THE CONSTRUCTION AND VALIDATION OF AN ARITHMETICAL COMPUTATION TEST

NORMAN FREDERIKSEN
Princeton University and Educational Testing Service

and

G. A. SATTER
University of Michigan

Introduction

THIS article is a condensation of two reports (2, 5) prepared by National Defense Research Committee Project N-106 during World War II under the direction of Harold Gulliksen, Project Director. The summary is presented with no thought that the particular test under consideration is of unusual interest, but rather to illustrate several methods of test construction and validation which are unique or rarely employed and which deserve to be brought to the attention of those interested in test development techniques. Some of the procedures employed in the study and some of the findings which are thought to be of interest are as follows:

1. The tabulation and analysis of free-answer responses as a source of distracters for multiple-choice items.

2. A comparison of measures of item difficulty under free-answer and multiple-choice conditions.

3. The use of item-analysis data for multiple-choice responses in selecting distracters.

4. The use of *answer not given* as an item-response.

5. The correlation between numerical and verbal ability in a heterogeneous population.

6. The use of machine scoring with a test which does not require a separate answer sheet and which provides space for figuring.

7. A method of validating a test which does not require waiting for criterion data to become available.

Purpose

A number of correlational studies (1, 4) have shown that performance on the *United States Navy Arithmetical Reasoning Test* was closely related to performance on the more purely verbal measures, such as the *General Classification Test* and the *Reading Test.* Such findings suggested the desirability of introducing a test of quantitative ability which would be "purer" in the sense that scores on it would reflect more of the ability to manipulate numbers and less of the ability to read directions or to manipulate verbal symbols. This report describes the procedure followed in constructing and validating a test of arithmetical computation designed to make fewer demands on verbal ability and to measure those types of computational skills which are demanded by service school curricula.

The Construction and Analysis of a Free-Answer Test, Form X-1

The first step in building the *Arithmetical Computation Test* was to assemble a group of items which would sample the types of computational operations most frequently demanded in service school training. From a large number of items, sixty were selected for incorporation into Form X-1. In the assembled test, the four fundamental arithmetical operations received approximately equal representation. Five problems involved the use of percentages. The problems called for a knowledge of whole numbers, fractions, and decimals, and for the ability to manipulate them. Taken as a group, these problems covered the major computational skills which should be at the command of any eighth-grade graduate.

These sixty items were assembled in a four-page test booklet under the title of *United States Navy Arithmetical Computation Test, Form X-1.* The items were presented in "free-answer" form, i.e., as individual computational problems with an accompanying space where the answer could be written. In order that computational errors might be analyzed, space was provided for figuring. The answers given in the trial administration provided suggestions for constructing the distracters for a revised multiple-choice test.

Form X-1 was administered on October 14–15, 1943, to some 1,430 recruits at the United States Naval Training Center at

Sampson, New York. The test was given during the regular periods assigned to selection testing. The directions were given orally and time was called after 45 minutes had elapsed.

The completed test booklets were assembled and random samples of them were used for the analysis of each item.[1] A tabulation was made of all of the answers given to each of the individual items. The difficulty (p') of each item was determined by computing the proportion who answered it correctly.[2] The distribution of these values of p' is shown in Table 1.

TABLE 1

Distributions of p'-Values for the Items of the Arithmetical Computation Test, Form X-1

p'-values	Items in Form X-1	Items accepted for Form X-2	Items rejected for Form X-2
.90 –.99	8	3	5
.80 –.89	5	5	
.70 –.79	8	8	
.60 –.69	7	7	
.50 –.59	5	5	
.40 –.49	8	8	
.30 –.39	7	6	1
.20 –.29	6	5	1
.10 –.19	5	3	2
.00 –.09	1		1
Total	60	50	10
Median	.555	.555	.645

The values of p' were distributed over practically the entire range of difficulty. The median item was one which 55 per cent of the recruits could pass. Only six were so difficult that fewer than 20 per cent could successfully answer them. Eight of the sixty items, however, were so easy that over 90 per cent of the recruits could answer them correctly. Such easy items, or the six which were so difficult that less than 20 per cent answered

[1] The size of these samples varied from 500 to 900. The stack of 1400 test booklets was broken up into piles of 100 each, and 5 or more of these piles were used for tabulation purposes. The number of piles used depended upon the variety of answers given to the item and the need for an adequate number in each of the answer categories.

[2] In this analysis it was assumed that all of the persons taking the test responded to all of the items; i.e., N's used in computing the p' values were the same as the number of test booklets used in the analysis. In subsequent analyses, N_t (number attempting the item) was used; N_t of course did not necessarily equal the total number taking the test (Base N or N_b).

them correctly, discriminate among only a relatively few recruits and are, as a consequence, of less utility than those near the center of the distribution. All factors considered, however, enough of the items were of satisfactory difficulty to permit the construction of another test.

