First Annual Report
of the
Commission on Scholastic Aptitude Tests
1926

Number of Candidates Examined

The scholastic aptitude test was taken on June 23, 1926, by 8,040 candidates, 4,829 boys and 3,211 girls. The centers at which the tests were given, and the number of candidates examined at each center are listed in Table III on pages 163-170. If the number of candidates to be examined at any center exceeded the seating capacity of the largest room available, the candidates were examined by different supervisors in separate rooms, designated in Table III as sub-centers. The tests were given in 353 centers and sub-centers in 318 examination centers.

In order that all candidates might have the opportunity of familiarizing themselves with the nature of the tests to be given, practice booklets containing samples of all the tests were sent out one week before the examination. These practice booklets constituted tickets of admission to the test. Eleven candidates, eight boys and three girls, were admitted to the test without practice booklets. Since these candidates did not take the test under conditions comparable with those of the other candidates, no scores were reported for them. It should be remembered that this action was made necessary because the scores obtained by candidates who had not studied the practice booklet could not be compared with the scores of candidates who had had ample opportunity to practice on the material of the examination.

The test of one candidate, a girl, was invalidated for the reason that she worked on the wrong sub-tests during some of the allotted periods of time, and her error was not detected by the supervisor in time to obtain a reliable test record.

The elimination of this faulty test record, and the eleven records invalidated for lack of practice, left 8,028 test records which were reported to the colleges. These records were classified by colleges, duplications caused by the designation of more than one college being avoided by taking the college named first.

The colleges for which the 4,821 boys were examined, arranged in the order of the number of candidates taking the test, are as follows:

TABLE I*

Yale University	1257
University of Pennsylvania	1176
Princeton University	918
Harvard University	536
Massachusetts Institute of Technology	396
Columbia University	171
Williams College	121
Brown University	55
Dartmouth College	26
Trinity College (Hartford, Conn.)	21
Tufts College	17
Amherst College	16
Stanford University	13

Ten candidates were examined for Hamilton College; seven each for Wesleyan University and for the Springfield Junior College; six each for Bowdoin College and the University of Michigan; five each for Cornell University, Lehigh University, New York University, and Swarthmore College; three candidates for Lafayette College; two candidates each for the University of Cincinnati, Haverford College, and Middlebury College; and one candidate each for Bates College, Carlton College, the College of the City of New York, Holy Cross College, Johns Hopkins University, Northwestern University, Pennsylvania State College, Rensselaer Polytechnic Institute, Rutgers University, Stevens Institute of Technology, Union College, the United States Naval Academy, the University of Indiana, the University of Virginia, the University of Wisconsin, and Worcester Polytechnic Institute. In addition to the boys planning to enter the colleges listed above, six candidates who had graduated from secondary schools and were eligible for the test were planning to take an additional year in a preparatory school before entering college. Two of these boys were planning to enter Phillips Academy, two were entering Phillips Exeter Academy, one was entering Lawrenceville School, and one was entering the Chauncy Hall School. Reports were withheld in these six cases. Eleven candidates had designated no college at the time the examination was given.

The colleges for which the 3,207 girls were examined, arranged in the order of the number of candidates taking the test, were as follows:

TABLE II*

Smith College	865
Wellesley College	742
Vassar College	602
Mount Holyoke College	373
Radcliffe College	233
University of Pennsylvania	123
Barnard College	88
Wells College	53
Connecticut College for Women	40
Bryn Mawr College	22
Albertus Magnus College	22

Eight candidates were examined for Brown University, six for Goucher College, three candidates each for Simmons College, Trinity College (Washington, D. C.), Springfield Junior College, and Wheaton College; two candidates each for Stanford University, Swarthmore College, and the University of Minnesota; and one candidate each for Mills College, Northwestern University, the New Jersey College for Women, Rice Institute, Syracuse University, and Tufts College. One girl intended to enter the Martha Washington Seminary. Five girls were undecided as to their college at the time of taking the examination.

TABLE IV

NUMBER OF MEN WORKING AND TOTAL NUMBER OF HOURS SPENT IN SCORING EACH DAY DURING WHICH THE SCORING UNIT OPERATED

DATE	NUMBER OF MEN	TOTAL HOURS
June 23	5	37.5
June 24	34	320.3
June 25	37	355.4
June 26	44	408.1
June 27	37	342.7
June 28	45	429.0
June 29	44	409.6
June 30	42	380.8
July 1	40	376.9
July 2	36	342.3
July 3	30	264.4
July 4	24	242.7
July 5	28	278.8
July 6	22	209.9
July 7	17	151.0
July 8	12	93.6
July 9	11	86.0
July 10	10	75.6
July 11	00	00.0
July 12	10	82.9
July 13	10	81.9
July 14	9	67.0
July 15	8	68.5
July 16	2	15.0

* This table was constructed on the basis of the colleges named by the candidates on their test booklets and does not agree with Table I on pages 8 and 9 of the Annual Report of the Secretary, which was compiled from the application blanks filed three weeks or more in advance of the examinations.

General Arrangements for Scoring the Tests

The test was given at 9:00 A. M. on Wednesday, June 23, and over fifteen hundred booklets were delivered to the scoring unit in New York City on that day. The scoring of the test papers was begun on Thursday, June 24.

The clerks for scoring the test were recruited from the ranks of the undergraduates at Princeton University and Columbia University, and worked under the supervision of a staff composed of instructors and graduate students in psychology in various institutions. The clerks drawn from the Princeton undergraduate body were selected by first obtaining a roster from the college employment office showing all students registered for self-help who lived in or near New York City, and by inviting those men to work whose Princeton intelligence test score was one sigma above the Princeton average. Practically all men finally employed met this condition. Clerks obtained through the Columbia Appointments Bureau had Thorndike scores over 90.

The total number of individuals employed during each day of scoring and the total number of hours spent each day are shown in Table IV.

The physical conditions for scoring were ideal. The clerks used large drawing-rooms which were adequately lighted and ventilated. Other rooms were available for receiving records, checking, sorting, and other work, and were well adapted to the various purposes. The Sundstrand Adding Machine Company loaned the scoring unit five adding machines. Two Monroe Calculators were available.

The bulk of the records had been received by Saturday, June 26. About fifteen hundred papers were received on Wednesday, June 23, about fourteen hundred on Thursday, and about twenty-six hundred on Friday. By Saturday, June 26, over seven thousand booklets had been scored. The reports had been mailed to the colleges for all examination centers except El Paso, Paris, Honolulu, Shanghai, and Peking by the evening of Tuesday, July 6.

Details of Checking and Scoring

The operations involved in the work of checking and scoring fell under five headings:

Receiving
Marking
Tallying test scores for the first 1500 cases
Converting
Reporting

In the receiving room, each package was checked against the supervisor's report for the following purposes:

(1) to see whether each candidate had submitted a practice booklet;
(2) to see whether all booklets either (*a*) had been used by the candidates or (*b*) were returned with seals unbroken;

and the supervisors' reports were examined for reported mistakes in timing the sub-tests.

