that is not regular is a: (a) square; (b) trapezoid; (c) rectangle; (d) rhombus.

*Items omitted in the second step of the validation.

**Items omitted in the third step of the validation.

Test II. Survey Test in Plane Geometry - Second Semester

Proportion, Similar Triangles, and Areas

Instructions: After each of the following questions or problems several suggested answers are listed. Choose the best answer to each question; notice the letter before your chosen answer, and draw a circle around that same letter on the line with the corresponding number of the answer sheet.

* 1. The first and last terms of a proportion are the: (a) antecedents; (b) means; (c) extremes; (d) consequents.

* 2. The base of a triangle is B and the altitude is H. The area may be represented by the formula: (a) BH; (b) $\frac{BH}{2}$; (c) 2BH; (d) $BH\sqrt{3}$.

3. If the legs of a right triangle are 6" and 8" respectively, the hypotenuse is: (a) 10"; (b) 48"; (c) 14"; (d) 24".

4. Two polygons having equal areas are said to be: (a) similar; (b) congruent; (c) convex; (d) equivalent.

5. The perimeters of two similar triangles are 18" and 15" respectively. If a side of the first is 12", the corresponding side of the second will be: (a) 10"; (b) 6"; (c) 6"; (d) 12".

* 6. If the sides of two triangles are respectively proportional, the triangles are: (a) scalene; (b) congruent; (c) similar; (d) isosceles.

7. The area of a trapezoid whose bases are 8" and 10" and whose altitude is 6" will be: (a) 48; (b) 60; (c) 54; (d) 108 square inches.

** 8. The fourth proportional to 2, 7, and 4 is: (a) 14; (b) 8; (c) 1 1/7; (d) 3 1/2.

** 9. If two polygons are similar, their corresponding angles are: (a) acute; (b) proportional; (c) equal; (d) obtuse.

*10. The area of a trapezoid equals the product of one-half its altitude and: (a) the perimeter; (b) one-half the sum of the bases; (c) the sum of the bases; (d) the difference of the bases.

11. The sides of a triangle are 3, 4, and 5. The shortest side of a similar triangle is 9. The longest side of the second triangle is: (a) 15; (b) 11; (c) 12; (d) 25.

*12. Two triangles are similar if they are: (a) right triangles; (b) mutually equiangular; (c) equal in area; (d) oblique.

13. Two chords intersect within a circle. The segments of one are 3" and 4". If one of the segments of the other chord is 6", the other segment will be: (a) 1"; (b) 2"; (c) 8"; (d) 7".

*14. The area of a parallelogram whose base is 18" and whose altitude is 8" is: (a) 72; (b) 26; (c) 52; (d) 144 square inches.

15. If a line is parallel to one side of a triangle, it will: (a) divide the other two sides proportionally; (b) bisect the other two sides; (c) be equal to one-half the parallel side; (d) bisect the triangle.

*16. The base of a triangle is 16" and the altitude to the base is 10". The area of the triangle is: (a) 160; (b) 40; (c) 80; (d) 320 square inches.

17. Triangles having equal bases and equal altitudes are always: (a) equilateral; (b) equivalent; (c) isosceles; (d) scalene.

18. The mean proportional between 4 and 9 is: (a) 36; (b) 20 1/4; (c) 1 7/9; (d) 6.

19. If a side of an equilateral triangle is 8", its area will be: (a) 32; (b) $16\sqrt{3}$; (c) $9\sqrt{3}$; (d) $32\sqrt{3}$ square inches.

20. The square of the hypotenuse of a right triangle is equal to: (a) the square of the sum of the two legs; (b) the square of the difference of the two legs; (c) the difference of the squares of the two legs; (d) the sum of the squares of the two legs.

**21. The sides of a triangle are 13, 12, and 5. The angle opposite the side 13 is: (a) right; (b) acute; (c) obtuse; (d) cannot be determined.

22. The sides of a triangle are 9, 18, and 21. The shorter of the two segments of the side 21 made by the bisector of the opposite angle is: (a) 10 1/2; (b) 14; (c) 6; (d) 7.