The tabulations of the incorrect answers to each of the sixty items of Form X-1 were the source of the distracters for Form X-2; along with the data on the difficulty of the items, they provided the basis upon which the multiple-choice form (Form X-2) was constructed.

The Construction and Analysis of a Multiple-Choice Test, Form X-2

The *United States Navy Arithmetical Computation Test, Form X-2* reflects the experience which was gained in administering and analyzing Form X-1. Form X-2 was constructed and administered primarily to provide information concerning the difficulty and discriminative value of arithmetical computation items in the multiple-choice form and to describe the reliability and validity of such a test. It was to serve as the basis for constructing a shorter multiple-choice computation test of known validity which could be considered for use as a selection measure.

The fifty items for this test (Form X-2) were selected from the sixty Form X-1 items which had been analyzed. They represent those items which the analysis of Form X-1 had shown to be neither extremely difficult nor extremely easy. Table 1 shows the distributions of the p'-values for those items selected and rejected for inclusion in the new form. Five of the rejected items were among the very easy ones and five were among the extremely difficult. The median p'-value of those items accepted (according to the Form X-1 analysis) was .55.

Distracters for the fifty selected items were chosen from the tabulations of answers to the Form X-1 problems. These tabulations showed that certain wrong answers to some problems were very popular; among other items, generally the more difficult ones, wrong responses were distributed over a wide variety of answers. For this reason primarily it was difficult, on the basis of answer count alone, to make a definite decision as to which

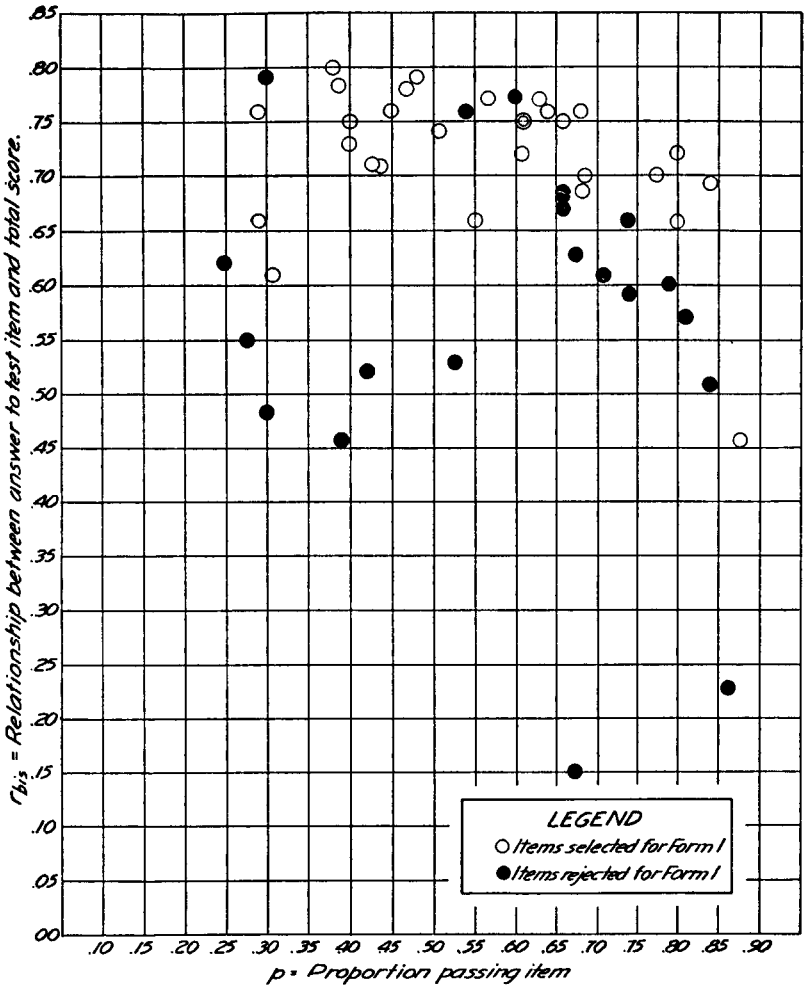
distracters should be selected for certain of the items; in cases of this sort, the data supplied by an error analysis were very helpful. Six alternatives were selected for most of the items; some were provided with seven, and a few with eight. To prevent persons from choosing an "approximate" answer without actually engaging in the operations necessary for the solution of the problem, "Answer not given" was introduced as one of the alternatives for each of the items; this was the correct answer for seven of the fifty problems. By making use of a surplus of distracters, it was hoped that additional selection could be made once tabulations were compiled for the X-1 test. This is the procedure which was followed for selecting the distracters for the items of the final form of the test (Form 1). In terms of the arithmetical operations which are represented, Form X-2 is almost identical with Form X-1.