Irregularities discovered were immediately reported to the Secretary of the College Entrance Examination Board, who communicated with the supervisors by telegraph. Eleven candidates were admitted without practice booklets. The practice booklets for one group of twenty-eight candidates were accidentally destroyed. The seals of seven booklets had been broken by supervisors who picked up regular test booklets instead of the dummies provided to illustrate the method of breaking the seals. The seals of five booklets were broken inadvertently. The seals of eight booklets were broken for the purpose of inspecting the tests. Two booklets were accidentally destroyed. Ten booklets were not returned and were either retained or given to other persons to study. Seven of these booklets have now been recovered and the other three booklets will be recovered. One booklet was missing and no check as to its location was possible.

The wishes of the committee in insisting on preserving the secrecy of the examination booklets were respected by most persons, as shown by the very small number of irregular cases recorded above. A more adequate presentation of the committee's reasons might have obviated the unpleasantness occasioned by these few flagrant violations of the rules. The cost of constructing diagnostic tests of this type quickly runs into thousands of dollars. If an examination is once generally released it cannot be used again. The committee plans to use these same examinations again either for purposes of examining candidates who did not take this particular form or for re-testing purposes. Consequently, the same rules for secrecy obtained in the case of this test after it had been taken as would obtain with any unpublished examination paper. After the usefulness of this particular examination has been exhausted, it will be released.

There were four errors in timing which affected the scores in one test, and one error in timing which affected the scores in two tests. Two other errors in timing were corrected and did not affect the scores in the tests. Since an error in timing would alter the results in only one or two tests, no complete test records were invalidated on this account. The timing errors reported amounted to about four-tenths of one per cent.

The actual work of marking the answers of the candidates proceeded rapidly. The average scoring time for practiced markers was six minutes for each booklet. The tests were sent from the receiving room in packets of fifty, the covering folder providing places for recording the names of the scorers, the time, and other information. No clerk worked on more than one subtest. Five booklets from each packet of fifty were completely re-scored, and errors in scoring thus detected. If two errors were discovered on any test, the entire packet was re-scored.

With regard to the accuracy of the marking, it is true that there were errors but it is thought that the system employed prevented the possibility of any large errors.

The average scoring error per subject was estimated from computations as follows for each of the nine sub-tests:

TABLE V

SUB-TEST	
1. Definitions	.08
2. Arithmetical Problems	.03
3. Classification	.12
4. Artificial Language	.40
5. Antonyms	.16
6. Number Series Completion	.07
7. Analogies	.08
8. Logical Inference	.05
9. Paragraph Reading	.14

Differences of construction in the tests and in the scoring stencils made differences in the average error of scoring. The average error of scoring per *test* for a sample of the first 3000 cases was .14, while the average error per *candidate* was 1.14. These figures are in terms of raw score and signify errors smaller than two points, on the average, in the final scale score reported to the colleges. The accuracy of scoring may be improved by better designing of tests, or by more complete re-scoring if this is considered necessary.

The particular manner in which the tests are constructed dictates the amount of time required in scoring. Five experienced scorers marked each test of a packet of fifty booklets in an average time of five hours, or six minutes to a booklet. The proportion of time spent on the nine sub-tests was as follows:

TABLE VI

SUB-TEST	
1. Definitions	.10
2. Arithmetical Problems	.06
3. Classification	.09
4. Artificial Language	.16
5. Antonyms	.17
6. Number Series Completion	.08
7. Analogies	.08
8. Logical Inference	.10
9. Paragraph Reading	.16

Since the tests are standardized on the basis of the particular group examined, there are many statistical operations involved

which are peculiar to this type of work. The first big task was to construct a standard scale for converting the scores in the nine separate sub-tests into a score in which the average was 50 and the value of the standard deviation was 10. The separate scores in the nine sub-tests and a total score were reported to the colleges. The total score was based on the sum of the nine converted scores and served to weight each sub-test at unity, in accordance with conservative practice.

Since it was impossible to wait until all eight thousand papers had been scored before constructing the conversion key, tallies of scores in single tests were made for the first 500 cases, the first 1000 cases, and the first 1500 cases. This operation of adding 500 cases would have been continued until the means and sigmas had become stable, had not the results of the first 1500 cases showed that these were sufficiently stable to give a sound basis for the conversion key.

The data showing the stability of means and sigmas for the first 500, first 1000, and first 1500 cases are shown in Table VII. Since the means and sigmas fluctuated little as additional cases were added, the conversion key was built on the basis of the means and sigmas given in the third column of Table VII. If the conversion key as standardized for this group is valid for the total group, the average of the sum of all converted tests for the total group of applicants should equal 50. This average, subsequently computed for 7992 candidates, was 50.1143.

TABLE VIII

CONVERSION KEY USED IN CONVERTING RAW SCORES IN NINE SUB-TESTS TO A STANDARD SCALE IN WHICH THE MEAN WAS 50 AND SIGMA 10

RAW SCORE	1	2	3	4	5	6	7	8	9
0	12	22	18	17	8	18	8	10	12
1	15	26	21	18	9	21	10	12	13
2	17	30	23	20	10	24	11	14	15
3	19	34	25	22	12	27	13	15	16
4	22	37	28	23	13	30	15	17	17
5	24	41	30	25	14	33	17	19	19
6	26	45	32	27	16	36	18	20	20
7	29	48	35	28	17	39	20	22	22
8	31	52	37	30	18	43	22	23	23
9	33	56	40	32	20	46	24	25	25
10	36	60	42	33	21	49	26	27	26
11	38	63	44	35	23	52	27	28	28
12	40	67	47	37	24	55	29	30	29
13	43	71	49	38	25	58	31	32	31
14	45	75	52	40	27	61	33	33	32
15	47	78	54	42	28	64	34	35	33
16	49	82	56	43	29	68	36	36	35
17	52	86	59	45	31	71	38	38	36
18	54	89	61	47	32	74	40	40	38
19	56	93	64	48	33	77	42	41	39
20	59	97	66	50	35	80	43	43	41
21	61		68	52	36	83	45	45	42
22	63		71	53	38	86	47	46	44
23	66		73	55	39	89	49	48	45
24	68		76	57	40	92	50	49	47
25	70		78	58	42	95	52	51	48
26	73		80	60	43		54	53	49
27	75		83	62	44		56	54	51
28	77		85	63	46		58	56	52
29	80		88	65	47		59	58	54
30	82		90	67	48		61	59	55
31			92	68	50		63	61	57
32			95	70	51		65	62	58
33			97	72	53		67	64	60
34			100	74	54		68	66	61
35			102	75	55		70	67	63
36			104	77	57		72	69	64
37			107	78	58		74	71	65
38			109		59		75	72	67
39			112		61		77	74	68
40			114		62		79	75	70
41					63				71
42					65				73
43					66				74
44					68				76
45					69				77
46					70				79
47					72				80
48					73				81
49									83
50									84

There were 74 scorable points in sub-test 4 and the score in this test was divided by 2, giving a maximum possible score of 37 points.

