RELATIVE VALUE OF DIFFERENT TYPES OF QUESTIONS IN READING TESTS

STERLING G. BRINKLEY, ILLINOIS COLLEGE

The Problem.

The subject of this study is narrower than is indicated by the title given it. "Reading tests" are limited to silent reading tests of the type of the Thorndike-McCall Reading Scale, designed to measure depth of power of comprehension in reading. Only three "types of questions" are considered: the true-false, the multiple choice or selective response, and the word or phrase answer type. The comparison of "value" is limited to one point, that of validity or the extent to which the test measures the ability it is supposed to measure. The question studied may then be more accurately stated as follows: What is the relative validity in a silent reading test of the kind indicated of true-false questions, multiple choice questions, and word or phrase answer questions?

This study of the types of questions for reading tests is a phase of a larger problem undertaken two years ago. When Professor W. A. McCall of Columbia University was serving as Director of Psychological Research for the Chinese National Association for the Advancement of Education, I undertook in coöperation with this organization the construction of an English reading test. As previous use of the Thorndike-McCall Reading Scale in Chinese schools had demonstrated its value for measuring reading and general English ability where English was a foreign language, my procedure was to adapt this test to the situation. In doing this all poetry was omitted, new reading selections were introduced, all duplication of selections in parallel forms of the test was eliminated, the vocabulary was zoned on the basis of Thorndike's Teachers Word Book, and idioms likely to cause unusual difficulty to the foreign student were changed. In order that the test might involve only reading, it was decided to construct questions of the multiple choice type. Thus the written answer was avoided. Four forms of the reading test were constructed and standardized for Chinese schools.

The problem of the validity of the multiple choice type of test was raised by this larger study. Data bearing on its solution were gathered, but it is only recently that there has been opportunity to go back to the original data collected and complete the study.

The Tests Used in the Study.

The following sample taken from Form I will illustrate the types of questions compared. The paragraph chosen is one of the easier paragraphs, containing only words from Thorndike's first 500 in frequency of occurrence.

Once there was a rich man who got up early one morning and placed a great stone in the road. He wanted to see what people would do when they saw it. That day he saw men, women, and children pass by. Each one looked at the stone and then walked or drove around it. At last, just at dark, a boy came along. He said to himself: "Someone may fall over this stone. I will try to move it." He worked and worked until he moved it away. Then he saw a box which had been lying under the stone. When he opened the box, he found that it was filled with gold.

Type A questions, (Multiple Choice, from Form 2A).

1. The stone was placed in the road (a) by the rich man, (b) by a boy, (c) by a group of people, (d) by some children.
2. The box of money was found (a) under the stone, (b) in a house, (c) by the side of the road, (d) in a hole in the ground.
3. Why did the boy move the stone? (a) Because he thought some one might fall over it. (b) Because he wished to see if he could move it. (c) Because the people would all know about it. (d) Because he wished to find some money.
4. Most of the people who passed along this road (a) tried to move the stone, (b) moved the stone, (c) fell over the stone, (d) went around the stone.

Type B questions, (Word or Phrase Answer, from Form 2B).

1. Who placed the stone in the road?
2. Where was the box of money found?
3. Why did the boy move the stone?
4. What did the men do when they saw the stone in the road?

Type C questions. (True-False, from Form 2C).

1. Was the stone placed in the road by the boy?
2. Was the box of money found under the stone?
3. Did the boy move the stone away?
4. Did the people who passed along this road fall over the stone?

It is evident from the reading of the different types of questions that for any given paragraph the questions of one type cover the same facts as are covered by the questions of the other two types. This is true of the tests as a whole. The paragraphs of Form 1 were thus used in three tests, Type B having the word or phrase answer questions, Type A having the multiple choice questions, and Type C having the true-false questions.

Similarly, there were the three types of questions grouped in three separate tests for Form 2 and for Form 3.

Forms 1A, 2A, 3A (that is, Forms 1, 2, and 3 of the reading test in each of which the Type A, or multiple choice questions, were used) were of equal difficulty, as determined by results got when they were administered to a group of 100 pupils. The arithmetical means were respectively 25.6, 24.9, 24.1. It was assumed that this same approximation of equality between the forms existed in the case of the Type B and Type C questions also. The three forms with the Type A questions had a length of 50 minutes (the average time used by the pupils in taking the test). The length in minutes for the Type B and Type C tests was not determined. The number of items was, of course, the same as in the Type A tests.

The Criterion.

The study undertook to determine the relative validity of the three types of questions. Validity is determined by comparison with a criterion. Two types of criteria were possible. (1) The first was a composite of the scores on the three reading tests which were being compared. The advantage of this criterion is that it is a criterion of *reading ability;* its disadvantage lies in the fact that spurious correlation is present due to the inclusion in the criterion of the test studied. (2) The second was a measure of *general ability* in English. The English department gave at the close of the semester six standard tests covering varying phases of English ability. These were: Thorndike-McCall Reading Scale, Kelley-Trabue Completion Test, a spelling test made up from Ayres Spelling Scale, and three tests of the English Mastery Series published in Shanghai, i. e., Grammar-Idiom, Word Opposites, and Auditory Comprehension. A composite of the results on these six tests was made, by giving to spelling and auditory comprehension each a weight of one-half while each of the other tests received a weight of one. This composite was, it is believed, a very accurate measure of general ability in English.