These fifty items were lithoprinted as an eight-page test booklet entitled the *United States Navy Arithmetical Computation Test, Form X-2*. Directions for taking the test were printed on the front of the booklet; alternate pages were left blank to provide space for figuring. The problems were printed in columns with the alternatives immediately to the right of each problem.

Form X-2 of the test was then administered to a sample of approximately 1,000 recruits at two Naval Training Centers during the month of January, 1944—to 506 at Great Lakes, Illinois, and to 536 at Bainbridge, Maryland. This administration provided the data used in analyzing the items and distracters of the X-2 test. The recruits were given 40 minutes to complete the test; as the analysis later showed, this time limit was generous enough to permit three-fifths of the recruits to complete the test.

From the group of 1,042 test papers, a random sample of 500 was selected for analysis. Two hundred and fifty of these were from the Bainbridge sample and 250 from Great Lakes. Tabulations of responses to each of the fifty items of these 500 tests were made and then analyzed. Each of the items was described in terms of two of its characteristics: its difficulty and its discriminative value.

For the Form X-2 analysis, p was used as the measure of item



Note: Item 1 of Form I was taken from Form X-1; for this reason it is not included in this plot.
The p's for this plot are based on the number attempting the item(N_p).

FIG. 1 Values of p and r_{bis} for Items of Form X-2

difficulty. These proportions are based on the number (N_t) who attempted to answer the item.[3] A distribution of the p-values (see Figure 1) shows that, in general, all levels of difficulty are

[3] p was computed by dividing the number of persons who answered the item correctly by the number (N_t) who *attempted* it. An individual was judged to have attempted an item if he answered either that item or any subsequent one.

represented fairly adequately. None of the items, however, is so difficult that less than 20 per cent succeed in passing it; none is so easy that less than 10 per cent fail it. It may be argued that the latter of these distribution characteristics is undesirable; in the final form (Form 1) of the computation test, one easy item was reintroduced from Form X-1 to overcome this minor deficiency. The greatest concentration of items of similar difficulty is in the range from .60 to .69; this is definitely a desirable feature when one's objective is to make discriminations among individuals near the middle of the ability range, for it represents a concentration of items where they can make the greatest number of discriminations. The median of this distribution of p's is .61.

Since the items of Form X-2 are in multiple-choice form, one might expect them to be somewhat easier than the same items in free-answer form. Actually, however, the differences are neither pronounced nor in the anticipated direction. To illustrate the relationship between the difficulties of the two groups of items in Forms X-1 and X-2, Figure 2 was prepared. Here the p'-values of the items of the first form are plotted against the p'-values for the same items in the second experimental form; both of these sets of p'-values are based on the total number of persons taking the test.[4] This plot shows that, for the most part, the shifts in difficulty are small. Secondly, it shows that if one allows for chance success, almost all of the items in the multiple-choice form are somewhat more difficult than their companion items in the free-answer form; this apparent inconsistency probably can be explained in large part by differences in the abilities of the recruit samples taking these two forms of the test. Several of the problems were changed slightly in transferring them from Form X-1 to Form X-2; the effect of these alterations, as Figure 2 shows, was to make the problems somewhat easier.

In describing test items, however, difficulty is only one consideration. The efficiency of the items as "discriminators" is also of concern. To make this type of evaluation, biserial r

[4] Note that in Figure 2 the p-values for the X-2 items are based on N_b while those which are shown in Figure 1 and cited in the text are based on the number (N_t) responding to the item.

(r_{bis}) was employed. This measure describes the relationship between performance on the item and performance on the whole test; it was computed for each of the items of Form X-2 and is shown graphically in Figure 1. Here it can be seen that, with few exceptions, the items are of uniformly high discriminative