TABLE VII

SUB-TEST		FIRST 500 N=502	FIRST 1000 N=999	FIRST 1500 N=1499
1. Definitions	Mean	16.02	15.96	16.24
	Sigma	4.33	4.36	4.33
2. Arithmetical Problems	Mean	7.37	7.44	7.42
	Sigma	2.74	2.73	2.68
3. Classification	Mean	13.32	13.18	13.32
	Sigma	4.19	4.13	4.17
4. Artificial Language	Mean	20.52	19.78	19.91
	Sigma	6.17	5.93	6.00
5. Antonyms	Mean	32.04	30.88	31.11
	Sigma	7.15	7.29	7.33
6. Number Series Completion	Mean	10.48	10.48	10.42
	Sigma	3.23	3.24	3.24
7. Analogies	Mean	23.64	23.48	23.73
	Sigma	5.60	5.49	5.61
8. Logical Inference	Mean	24.69	24.31	24.34
	Sigma	6.09	6.06	6.16
9. Paragraph Reading	Mean	26.96	26.13	26.39
	Sigma	6.44	6.74	6.88

The 1499 cases used in this computation included 921 boys and 578 girls. Average scores and standard deviations were computed for boys and girls separately and will be reported later. The same conversion key was used for boys and girls. This key is shown in Table VIII.

The distribution of raw scores in the nine sub-tests for the 1499 cases is shown in Table IX. On account of a timing error, sub-test 2 shows 1359 cases instead of 1499.

TABLE IX

DISTRIBUTION OF RAW SCORES IN THE NINE SUB-TESTS FOR 1499 CASES (1359 IN SUB-TEST 2)

RAW SCORE	1	2	3	4	5	6	7	8	9
0	2	1	1			1			
1		8	1			3	1		1
2		16	2			6			
3	1	55	3			7			2
4	1	109	9	1		23			1
5	7	132	25	1		40			
6	7	206	32			75		1	1
7	12	194	42	4		104	3	2	1
8	27	183	78	13	1	157		5	4
9	41	178	80	16	4	181		5	8
10	34	116	119	24	1	193	8	11	7
11	69	70	128	39	2	186	4	17	10
12	87	43	125	45	1	146	7	17	7
13	122	18	132	55	8	153	10	19	18
14	112	18	143	67	7	70	25	28	17
15	118	7	128	89	7	70	38	28	17
16	128		106	99	16	22	62	33	30
17	146	2	108	90	15	22	56	42	34
18	124	3	74	98	18	14	66	50	41
19	121		56	105	20	20	84	50	44
20	99		41	102	44	3	68	67	65
21	70		26	126	30	1	96	86	42
22	57		14	128	40	1	108	73	49
23	47		14	69	38		96	73	75
24	30		7	45	42	1	109	97	81
25	19		1	29	50		89	98	87
26	9		4	36	45		100	107	85
27	3			42	58		96	100	72
28	4			33	61		73	93	99
29	2			23	63		58	81	96
30				14	76		56	69	82
31				20	78		46	51	82
32				32	92		41	55	45
33				12	71		33	45	68
34				12	85		26	36	55
35				14	76		14	20	36
36				7	78		9	25	40
37				9	66		10	9	42
38					63		7	6	18
39					56				14
40					52				5
41					41				10
42					27				3
43					20				3
44					23				
45					15				1
46					5				
47					1				
48					3				
49									1
50									

As soon as the conversion key shown in Table VIII had been constructed, the scores of all booklets which had been marked were converted in accordance with this key, and the total score in the nine tests obtained. This operation was carried through for every booklet. The adding-machine strips were saved, so that single test distributions for the entire group may be obtained at a later date, if required.

The sum of nine tests on the standard scale in which the mean is 50 and sigma 10 will give a total score for the whole group in which the mean is 450 and sigma less than 90, the restriction of sigma being due to the inter-correlation of the nine tests. Consequently, it was necessary to tally all total scores to furnish a basis for converting these scores into the final scale score in which the mean was 500 and sigma 100. This procedure automatically corrected any error in the first conversion key. Tallies of total scores were accumulated by thousands with the following results:

TABLE X

	ACTUAL NUMBER	MEAN	SIGMA
First 1000.....	999	447.4224	63.48
First 2000.....	1949	451.7320	63.53
First 3000.....	2927	452.8203	63.40
First 4000.....	3928	450.8707	63.26
First 5000.....	4928	450.8730	62.72
First 6000	5928	450.1586	63.16
First 7000.....	6928	450.9790	62.73
First 8000.....	7992	451.0290	62.94

It was, of course, impossible to wait until 8000 books had been scored before beginning the work of writing reports to the colleges. The key for converting the sum of the sigma scores in the nine tests to the final scale score was based on the first 6000 cases.

Tallies of final scale scores were made as a final check. If the first 6000 cases were typical of the whole group, the average of all scale scores should have been 500, and the value of sigma should have been 100. The actual calculated value of the mean of 7990 cases was 500.8 and the calculated value of sigma was 99.8, which constitutes a sufficiently close check of the method.

The means and sigmas of the group of 921 boys and 578 girls in the nine sub-tests are given in Table XI. This table also shows the sigma of the means and significance of the difference as obtained by the ratio between the difference and the probable error of the difference. If this ratio is greater than 4.00, the difference is conventionally accepted as significant. Significant differences are shown in all tests except #1, Definitions, and #7, Analogies. Boys are significantly better in #2, Arithmetical Problems, and #6, Number Series Completion. Girls are significantly better in the other five tests, the differences being very great in the Artificial Language Test, Antonyms, and Paragraph Reading.

TABLE XI

MEANS AND SIGMAS OF 921 BOYS AND 578 GIRLS IN THE NINE SUB-TESTS

SUB-TEST		MEANS AND SIGMAS: BOYS	MEANS AND SIGMAS: GIRLS	RELIABILITY (SIGMAS OF THE MEAN): BOYS	RELIABILITY (SIGMAS OF THE MEAN): GIRLS	SIGNIFICANCE OF DIFFERENCE: DIFFERENCE / P. E. OF DIFFERENCE
1. Definitions	Mean	16.09	16.48	.1527	.1567	2.58
	Sigma	4.63	3.77			
2. Arithmetical Problems	Mean	8.14	6.44	.0971	.0956	18.49
	Sigma	2.71	2.30			
3. Classification.....	Mean	12.98	13.85	.1425	.1611	6.03
	Sigma	4.32	3.87			
4. Artificial Language..	Mean	18.39	22.35	.1904	.2301	19.66
	Sigma	5.78	5.53			
5. Antonyms	Mean	29.80	33.35	.2496	.2618	14.55
	Sigma	7.57	6.29			
6. Number Series Completion	Mean	10.70	9.98	.1130	.1190	6.54
	Sigma	3.43	2.86			
7. Analogies	Mean	23.84	23.64	.1907	.2213	1.00
	Sigma	5.79	5.32			
8. Logical Inference....	Mean	24.03	25.16	.2160	.2227	5.41
	Sigma	6.56	5.35			
9. Paragraph Reading..	Mean	25.20	28.27	.2304	.2578	13.16
	Sigma	6.99	6.20			

NOTE:—Because of a timing error, the number of boys included in test 2 is 781.

The girls were superior on the scale as a whole. 3192 girls had an average final scale mean of 512.6738 with a sigma of 87.2334, while 4800 boys had an average scale mean of 493.8644 with a sigma of 106.4434. The mean of the boys is lower and the variability higher. The difference between the boys and girls is 12.803 times the probable error of the difference.