23. The altitude to the hypotenuse of a right triangle divides it into two triangles which are: (a) equal; (b) congruent; (c) 30°-60° right triangles; (d) similar.

24. The corresponding bases of two similar triangles are 2" and 3" respectively. If the area of the first is 8 square inches, the area of the second is: (a) 18; (b) 16; (c) 9; (d) 12 square inches.

25. The area of a trapezoid is 36 square inches, and the parallel sides are 10" and 8". The altitude is: (a) 2"; (b) 8"; (c) 4"; (d) 3.6".

26. If the vertex angle of an isosceles triangle is 50°, one of the equal base angles of a similar triangle will be: (a) 50; (b) 75; (c) 40; (d) 65 degrees.

27. The areas of two similar triangles have the same ratio as: (a) two corresponding sides; (b) two corresponding altitudes; (c) their perimeters; (d) the squares of two corresponding sides.

28. A tower on level ground casts a shadow 150 feet long while a pole 12 feet high casts a shadow 10 feet long. The height of the tower is: (a) 100'; (b) 125'; (c) 200'; (d) 180'.

*29. If the antecedents of a proportion are equal, then the: (a) means are equal; (b) consequents are equal; (c) extremes are equal; (d) the proportion is true.

30. The area of a triangle is 36 square inches, and its base is 9". Its altitude is: (a) 2"; (b) 4"; (c) 8"; (d) 16".

**31. In triangle ABC D and E are points in AB and AC respectively, and DE is parallel to BC. AD is 3", AE is 5", AB is 12", and AC is: (a) 20"; (b) 15"; (c) 10"; (d) 7 1/5".

32. A median of a triangle will divide the triangle into two triangles which are: (a) right; (b) similar; (c) equivalent; (d) congruent.

33. If the altitude of an equilateral triangle is 12", the side is: (a) $8\sqrt{3}$ "; (b) 6"; (c) 12"; (d) 24".

**34. In a right triangle the ratio of the side opposite an acute angle to the hypotenuse is called the: (a) tangent; (b) cotangent; (c) sine; (d) cosine.

35. A diagonal of a rectangle is 10" long and makes an angle of 30° with the base. The area of the rectangle is: (a) $25\sqrt{3}$; (b) 25; (c) $50\sqrt{3}$; (d) 50 square inches.

36. If the square of one side of a triangle is greater than the sum of the squares of the other two sides, the angle opposite it is: (a) cannot be determined; (b) right; (c) acute; (d) obtuse.

37. If the sides of a right triangle are 17, 15, and 8, the tangent of the angle opposite the side 8 is: (a) .533; (b) .882; (c) .471; (d) 1.875.

38. Two isosceles triangles with equal vertex angles are: (a) congruent; (b) equivalent; (c) equilateral; (d) similar.

**39. If a side of a square is 10', the diagonal is: (a) $10\sqrt{2}$; (b) $5\sqrt{2}$; (c) $10\sqrt{3}$; (d) 15 feet.

40. The bisector of an angle of a triangle divides the opposite side into segments proportional to: (a) the other two sides; (b) the longest side; (c) each other; (d) the base and the altitude.

41. From a point 4" from a circle a tangent to the circle is 10" long. The diameter of the circle is: (a) 25"; (b) 15"; (c) 21"; (d) 14".

42. The sides of a triangle are 3, 25, and 26. The area will be: (a) $12\sqrt{3}$; (b) 75; (c) 48; (d) 36.

43. The product of the legs of a right triangle is equal to the product of: (a) the square of the hypotenuse; (b) the segments of the hypotenuse made by the altitude to the hypotenuse; (c) the hypotenuse and the segment adjacent to one leg; (d) the hypotenuse and the altitude to the hypotenuse.

44. The area of an equilateral triangle is $25\sqrt{3}$. The length of a side is: (a) $5\sqrt{3}$; (b) 10; (c) 12 1/2; (d) 5.

*Items omitted in the second step of the validation.

**Items omitted in the third step of the validation.