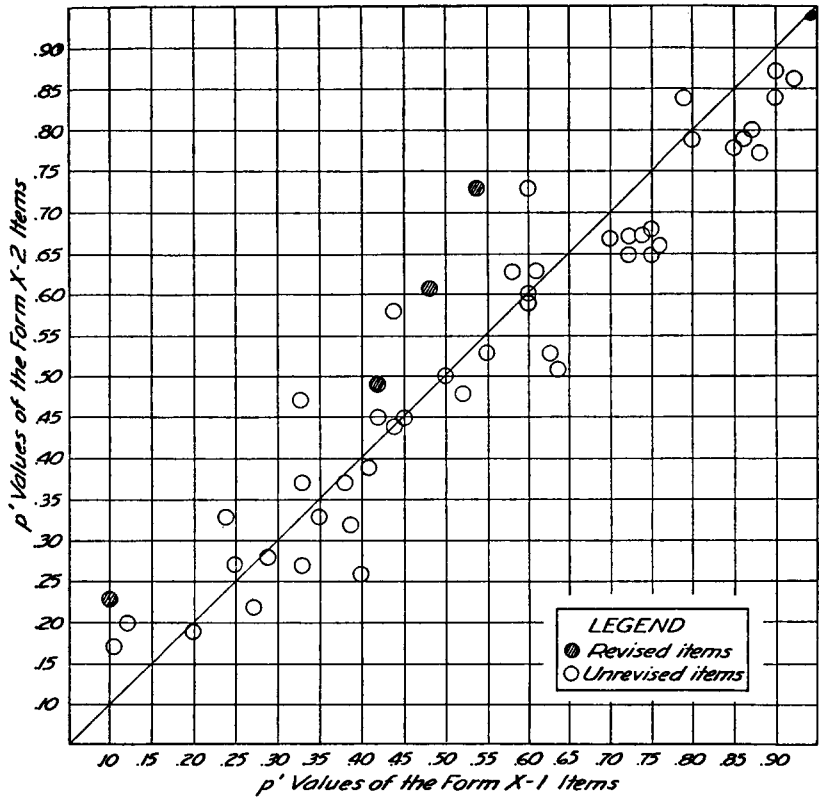


FIG. 2 Comparison of p' values for Forms X-1 and X-2

value—much more so than one ordinarily expects to find. The median of the distribution of biserial r's is .70; only five of the items have an r_{bis} of less than .50. It would seem that in terms of this criterion little more could be desired.

Thus far, two of the internal characteristics (item difficulty and discriminative value) of the test have been described; the

test as a whole remains to be considered. For this purpose, the test scores of a sample of 227 Bainbridge recruits (part of the original group) were used for analysis. A distribution was made of the X-2 scores; its mean and standard deviation were computed; the X-2 scores were correlated with scores on both the General Classification Test (GCT) and the Arithmetical Reasoning Test (AR); and the split-half reliability of the X-2 test was estimated.

For this sample of recruits the mean Form X-2 score was 24.2, slightly less than the mid-score. If the sample is typical of recruits nation-wide, it might be argued that the mean score should be somewhat higher—somewhere nearer the middle of the range between a chance score and a perfect one. There

TABLE 2

Means and S.D.'s of the General Classification Test, Arithmetical Reasoning Test, and Arithmetical Computation Test (227 Bainbridge Recruits)

Test	Mean	S.D.
AC—Form X-2	24.2	11.9
GCT—Form 2	45.0	10.3
AR—Form 2	45.9	11.4

is evidence to indicate, however, that this particular sample is somewhat inferior (see Table 2). Its mean GCT and AR scores are definitely lower than the national norms (Navy standard scores are based on a mean of 50 and a standard deviation of 10). At any rate, this particular deficiency is hardly pronounced enough to detract materially from the value of the test.

A more serious consideration is suggested by the correlation coefficients which describe the relationship between performance on the Arithmetical Computation Test and two of the tests of the Basic Battery. In constructing the AC test, a definite attempt was made to design a measure which would yield scores relatively unrelated to verbal abilities—a special attempt was made to make the directions as simple as possible, and the items themselves were prefaced with only a word to indicate the type of operation involved. In spite of these precautions, the objective was not attained. When compared to those of the Arithmetical Reasoning Test, AC scores hardly

correlate appreciably lower with General Classification Test scores (see Table 3).[5] Unquestionably, these relationships would be considerably lower for a sample of service school men because of a smaller range of ability; but the fact remains that now it is doubtful whether any type of computation test can be designed which will be uncorrelated with the "pure" verbal measures in a group which is as heterogeneous as a Navy recruit population.

TABLE 3

Intercorrelations of the General Classification Test, Arithmetical Reasoning Test, and Arithmetical Computation Test (227 Bainbridge Recruits)

Test	GCT (Form 2)	AR (Form 2)
AC—Form X-2	.758	.782
GCT—Form 2		.820

TABLE 4

Odd-Even Reliability of the Arithmetical Computation Test (200 Recruits)

Length of Test	Reliability Coefficient
25 items	.905
30 items	.920
50 items	.950

The reliability of the X-2 test was computed by correlating the scores from the odd-numbered items with those from the even-numbered ones and then applying the Spearman-Brown prophecy formula to estimate the reliability of both the whole (50 item) test and tests of 25 and 30 items—the latter being the proposed length of the final form (see Table 4). For a test of its length, the *Arithmetical Computation Test* is of exceptionally high reliability; in this respect it compares quite favorably with the existing measures of the Basic Battery.