Average total scores and sigmas were computed for all candidates by college of choice, but the data are held confidentially and will not be reported. These computations were made to obtain an understanding of the sampling factors determining total score. Since the test is each year standardized on the basis of the group taking the examinations, the different sampling from year to year must be known to understand the nature of the factors selecting the population taking the test.

There is no apparent reason for questioning the sampling of individuals who took the test this year. Over 6000 of the 8000 candidates were applicants for colleges requiring the scholastic aptitude test of all candidates for admission. This proportion should give stability to the standards. If, in the future, many colleges require only their doubtful applicants to take the test, and if the number of candidates thus selected for examination becomes proportionately large, the standardizing of the test on the basis of such a group would be somewhat uncertain. The results would be much more stable if the colleges would require either all or none of their candidates to take the test. A geographical selection of candidates would probably not affect the result, but a selection by failure to receive school certificates might alter the standardization.

The present situation is satisfactory, and enough control tests are in the possession of the committee to watch the possibility of a sampling error entering into the experiment.

Tentative Estimates of Validity

The validity of the single tests is derived from their correlations with criteria of academic achievement and with other measures of success. At the present time, of course, no correlations with academic attainment are possible. A tentative effort to test the validity of the tests was made by correlating scores of each of the nine tests with age. One of the conditions of eligibility for the test was graduation from a secondary school. There are many individual factors determining age of graduation from a secondary school, but by and large, the candidates graduating under age are brighter. The tests, by this criterion, should show negative correlations with age.

The sample of the total group selected for correlating tests with age was made by taking every boy and girl from the total group who was born in September of any year. This procedure should give a random selection of cases. The data are shown in Table XII. Scores in the nine sub-tests are given in terms of the scale in which the mean is 50 and sigma 10. Total score is in terms of the final scale in which the mean is 500 and sigma 100. Since the sigmas of ages and test scores of boys and girls are not alike the coefficients are not comparable. The last column gives the coefficients of girls expanded to correct for the restriction in range or the influence of double selection. These coefficients are comparable with those of boys. The average age for 406 boys included in this table was 18.4729, and sigma was 1.2488. On account of timing errors the number of boys included in test 2 is 388 (M=18.4974 and sigma=1.2508). The average age of 265 girls included was 18.0265 and their sigma was .9572.

From experience, it may be stated that the range of coefficients shown in Table XII indicates that the test battery will probably have satisfactory validity. The differences between the possible differential selection of boys and girls by age of graduation is not known and not understood. These preliminary findings suggest that the same test may have differential validities for the two sexes.

TABLE XII

CORRELATIONS OF TESTS WITH AGE, AND MEANS AND SIGMAS OF TEST SCORES

SUB-TEST	MEANS AND SIGMAS OF TEST SCORES BOYS	GIRLS	COEFFICIENT OF CORRELATIONS BOYS	GIRLS	GIRLS (CORRECTED)
1. Definitions	51.45	49.87	−.3021	−.1101	−.1758
	11.12	9.02			
2. Arithmetical Problems	54.30	46.46	−.2842	−.1181	−.1831
	9.50	7.94			
3. Classification	50.28	50.86	−.2177	−.0904	−.1361
	10.28	8.87			
4. Artificial Language	49.84	55.93	−.3615	−.2628	−.3145
	9.11	9.74			
5 Antonyms	49.04	52.25	−.2618	−.0294	−.0467
	10.59	8.70			
6. Number Series Completion	52.11	50.25	−.2400	−.1203	−.1523
	9.70	9.96			
7. Analogies	50.28	48.97	−.2450	−.1776	−.2550
	10.49	9.39			
8. Logical Inference	50.53	50.93	−.3120	−.1773	−.2831
	10.38	8.31			
9. Paragraph Reading	49.69	52.61	−.2139	−.0231	−.0350
	11.07	9.53			
TOTAL SCALE	506.11	506.53	−.3867	−.1912	−.2985
	108.30	88.56			

A tentative estimate of "academic validity" may be made at this time by finding the correlation of the nine sub-tests with the so-called Princeton "bogie-grade". This grade is a predicted academic grade based on the weighted average of the College Entrance Examination Board's regular examinations, school grades, and the scholastic aptitude test, these three factors being combined by a multiple regression equation. This predicted grade correlates about .75 with future academic standing. As college grades will not be available for checking the tests for another year, or perhaps for two years, this provisional criterion is the only one now available. The correlation of each of the nine sub-tests with Princeton "bogie" is given in Table XIII.

TABLE XIII

CORRELATION OF THE NINE SUB-TESTS AND THE SCALE TOTAL WITH THE PRINCETON "BOGIE GROUP"

SUB-TEST	CORRELATION	MEAN	SIGMA
1. Definitions	.4151	55.4726	8.8378
2. Arithmetical Problems	.3035	56.0110	9.0460
3. Classification	.2793	52.5930	10.3964
4. Artificial Language	.3253	52.8720	9.5150
5. Antonyms	.4060	53.9122	8.3842
6. Number Series Completion	.3175	54.7536	10.3538
7. Analogies	.3304	52.9016	8.8844
8. Logical Inference	.3247	53.8276	9.0240
9. Paragraph Reading	.4096	53.4472	9.1170
SCALE TOTAL	.5189	555.74	88.4100

NOTE:—The number of cases for all sub-tests except No. 2 is 473. The mean "bogie group" for these cases is 2.7635 and sigma is .5721. There are 454 cases used for sub-test 2, the mean being 2.7561 and sigma .5740.

It might happen that the tests used would have different validities in boys' engineering colleges and boys' liberal arts colleges. The study of the correlations of various tests with varying curricula would again lead into the determination of optimum weights for predicting different types of academic work. At some time in the future, then, it may be expected that differential weights will be applied to the tests for varying purposes. Boys and girls, engineering applicants, and liberal arts applicants would all take the same examination under identical conditions and their scores would all be expressed on a scale in which the mean is 500 and sigma 100, but the various tests determining the total score would be differentially weighted depending on the sex of the applicant or the type of curriculum to be taken.

Inter-Correlations of the Tests

The amount of inter-correlation between the nine sub-tests is of prime importance. Theoretically, the best results would be obtained from a group of tests each of which had a high correlation with academic criteria and low inter-correlations. One sample of 300 boys and another sample of 300 girls were selected for the purpose of obtaining inter-correlations. A normal distribution was calculated for two groups having a mean and sigma typical of the mean and sigma of the boys and girls respectively. Each group selected was strictly normal, the distribution being truncated at a distance of three sigma from the mean. Each group had the same mean and sigma (to one decimal) as the sample of 6000 cases on which the final scale conversion was based. The inter-correlations for boys are given in Table XIV, and for girls in Table XV. Since the sigmas for the boys and girls are not the same, the coefficients for the girls shown in Table XVI have been calculated as they would be if girls had the same sigma as boys. The coefficients in Table XVI are comparable with those in Table XIV. The means and sigmas of the 300 boys and 300 girls in the nine sub-tests are given in Table XVII. These values are in terms of the scale in which the mean is 50 and sigma 10.

TABLE XIV

COEFFICIENTS OF CORRELATION BETWEEN NINE SUB-TESTS FOR A GROUP OF 300 BOYS

	1	2	3	4	5	6	7	8	9
1. Definitions		.4921	.5804	.4995	.5866	.3130	.6310	.5668	.5199
2. Arithmetical Problems	.4921		.4164	.2786	.3412	.5141	.4147	.5164	.4103
3. Classification	.5804	.4164		.4314	.5124	.2426	.5640	.4805	.5217
4. Artificial Language	.4995	.2786	.4314		.5295	.3201	.4471	.3996	.4527
5. Antonyms	.5866	.3412	.5124	.5295		.2823	.5793	.5651	.7422
6. Number Series Completion	.3130	.5141	.2426	.3201	.2823		.2525	.4082	.3216
7. Analogies	.6310	.4147	.5640	.4471	.5793	.2525		.6226	.5933
8. Logical Inference	.5668	.5164	.4805	.3996	.5651	.4082	.6226		.6212
9. Paragraph Reading	.5199	.4103	.5217	.4527	.7422	.3216	.5933	.6212	

TABLE XV

COEFFICIENTS OF CORRELATION BETWEEN NINE SUB-TESTS FOR A GROUP OF 300 GIRLS

	1	2	3	4	5	6	7	8	9
1. Definitions		.3078	.4710	.2980	.5216	.3288	.5189	.4039	.4729
2. Arithmetical Problems	.3078		.2525	.2868	.2859	.5356	.3234	.4295	.3826
3. Classification	.4710	.2525		.3244	.4543	.2124	.5182	.4584	.3826
4. Artificial Language	.2980	.2868	.3244		.3663	.3095	.4372	.3538	.2375
5. Antonyms	.5216	.2859	.4543	.3663		.3001	.5219	.4803	.6887
6. Number Series Completion	.3288	.5356	.2124	.3095	.3001		.2755	.4621	.3698
7. Analogies	.5189	.3234	.5182	.4372	.5219	.2755		.5589	.4653
8. Logical Inference	.4039	.4295	.4584	.3538	.4803	.4621	.5589		.5506
9. Paragraph Reading	.4729	.3826	.3826	.2375	.6887	.3698	.4653	.5506	

TABLE XVI

COEFFICIENTS OF CORRELATION FROM TABLE XI CORRECTED FOR DOUBLE SELECTION

	1	2	3	4	5	6	7	8	9
1. Definitions		.4589	.6196	.3872	.6794	.4183	.6440	.5243	.6121
2. Arithmetical Problems	.4589		.3493	.3674	.3981	.6371	.4187	.5456	.5029
3. Classification	.6196	.3493		.3845	.5656	.2511	.6024	.5440	.4708
4. Artificial Language	.3872	.3674	.3845		.4372	.3384	.4872	.4014	.2788
5. Antonyms	.6794	.3981	.5656	.4372		.3565	.6125	.5736	.7840
6. Number Series Completion	.4183	.6371	.2511	.3384	.3565		.3075	.5107	.4219
7. Analogies	.6440	.4187	.6024	.4872	.6125	.3075		.6245	.5387
8. Logical Inference	.5243	.5456	.5440	.4014	.5736	.5107	.6245		.6305
9. Paragraph Reading	.6121	.5029	.4708	.2788	.7840	.4219	.5387	.6305	

TABLE XVII

MEANS AND SIGMAS OF EACH SUB-TEST FOR THE NORMAL GROUPS OF 300 BOYS AND 300 GIRLS USED IN COMPUTING INTER-CORRELATIONS

SUB-TEST		300 BOYS	300 GIRLS
1. Definitions	Mean	50.18	50.13
	Sigma	10.06	7.95
2. Arithmetical Problems	Mean	53.52	44.70
	Sigma	9.92	7.99
3. Classification	Mean	49.23	50.64
	Sigma	10.89	9.52
4. Artificial Language	Mean	48.50	54.59
	Sigma	9.10	8.59
5. Antonyms	Mean	49.03	52.03
	Sigma	10.54	9.07
6. Number Series Completion.	Mean	50.85	49.67
	Sigma	10.01	9.60
7. Analogies	Mean	40.83	48.95
	Sigma	10.30	9.53
8. Logical Inference	Mean	49.43	50.36
	Sigma	9.96	9.13
9. Paragraph Reading	Mean	48.90	52.07
	Sigma	10.48	9.36

Summary of Sex-Differences

The differences between boys and girls are significant in average score in several of the sub-tests (Table XI). There are differences shown in the tables of inter-correlations, and there may be differences in the validities of the tests for boys and girls. When the scores on the sub-tests are converted into a scale in which the mean is 50 and sigma 10, and the nine tests are added to obtain the total score, each test enters into the table with a weight of one. Unit weight should always be used before empirical evidence showing better weights is obtainable, but such weights eventually will be obtainable after validities in boys' and girls' colleges are known. A study of these would lead to the discovery of optimum weights to be applied to test scores for boys and girls separately.

The Otis Self-Administering Test

The advice of the committee has frequently been requested concerning tests equivalent to the scholastic aptitude test of the College Entrance Examination Board which may be given in secondary schools. There are many tests available, but the committee has no information concerning the relative merits of these tests and can give no advice. For experimental purposes, the committee decided to include one of the available tests in the thirty-minute time limit provided for the experimental section (sub-test 10). The Thurstone Test IV and the Otis Self-Administering Tests of Mental Ability (higher form) are both thirty-minute tests. Permission was obtained from both the authors and the publishers of these tests to use them, but the amount of space required by the Thurstone Test IV made it impossible to include it in the 32-page test booklet. The Otis Higher Form A was reprinted through the courtesy of the World Book Company, Yonkers, New York.

It would be expected that a test designed for use in high schools would not differentiate the upper range of college applicants and this proved to be the case, although this is not a reflection upon the Otis test. 1080 candidates were given the Otis test with a half-hour time limit. The mean scholastic aptitude score of these candidates (randomly sampled) was 499.213 and the sigma was 99.272. The mean Otis score was 58.7435, sigma 8.6839. The correlation between the two examinations was .7931. As the maximum possible score on the Otis examination was 75, the distribution of Otis scores was cut off sharply at the upper end and high score candidates probably were not completely differentiated.

Table XVIII gives the most probable scholastic aptitude test score, percentile, and letter grade for each Otis total score. The Otis I. Q. conversions were not used, and this table applies only to total raw scores. The coefficient of correlation between the Otis test and the Board test (.7931) means considerable uncertainty in predicting one score from another. The probable error of estimating a scholastic aptitude test score from an Otis score is about 40. A "predicted" S. A. T. score from an Otis score of 50, for example, would be 420 ± 40. This means that the chances are even that the true S. A. T. score is between 380 and 460. The chances are also even that the score is greater than 460 or less than 380, but become increasingly smaller as the distance from these numbers becomes larger.