[5] Note, however, that these coefficients are somewhat higher than those which were found in earlier correlational studies, and that the differences between "recruit sample" and "school sample" *r*'s are larger than one would anticipate in terms of the amount of curtailment in school samples. These two facts would suggest that the recruit sample used in this study is not altogether typical of recruits in general.

The Construction of the Arithmetical Computation Test, Form 1

The analysis of Form X-2 has demonstrated that in terms of internal characteristics, the test is of superior quality. With data on the difficulty and discriminative value of the items and with tabulations of the responses to the individual distracters, it was possible to select items for still another form of the test which embodies only the best of the AC items and the best of the distracters to accompany them. This new test is the *Arithmetical Computation Test, Form 1*. It was designed on the basis of the information gained from the analysis of two experimental forms of the *Arithmetical Computation Test.*

The items for this test were selected on a multiple criterion:

1. In general, an attempt was made to select those items which bore the highest correlation (r_{bis}) with the total score on the test.

2. Items were selected in such a way that their difficulties were distributed over a wide range (from about .90 to the level of chance), with the majority of the items falling in the middle part of this range. In some instances, where there was a surplus of items at a particular level of difficulty, an attempt was made to thin out the concentration.

3. Before any of these tests were constructed, a decision had been reached as to the types of arithmetical operations which should be represented in the test; as far as possible an attempt was made to preserve this representation in the selection of the items for Form 1.

To illustrate how the criteria outlined above were applied, Figure 1 is presented. In this figure, each of the items of Form X-2 is positioned with respect to its difficulty and its discriminative value. The white circles represent the items selected for Form 1 and the black, the rejected items.

An inspection of Figure 1 will reveal several characteristics of the selected items. Probably the most outstanding of these characteristics is the uniformly high r_{bis} of the Form 1 items—only one of the items has an r_{bis} of less than .60. The plot also illustrates that discriminative value was not the only basis upon which items were evaluated. In several cases, the item with the lower r_{bis} was retained in preference to other items

with higher values; such a procedure was necessary on several occasions in order to maintain the agreed-upon content of the test. The items represent all difficulty levels from .29 to .87; on the ground that at least one very easy introductory item should be used, an item which had been used in the earlier Form X-1 was reincorporated. The general effect of selection was to increase both the difficulty of the items and their average discriminative value. The median r_{bis} of accepted items is .74—much higher than the average r-values of items which are used in making the typical test. The median p-value is .58, which meets the criterion of being approximately halfway between the level of chance success (.20) and perfect performance.

In spite of the fact that Form 1 includes only three-fifths of the original X-2 items, the proportions devoted to each of the arithmetical operations of the initial form are changed but little. In general those problems which dealt with decimals, or with fractions and decimals in combination, performed most poorly and as a consequence their representation is reduced the most.

A decision having been reached as to which of the fifty X-2 items should be incorporated into Form 1, a similar decision had to be made concerning the selection of the distracters which were to accompany the selected items. The item analysis of Form X-2 provided two types of data which were used in making this selection: (1) a count of the number of persons choosing each distracter and (2) their average (mean) total score on the test. Choice of distracters was thus governed by two major considerations:

1. In general, the distracters selected for the final form were the popular ones. Obviously there is little to be gained by using distracters which fail to "distract," i.e., are chosen by only a few persons.

2. The selected distracters discriminated between those persons who made "high" scores and those who made "low" ones. The mean scores of those who chose the correct answer were compared with the means of those who chose the incorrect ones. Distracters with the lowest mean scores were chosen to represent the incorrect responses; those with means which approached the mean of the "correct" group were discarded.

By following this procedure, the discriminative value of the item as a whole is improved.

Five alternatives were chosen for each of the selected thirty items. For each of the items, one of these five is "Answer not given"; the item-analysis statistics derived from Form X-2 show that this particular type of alternative functions very well both as the "correct" as well as the "incorrect" answer.