TABLE XVIII

THE MOST PROBABLE SCHOLASTIC APTITUDE TEST SCORE, PERCENTILE AND LETTER GRADE FOR EACH TOTAL SCORE IN THE OTIS SELF-ADMINISTERING TESTS OF MENTAL ABILITY (HIGHER EXAMINATION, FORM A, THIRTY-MINUTE TIME LIMIT)

OTIS SCORE	S. A. T. SCORE	PERCENTILE	LETTER GRADE
25	193	.1	E
26	202	.1	E
27	211	.1	E
28	220	.2	E
29	230	.3	E
30	239	.4	E
31	248	.5	E
32	257	.7	E
33	266	.9	E
34	275	1	E
35	284	1	E
36	293	1	E
37	302	2	E
38	311	2	E
39	320	3	E
40	329	4	E
41	338	5	E
42	347	6	E
43	356	7	D
44	366	8	D
45	375	10	D
46	384	12	D
47	393	14	D
48	402	16	D
49	411	18	D
50	420	21	D
51	429	24	D
52	438	26	D
53	447	29	D
54	456	32	C
55	465	36	C
56	474	39	C
57	483	43	C
58	492	46	C
59	502	50	C
60	511	54	C
61	520	58	C
62	529	61	C
63	538	64	C
64	547	68	C
65	556	71	B
66	565	74	B
67	574	77	B
68	583	79	B
69	592	82	B
70	601	84	B
71	610	86	B
72	619	88	B
73	628	90	B
74	638	91	B
75	647	92	B

Special Study of Reliability

There are two approaches to the problem of determining the reliability of a battery of tests, each requiring different techniques. One method of determining reliability of tests which has come into general usage is the split-half technique, or the correlation of random halves of a single test paper. To determine reliability by this method, it is only necessary to divide a candidate's score in a test into two random halves, and this is conventionally done by computing his total score for odd-numbered items and for even-numbered items. The correlation r between the total score on odd and even numbered items is then conventionally changed by the Spearman-Brown formula $2r/(1+r)$ to represent the reliability of the test. This method brings into relief the factors of unreliability due to test construction.

The other approach to the problem of estimating the reliability of an examination adds to the factors due to test construction those factors causing unreliability due to variable factors in the candidates taking the test. The essential procedure, in this case, consists in giving alternate forms of the same examination to the same subjects. The reliability of the examination is the correlation between first and second trials of the test, the paper given at the second trial being an alternate form of the paper given at the first trial. In other words, the subject is given two tests and the scores made in the different trials are correlated to determine reliability.

There are several controversial issues involved in estimating reliability, particularly in assuming identity of the two methods. The committee arranged an experiment for determining reliability and will report the results by the two methods without comment. In reporting reliability coefficients obtained by the split-half technique, both the value of r and the values of $2r/(1+r)$ will be given.

Through the cooperation of the War Department an arrangement was made with the authorities of the United States Military Academy at West Point to give the scholastic aptitude tests to the incoming (Fourth Year) class on July 22 and 23, 1926. The committee, while preparing Form A of the scholastic aptitude test for use in June 1926, also prepared an alternate form of this test

called Form B. Forms A and B contain different items, but the tests are of the same type so that one practice booklet may be used in preparation for both forms, and the same instructions in regard to time limits may be followed.

The incoming class at the United States Military Academy is divided into six companies, assignments to companies being made according to height, a procedure which should give a random intellectual sampling in the six companies. On July 22, Form A was given to the first, second, and third companies, and Form B to the fourth, fifth, and sixth companies. On July 23, Form B was given to the first, second, and third companies, and Form A to the fourth, fifth, and sixth companies. The group taking Form A the first day and Form B the second day will be called the A-B group, and the group taking Form B the first day and Form A the second day will be called the B-A group. The A-B group included 168 men, the B-A group, 164 men.

All procedures concerning the conduct of the examination were carefully followed. Practice booklets were issued in advance of the examination and carefully studied. The group examined proved to be typical of boys who took the test in June. The two forms used were given in different order to the two groups in order to equalize the practice effect from Form A to Form B, and *vice versa*. The practice effect being eliminated in this manner, the correlation between first and second trials may be taken as one index of the reliability of the tests.

The correlations between the first and second trials of the tests are given for the A-B and B-A groups separately in Table XX, and the means and sigmas for Forms A and B for these groups are given in Table XXI. Since the practice from Form A to Form B, and from Form B to Form A may be assumed to be equal, the correlation between the first and second trials, regardless of form used, may be taken as an indication of the reliability of the tests. These nine coefficients are as follows:

TABLE XIX

SUB-TEST	
1. Definitions	.6981
2. Arithmetical Problems	.7766
3. Classification	.6984
4. Artificial Language	.8350
5. Antonyms	.7468
6. Number Series Completion	.7827
7. Analogies	.6820
8. Logical Inference	.8290
9. Paragraph Reading	.8028

TABLE XX

CORRELATIONS BETWEEN FORMS A AND B FOR THE A-B GROUP (N=168) AND FOR THE B-A GROUP (N=164)

SUB-TEST	TIME ALLOWANCE	A-B GROUP CORRELATIONS FORM A AND B	B-A GROUP CORRELATIONS FORM A AND B
1. Definitions	9 min.	.7176	.7607
2. Arithmetical Problems	8 min.	.7752	.7971
3. Classification	6 min.	.6937	.7144
4. Artificial Language	9 min.	.8870	.7903
5. Antonyms	10 min.	.8003	.8059
6. Number Series Completion	9 min.	.8247	.8163
7. Analogies	6 min.	.8151	.7434
8. Logical Inference	10 min.	.8392	.8358
9. Paragraph Reading	30 min.	.7863	.8351

TABLE XXI

MEANS AND SIGMAS OF FORMS A AND B FOR THE A-B AND B-A GROUPS

SUB-TEST		A-B GROUP FORM A	A-B GROUP FORM B	B-A GROUP FORM B	B-A GROUP FORM A
1. Definitions	Mean	15.9762	18.6786	14.7318	19.2560
	Sigma	4.7422	4.9464	4.8208	4.6344
2. Arithmetical Problems	Mean	8.3690	9.4821	7.7744	9.6829
	Sigma	2.4362	3.0608	2.5239	3.2134
3. Classification	Mean	13.2976	14.2738	12.6220	14.3780
	Sigma	4.5416	4.5010	4.2122	4.9814
4. Artificial Language	Mean	16.1190	21.2858	15.1524	20.3598
	Sigma	5.2014	6.5554	5.1094	6.3932
5. Antonyms	Mean	27.3394	25.3930	24.5062	27.9268
	Sigma	7.4958	7.5891	7.6146	7.2732
6. Number Series Completion	Mean	10.1904	12.5834	10.9146	11.7196
	Sigma	3.4286	3.7566	4.0682	4.0636
7. Analogies	Mean	23.3096	22.3572	20.6342	24.8292
	Sigma	6.2764	5.1250	5.1992	6.0778
8. Logical Inference	Mean	23.6428	25.6190	22.2440	25.9756
	Sigma	6.3680	6.2744	6.8616	6.3008
9. Paragraph Reading	Mean	23.7322	26.1430	23.8660	24.4147
	Sigma	7.8114	8.4081	8.9652	8.0259

A question of interest, but one not concerned with reliability, is the relative difficulty of the two forms of the test which had been constructed so that they would be of approximately the same difficulty. The ratio

$$\frac{\text{First Trial}}{\text{Second Trial}}$$

indicates the practice effect. The smaller the ratio the larger the effect of practice. The means and sigmas of the first and second trials and the ratios between these means and sigmas are shown in Table XXII.

TABLE XXII

MEANS AND SIGMAS OF FIRST AND SECOND TRIALS, AND RATIOS OF FIRST TRIAL AND SECOND TRIAL

SUB-TEST		FIRST TRIAL	SECOND TRIAL	RATIOS MEANS	RATIOS SIGMAS
1. Definitions	Mean	15.3615	18.9638	.8100	1.0037
	Sigma	4.8214	4.8036		
2. Arithmetical Problems	Mean	8.0753	9.5813	.8428	.7957
	Sigma	2.4977	3.1388		
3. Classification	Mean	12.9639	14.3253	.9050	.9263
	Sigma	4.3948	4.7446		
4. Artificial Language	Mean	15.6415	20.8284	.7510	.7977
	Sigma	5.1788	6.4924		
5. Antonyms	Mean	25.9399	26.6446	.9736	1.0191
	Sigma	7.6863	7.5420		
6. Number Series Completion	Mean	10.5481	12.1567	.8677	.9805
	Sigma	3.8586	3.9350		
7. Analogies	Mean	21.9880	23.5783	.9325	1.0299
	Sigma	5.9224	5.7502		
8. Logical Inference	Mean	22.9518	25.7952	.8898	1.0577
	Sigma	6.6532	6.2900		
9. Paragraph Reading	Mean	23.7983	25.2893	.9410	1.0162
	Sigma	8.4012	8.2668		

TABLE XXIII

DIFFERENCES BETWEEN FORMS A AND B AFTER ELIMINATING THE EFFECT OF PRACTICE

SUB-TEST		FORM A	FORM B
1. Definitions	Mean	15.7868	14.9307
	Sigma	4.6793	4.8740
2. Arithmetical Problems	Mean	8.2649	7.8830
	Sigma	2.5128	2.4951
3. Classification	Mean	13.1548	12.7699
	Sigma	4.5779	4.1907
4. Artificial Language	Mean	15.7046	15.5690
	Sigma	5.1506	5.1693
5. Antonyms	Mean	27.2645	24.6144
	Sigma	7.4540	7.6743
6. Number Series Completion	Mean	10.1797	10.9166
	Sigma	3.7065	3.8758
7. Analogies	Mean	23.2314	20.7411
	Sigma	6.2762	5.2456
8. Logical Inference	Mean	23.3779	22.5199
	Sigma	6.5162	6.7490
9. Paragraph Reading	Mean	23.3532	24.2333
	Sigma	7.9837	8.7548

The means of Form A and B second trial may be multiplied by the ratios listed in Table XXII to reduce them to first trial means. Similarly the second trial sigmas may be reduced by the ratios in the last column of Table XXII. When this has been done, the average of the first trial of Form A and B as found, and the second trial changed as described will show whatever differences may exist between the two forms. These differences are listed in Table XXIII.

The differences between forms are not large, but they illustrate the difficulty of constructing tests that are alike. The two forms of sub-tests 1, 3, 6, 7, and 8 were constructed from data obtained by giving all items to the same groups or matched groups. The two forms of sub-tests 2 and 4 were made equivalent on an *a priori* basis. No special effort was made to match the two forms of sub-test 5. Two forms of sub-test 9 were made equivalent by a very elaborate but somewhat roundabout method.

By means of the equation

$$A = \frac{\sigma A}{\sigma B} B + M_A - \frac{\sigma A}{\sigma B} M_B$$

the A score corresponding to any given B score was computed for each of the nine tests, and a conversion table constructed

similar to that in Table VIII for converting any B score to a standard scale in which the mean was 50 and sigma 10, the standards for the two forms being alike. Inequalities between forms are thus eliminated, and standard scores obtained for Form B equivalent to the scores for Form A standardized on the basis of 1500 cases.

The scores of all cases in the A-B and B-A groups were converted by either the A or B conversion keys, and the sum of the sigma scores for the nine tests computed. The correlations between the total scores were computed for the A-B and B-A groups separately, and for the first and second trial. The correlations were as follows:

TABLE XXIV

A-B group	.9428
B-A group	.9577
First and Second Trial	.9497

The means and sigmas were as follows:

	MEAN	SIGMA
A-B GROUP		
Form A	435.8100	69.1500
Form B	472.8340	70.9020
B-A GROUP		
Form B	430.1100	70.2760
Form A	469.1340	77.8820
First Trial	432.9940	69.7660
Second Trial	471.0060	74.4540

The effect of practice in taking the tests amounts to 38 points on the nine tests, or about 60 points on the final scale used in which the mean is set at 500 and sigma at 100. By the method outlined before for the single tests, another key was constructed for converting scores from Form B to the previously standardized scale in which the mean was 500 and sigma 100. This being done, the committee was in a position to use Form B for the purposes of re-examination in September 1926, if such re-examinations were required.

The method of estimating the reliability of tests by giving alternate forms of the same test on successive days includes the variable factors due to the subject as well as those due to the tests. The estimate of the reliability of the examination obtained in this way is the correlation of .95 (.9497) between first and second trials. This yields a probable error of estimate of 21 points indicating that the committee's guess on page 20 of the "Manual for the Use of College Officers" was too conservative (25 to 30 points).

The method of estimating reliability by correlating odd and even numbered items was tried for each of the 332 cases who took Form A and each of the 332 cases who took Form B at West Point. The nine coefficients are given in Table XXV. The means and sigmas for odd and even numbered items for Forms A and B are given in Table XXVI.

TABLE XXV

RELIABILITIES OF TESTS ESTIMATED BY CORRELATING ODD AND EVEN NUMBERED ITEMS

SUB-TEST	NUMBER OF ITEMS	FORM A r	FORM A $\frac{2r}{1+r}$	FORM B r	FORM B $\frac{2r}{1+r}$
1. Definitions	30	.8050	.8920	.7658	.8674
2. Arithmetical Problems	20	.6608	.7958	.6336	.7757
3. Classification	40	.6647	.7986	.6451	.7843
4. Artificial Language	74	.9623	.9808	.9544	.9767
5. Antonyms	48	.7601	.8637	.7888	.8819
6. Number Series Completion	25	.7634	.8658	.7803	.8766
7. Analogies	40	.8175	.8996	.7107	.8309
8. Logical Inference	40	.7226	.8390	.8049	.8919
9. Paragraph Reading	50	.7910	.8833	.8184	.9001

The correlation between the sum of all odd numbered items and all even numbered items for the nine tests is .9597 for Form A and .9572 for Form B. These coefficients expanded by the formula $2r/(1+r)$ become .9794 for Form A and .9781 for Form B. When odd and even numbered items are added in this way, the tests enter into the total with weights proportional to the sigmas of the tests. To obtain a total in which the tests were weighted equally, the scores in the odd and even numbered items were doubled and converted by the conversion keys for Forms A and B into the scale in which the mean is 50 and sigma 10. The correlation between these equally weighted halves of Form A was .9528, and of Form B, .9366. Expanded by the Spearman-Brown formula, these coefficients become .9758 for Form A and .9672 for Form B.