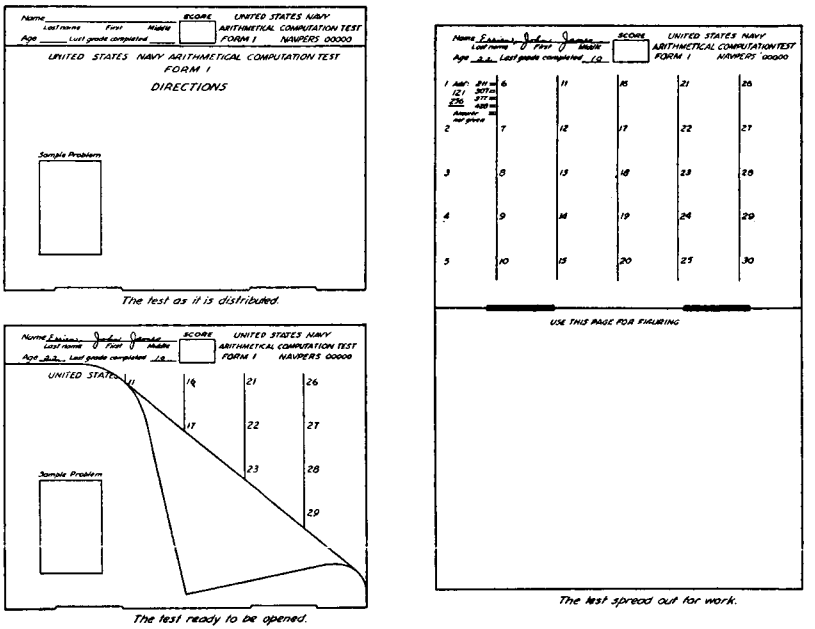


FIG. 3 Illustration of Form 1 Printed on an IBM Answer Sheet

A model layout of this test was prepared and submitted to the Bureau of Naval Personnel. The thirty items can be printed on one side of a standard IBM answer sheet (see Figure 3). Directions for taking the test can be printed on a cover page; the back of this page provides "figuring" space.

The Validation of the Arithmetical Computation Test

The question of the validity of the *Arithmetical Computation Test* remains to be answered. One of the most commonly employed procedures involves the following steps: (1) administering the test to a group of service school candidates; (2) waiting

until the men have completed their service school training; and then (3) correlating test scores with final achievement (grades) in the school. Obviously, this procedure demands that one wait as long as three or four months before it becomes possible to make a statement concerning the validity of the test in question.

An alternative procedure makes an earlier decision possible. If it can be assumed that training does not have a *differential* effect on the AC distribution, then the validity coefficients derived from correlating test scores of graduating classes with final grades should be essentially the same as those derived from classes taking the tests at the time of their entrance. One way of evaluating the effect of training on AC test scores is to compare the regressions of the AC on the GCT and AR for entering and for graduating classes. If there is no significant difference in these two groups of regression coefficients, there is some basis for assuming that the validities derived from graduating classes will be the same as those obtained from entering classes—this in spite of the fact that in the case of graduating classes the AC is given at the end of the service school course rather than at the beginning. This was the assumption under which the following steps were carried out.

1. The *Arithmetical Computation Test, Form X-2* was administered to both entering and graduating classes at various types of service schools where it seemed that the *Arithmetical Computation Test* might be a valid selection measure.

2. Final service school grades were obtained for the graduating classes. At the same time, the Basic Battery test scores of both entering and graduating men were also obtained.

3. Basic Battery and AC test scores were correlated with the final service school grades of the graduating classes.

4. Regression coefficients (b's) of the AC on the GCT and AR, of the GCT on the AR, and of the AR on the GCT were computed. The differences in these values (b's) for entering and graduating classes were evaluated to determine whether or not the assumption underlying the validation procedure was met.

5. Finally, provision has been made for correlating AC scores of the entering classes with service school grades at graduation.

The five steps of the procedure outlined above were completed for five schools at the Naval Training Center at Bainbridge, and for five schools located at Great Lakes. In addition, validity coefficients were computed for classes from three other schools. The N's of the schools used range from 48 to 214 (see Table 5).

The validity coefficients obtained by correlating test scores and service school grades for the classes tested at graduation are shown in Table 5 (in the columns headed *Grad.*). The twelve coefficients for the AC test range from .33 to .69.[6] In six of the ten schools where comparisons can be made, the AC has a validity which is as high as or higher than for any other single Basic Battery test. AC has a higher validity than AR in seven of the ten schools. If median validity is considered, the validity of the AC outranks those of the tests of the Basic Battery; in this connection, however, it should be remembered that the schools which were chosen for this study were those where the AC was expected to perform best.