TABLE XXVI

MEANS AND SIGMAS OF ODD AND EVEN NUMBERED ITEMS OF FORMS A AND B

SUB-TEST		FORM A EVEN	FORM A ODD	FORM B EVEN	FORM B ODD
1. Definitions	Mean	9.1928	8.3855	8.6355	8.0873
	Sigma	2.5693	2.5897	2.8827	2.6340
2. Arithmetical Problems	Mean	4.0120	5.0211	3.6627	4.9729
	Sigma	1.6354	1.5550	1.6283	1.6232
3. Classification	Mean	6.4970	7.1205	6.1084	7.5030
	Sigma	2.4964	2.6213	2.2509	2.4952
4. Artificial Language	Mean	18.4698	18.1868	18.2410	18.2530
	Sigma	6.0688	6.1896	6.6000	6.7324
5. Antonyms	Mean	13.2410	14.4698	11.1746	13.5362
	Sigma	3.6328	4.2974	3.9822	3.9468
6. Number Series Completion	Mean	5.4247	5.5000	5.7982	5.9398
	Sigma	1.9659	2.1684	1.9921	2.1682
7. Analogies	Mean	11.7229	12.3343	10.8464	10.6355
	Sigma	3.1396	3.4159	3.1533	2.5084
8. Logical Inference	Mean	11.9066	12.8675	10.8283	13.1627
	Sigma	4.0233	2.8812	3.5994	3.4261
9. Paragraph Reading	Mean	11.8916	12.2290	12.8254	12.3072
	Sigma	3.8656	4.4734	4.3536	4.7654

In summarizing the data on reliability it may be stated that the general situation as revealed by this analysis is satisfactory. There should be no reason why tests could not be built so that the value of $2r/(1+r)$ (r between random halves) would be as high as .995 even in the highly restricted upper range of ability in which the tests are used. The program of building such tests is a long one but not an impossibility.

The Experimental Sections

The time schedule followed in giving Form A of the scholastic aptitude tests in June 1926 was as follows:

TABLE XXVII

Sub-test 1. Definitions	9 minutes
Sub-test 2. Arithmetical Problems	8 minutes
Sub-test 3. Classification	6 minutes
Sub-test 4. Artificial Language	9 minutes
Rest Period	6 minutes
Sub-test 5. Antonyms	10 minutes
Sub-test 6. Number Series Completion	9 minutes
Sub-test 7. Analogies	6 minutes
Sub-test 8. Logical Inference	10 minutes
Rest Period	10 minutes
Sub-test 9. Paragraph Reading	30 minutes
Rest Period	6 minutes
Sub-test 10. Experimental Section	30 minutes

The first nine tests were the same for all candidates. Seven forms designated A_1, A_2, A_3, A_4, A_5, A_6, and A_7, were printed, each of the seven forms containing a different sub-test 10. One of these forms, for example, included the Otis test as explained. The papers were handed to the candidates in rotation, thus insuring a random sampling of candidates taking any given experimental section.

The purpose of the experimental sections is to provide test items, which are known to be diagnostic, for use in future forms. Such items can only be discovered empirically by the analysis of the results of actual testing of the item on the group for which it was designed. Any person can invent test items, but the merit of any given item may only be determined objectively. The value of an item must not be dependent on the ability of its inventor.

As an example of the method of determining the value of an item empirically, one of the items of the two hundred included in sub-test 10 of Form A-6 consisted of a definition with the word defined omitted, and listing five choices from which the candidate selected the word which best fitted the definition. The definition was "A..... is an abrupt change in feeling, opinion, or action, due to some fancy." The five choices

were "decision", "temperament", "dream", "convulsion", and "caprice". Table XXVIII shows the distribution of the total scores in the scholastic aptitude test, as reported to the colleges on the scale in which the mean is 500 and sigma 100, of 583 candidates who answered the item in various ways. The omissions were few in number (9) and "dream" proved unpopular, but the other possible responses were chosen frequently.

TABLE XXVIII

DISTRIBUTION OF RESPONSES OF 583 CANDIDATES IN ITEM 20, SUB-TEST 10, FORM A-6

S. A. T. SCORES	OMITTED	1 "DECISION"	2 "TEMPERAMENT"	3 "DREAM"	4 "CONVULSION"	5 "CAPRICE"
788 & above						1
763-787						1
738-762						12
713-737						8
688-712					1	10
663-687					1	14
638-662					3	25
613-637			1		1	26
588-612		7	3			40
563-587		1	5		1	39
538-562		7	3		1	44
513-537		4	5		4	53
488-512	1	13	5	1	7	45
463-487		8	10	1	6	33
438-462		6	3		2	14
413-437	2	7	5	1	1	21
388-412	1	10	6	3	2	12
363-387	1	4	3			3
338-362	1	5	1		1	2
313-337	2	3	2			
288-312					1	1
263-287	1		3			1
238-262						
213-237						
212 & below						1
Number of Cases	9	75	55	6	32	406
Mean	377	465	467	432	510	552

The coefficient of correlation (bi-serial r) between the right answer (#5) and all other answers is .53. The average S. A. T. score of those giving the right answer was 552 (sigma 90.5), while the average score of those omitting the item or choosing a wrong response was 467 (sigma 85.0). The difference between the average S. A. T. scores of candidates having the right and wrong solutions was 85 points, a difference which is 16 times the probable error of the difference. This means that this particular item very sharply differentiates the group.

Similar studies of items by this method, and other methods which are available, show that items vary in their power to differentiate the group. No matter how much care is exercised in inventing test items, their diagnostic values can be determined only by methods similar to that illustrated—they can not be predicted. A very large number of test items must therefore be discarded, after analysis, as worthless, and it is only by such laborious methods that valid tests may be constructed. A test containing 50 items of the differentiating power of the item illustrated would give very good results. As the complete series of tests taken by the candidates includes four or five hundred items, the process of constructing such a series requires a large amount of experimental work. The purpose of the experimental sections is to collect the data to be used for studies designed to improve the tests.

The experimental sections included, in addition to the Otis test, the following tests given in the thirty-minute time limit:

Verbal Analogies	200 items
Pictorial Analogies	90 items
Antonyms	200 items
Written Directions	60 items
Definitions Completion	200 items
Pictorial Series	88 items

About one-third of the work of analyzing these experimental sections by separate test items in the manner illustrated has been completed.

The General Plan of Testing

The general plan of testing employed in this work follows conventional procedures rather closely. One departure from current practice is that of standardizing the test on the basis of the group examined instead of building an entire series of alternate forms of equal difficulty. This procedure was, of course, dictated by the desirability of making a general test of a nature such that the specific tests might be changed when the necessity arose. The general effect of special coaching on tests is not known, but will be studied, and should results be influenced in this way, entirely new test techniques will be available for use.

This elasticity of procedure will enable the committee to improve the tests by the use of data from the experimental sections and from other specially arranged experiments. It is almost certain that these methods of analysis will result in the gradual improvement of test techniques. The elasticity of procedure, further, enables the committee to use regression weighting of tests for different types of curricula or for the two sexes without altering the general standardization. It does not seem unreasonable to expect improvement in both the validity and reliability of the tests as time goes on. The committee, at the start, finds no problem involved that does not show promise of a satisfactory empirical solution.

Respectfully submitted,

ROSWELL P. ANGIER, Yale University
ANDREW H. MACPHAIL, Brown University
DAVID C. ROGERS, Smith College
CHARLES L. STONE, Dartmouth College
CARL C. BRIGHAM, Princeton University, (*Chairman*)