Can these validity coefficients be taken to represent the coefficients which would have been obtained if the men tested prior to school training had been used? The regression coefficients (*b*'s) of AC on GCT and on AR for entering and graduating classes must be compared in order to predict the results which would have been obtained if AC were administered prior to entry in the schools. Since complete data at the time were available for only nine of the service school groups, the regressions could be compared for only those nine schools. The regressions of AR on GCT and of GCT on AR for both entering and graduating classes were also computed. Since the latter two tests were both administered prior to entry in the service school, they served as an indication of the difference in regressions which might be expected as a result of sampling differences between the two classes.

None of the differences between regression coefficients of the AC on the AR or GCT for entering and graduating classes approached statistical significance. The regression of AC on AR

[6] An estimate of what these coefficients would be for a test which is shortened to thirty items is supplied by a formula given by Guilford (3, p. 422). Application of this formula reduces the lowest AC coefficient from .33 to .31 and the highest from .69 to .65.

TABLE 5

Comparison of Validity Coefficients

Service School*	Center	Arithmetical Computation			General Classification			Reading			Arithmetical Reasoning			Mechanical Aptitude			Mechanical Knowledge (Mechanical)			Mechanical Knowledge (Electrical)			N	
		Grad.	Ent.	Diff.	Grad.	Ent.	Diff.	Grad.	Ent.	Diff.	Grad.	Ent.	Diff.	Grad.	Ent.	Diff.	Grad.	Ent.	Diff.	Grad.	Ent.	Diff.	Grad.	Ent.
AM	N.P.†	.33	.44	—	—	.36	—	—	.27	—	—	.42	—	—	.55	—	—	.62	—	—	.59	—	110	92
AMM	N.P.†	.53	.47	—	—	.33	—	—	.37	—	—	.46	—	—	.47	—	—	.52	—	—	.48	—	99	130
BEng	G.L.	.69	.43	−.26	.54	.31	−.23	.44	.30	−.14	.52	.38	−.14	.50	.50	.00	.60	.41	−.19	.51	.45	−.06	214	188
EM	Bain.	.50	.48	−.02	.01	.45	.44	.14	.48	.34	.26	.51	.25	.18	.34	.16	.24	.28	.04	.29	.46	.17	67	199
	G.L.	.41	.59	.18	.45	.55	.10	.50	.42	−.08	.35	.57	.22	.28	.42	.14	.23	.43	.20	.28	.53	.25	88	77
FC	Bain.	.38	.38	.00	.46	.30	−.16	.51	.28	−.23	.42	.35	−.07	.36	.32	−.04	.24	.17	−.07	.12	.24	.12	125	117
	G.L.	.39	.29	−.10	.35	.25	−.10	.39	.34	−.05	.27	.34	.07	.25	.22	−.03	.08	.04	−.04	.01	.29	.28	73	67
QM	Bain.	.33	.39	.06	.21	.46	.25	.09	.43	.34	.16	.46	.30	.20	.12	−.08	−.00	.22	.22	.02	.16	.14	48	137
	G.L.	.42	.49	.07	.27	.37	.10	.22	.36	.14	.29	.53	.24	.24	.36	.12	−.02	.18	.20	.18	.13	−.05	63	100
SC & Bkr	Bain.	.34	.40	.06	.11	.11	.00	.13	.21	.08	.28	.20	−.08	.10	.05	−.05	.18	−.02	−.20	.23	.18	−.05	84	72
	G.L.	—	.39	—	—	.54	—	—	.40	—	—	.33	—	—	.12	—	—	−.06	—	—	.25	—	—	104
SK	Bain.	.39	.37	−.02	.38	.44	.06	.23	.37	.14	.50	.37	−.13	.20	.22	.02	.35	.15	−.20	.37	.24	−.13	115	149
	G.L.	.40	.44	.04	.50	.16	−.34	.43	.26	−.17	.46	.43	−.03	.35	.08	−.27	.17	.18	.01	.21	.23	.02	111	63
Mean diff.				.001			.012			.037			.063			−.003			−.003			.069		
Mean absolute diff				.081			.178			.171			.153			.091			.137			.127		
Mdn. of all r's		.395	.430		.365	.360		.310	.360		.320	.420		.245	.320		.205	.180		.220	.250			

* The schools are as follows: AM, Aviation Metalsmith; AMM, Aviation Machinist's Mate; BEng, Basic Engineering; EM, Electrician's Mate; FC, Fire Controlman; QM, Quartermaster; SC & Bkr, Ship's Cook and Baker; SK, Storekeeper.

† Navy Pier, Chicago.

was found to be very similar for entering and graduating classes in the case of each of the nine schools. The regressions of AC on GCT differed to a somewhat greater extent, especially for Electrician's Mate schools. In most cases the effect of training in service school apparently was to increase the AC test scores, especially for those men with low GCT and AR scores; for Electrician's Mate schools the increase was especially marked for men with low GCT scores. The differences between entering and graduating classes with respect to regressions of AC on GCT and AR may be compared with similar regressions of GCT on AR, and AR on GCT. On the average the difference is smaller for regressions involving AC than for regressions of AR on GCT or of GCT on AR, in spite of the fact that GCT and AR were both administered prior to entry in a service school, while AC was administered in one case before and in another case following the service school course. In view of the similarity of the regressions of AC on GCT and on AR for entering and graduating classes, it was expected that the validity coefficients obtained for the classes tested at entry would be approximately the same as for the classes tested at graduation.

After graduation of the classes to which the AC test was given at the time of entry and computation of validity coefficients for these classes, it became possible to compare the validity coefficients for classes tested at graduation with those for classes tested at entry; this furnishes an evaluation of the dependability of the time-saving method of studying validity.

The statistics to be compared are presented in Table 5. Validity coefficients obtained from classes to which the AC test was administered at the time of graduation are shown in the columns headed *Grad.*; validity coefficients for classes to which the test was given at the time of entry are in the columns headed *Ent.* (The corresponding means and standard deviations are reported in 2.) The differences between the pairs of validity coefficients are given in the columns headed *Diff.*

The AC validity coefficients for classes tested at entry and at graduation differ by no more than .10, except for two classes. The mean of the differences is only .001. The validity coefficients for most of the Basic Battery tests differ somewhat more; the mean differences range from .012 to .069, although

the time of test administration was not a variable for the Basic Battery tests. The mean difference (disregarding signs) is smaller in the case of AC than for any Basic Battery test, the average magnitude of the difference being .08 for AC and varying from. 09 to .18 for the other tests. It may be concluded that the differences in AC validity between classes tested at graduation and classes tested at entry are due to differences between the samples and not to the influence of training on the AC score distributions.

Median validities for each test are reported in the last row of Table 5. The statistics obtained from graduating classes are still pertinent. On the basis of these median values, it can be stated that the AC is as good as or better than AR, and the AC is superior to all the Basic Battery tests except AR for predicting grades in the types of service schools studied. The AC possesses the additional advantages of higher reliability than AR and shorter time for administration.

Conclusions

1. The tabulation of free-answer responses to test items and the analysis of types of error provide a useful method for obtaining efficient distracters for arithmetical computation test items.

2. Shifts in item-difficulty from free-answer form to multiple-choice form are relatively small, when the multiple-choice distracters are chosen on the basis of tabulations of free-answer responses and analysis of types of error.

3. Item-analysis data which include frequency of choice of each response and mean score of those choosing each response constitute a useful basis for choosing among a surplus of distracters in a trial form of a test.

4. "Answer not given" functions well, both as a correct answer and as a distracter, when consistently used as a response to multiple-choice arithmetical computation items.

5. In a heterogeneous group, a substantial correlation is found between arithmetical computation test scores and scores on a measure of verbal ability, even when efforts are made to reduce the verbal component of the arithmetic test to a minimum.

6. A workable plan was developed to print a 30-item multiple-choice arithmetical computation test on one side of an IBM answer sheet which could be machine scored, eliminating the necessity for a separate answer sheet.

7. A method of validating a test which does not require waiting for criterion data to become available was developed. This method involves administering the test at the time of graduation rather than at the time of entrance. If evidence can be obtained that training does not have a differential effect on the test scores, it can be assumed that the validities obtained from entering classes will be about the same as those for graduating classes. Satisfactory evidence can be provided by showing that the regressions of the test on other measures of ability known to be related to the criterion are about the same for groups tested at entrance and at the time of graduation.

REFERENCES

1. Conrad, H. S. *A Statistical Evaluation of the Basic Classification Test Battery (Form 1)*. (OSRD, 1945; Publ. Bd., No. 13294.) Washington, D. C.: U. S. Department of Commerce, 1946.
2. Frederiksen, Norman. *A Further Study of the Validity of the Arithmetical Computation Test*. (OSRD, 1945; Publ. Bd., No. 13306.) Washington, D. C.: U. S. Department of Commerce, 1946.
3. Guilford, J. P. *Psychometric Methods*. New York: McGraw-Hill, 1936.
4. Peterson, D. A. *Factor Analysis of the New United States Navy Basic Classification Test Battery*. (OSRD, 1943; Publ. Bd., No. 13317.) Washington, D. C.: U. S. Department of Commerce, 1946.
5. Satter, G. A., and Frederiksen, Norman. *The Construction and Validation of an Arithmetical Computation Test*. (OSRD Report No. 4556.) Princeton, N. J.: College Entrance Examination Board, 1945. (Restricted.)