A previous study[1] had shown fairly high correlation between ability in reading as measured by the Thorndike-McCall Reading Scale and general English ability. This fact was taken to justify the use of the composite as criterion, while the fact that none of the tests being compared was directly included in the composite led to its acceptance in preference to the more specific reading criterion.

The Subjects.

The subjects in the experiment were the pupils of the Soochow University Middle School[2]. Only those who took all the tests were included in the final computations. The number was 226. The school is a college preparatory school for boys. The course of study covers seven years. English is begun the first year and is compulsory throughout the course. For the purpose of the experiment the pupils were divided into two major groups, Group I and Group II. Group I was composed of the pupils who had taken the original forms of the reading test, Type A, in April, when it was administered for purposes of try-out and scaling. The remainder of the pupils of the school made up Group II.

The pupils of Group I were divided into two sub-groups, designated as V and W. The division was carried out by the method of pairing so as to render the groups of equal ability in English. The pupils of Group II were divided into three sub-groups in the same way. These groups were designated as X, Y, and Z. That the sub-groups within each major group were closely similar groups is seen by the fact that the averages and standard deviations of the scores made by these groups on the criterion composite are almost identical. (Table I.)

1Brinkley, The use of American tests to measure English teaching in China, *Journal of Educational Research*, Sept. 1923, Vol. 8, No. 2

2Located at Soochow, China

TABLE I.

AVERAGES AND STANDARD DEVIATIONS.

	No. of Pupils	Citerion Mean	Citerion S. D.	Type B Mean	Type B S. D.	Type A Mean	Type A S. D.	Type C Mean	Type C S. D.
Group I.									
Sub-Group V	50	104.7	25.9	28.3	10.4	25.0	12.1	28.3	12.1
Sub-Group W	49	104.5	26.5	32.1	11.9	25.9	12.1	28.1	12.3
Group II.									
Sub-Group X	40	104.2	16.9	23.2	6.3	27.9	8.2	24.2	7.0
Sub-Group Y	44	104.6	17.0	29.4	6.3	23.6	7.1	24.9	8.9
Sub-Group Z	43	104.1	16.7	27.5	7.0	25.0	7.0	26.8	7.8
Arith. Mean (unweighted)				28.1		25.5		26.5	

NOTE.—The pupils in each Group were drawn from all the seven classes in the school, except that in Group II. there were no pupils from either the lowest or the highest class.

Administration of the Tests.

With one exception the tests were administered at the close of the semester in June, 1923. The exception was in the case on Group I which had taken the Type A tests in April. In June this group was given Form 1B and Form 1C, and the scores made on Form 1A in April were compared with these. The two tests in June were given during one two-hour period. The order of giving the two tests was rotated in order to eliminate the probable practice effect due to the identity of content in the tests and any fatigue effect due to the taking of the two tests at one sitting.

To Group II were administered three tests on two consecutive days, two being given at a two-hour period on the first day. Any specific practice effect was eliminated with this group by using all three forms of the test, making it possible to have different content on each occasion. Thus Sub-group X took as its three tests Form 1A, Form 2B, and Form 3C. In this way no group of pupils had the same paragraphs twice and no two groups of pupils had the same questions. Moreover, there was rotation of types so that there might be eliminated whatever advantage or disadvantage was

present due to the time and order in which the tests were taken. Thus Type A was administered first to Group Z, second to Group Y, and third to Group X. For the first test Group X had Form 2B, Group Y had Form 2C, and Group Z had Form 2A. These illustrations should make clear the type of rotation used.

To sum up: Group I had the same content in all three of the tests taken. For this group the variable factors were type of questions and time and order of taking the tests. The rotation in administering the tests was such as to give the advantage of practice now to one type and now to another. Group II had different content in each of the tests taken, but the difference was only that between parallel forms of the same test; the rotation in administration here served to balance advantages of position and order of tests in the series. Correlations were computed for each of the five groups separately.

As a result of this method of administering the tests, the three types of questions are compared under the following conditions: (1) When the content is the same and the pupils are the same, but the time of taking the test varies (Table II, Group I.); (2) When the content is the same and the time of taking the test is the same, but the pupils are equal groups rather than the same pupils (Table III.); (3) When the pupils are the same, but the content instead of being the same is equivalent (parallel forms of the test), and the time of taking the test varies (Table II, Group II.). It would seem that differences between the tests shown after all these comparisons are combined must be due to the type of question used.

Results of the Study.

The results of the comparisons are shown in Tables II and III. The word or phrase answer test gives in every

TABLE II.

CORRELATIONS WITH CRITERION.

(Pupils, the same: Group I. Content, identical; Group II. Content, equivalent forms).

	Correlations			P. E. of Correlation			No. of
	Type B Word or Phrase Ans.	Type A Multiple Choice	Type C True-False	Type B	Type A	Type C	Pupils
Group I.							
Sub-Group V.	.92	.91	.88	.015	.016	.022	50
Sub-Group W.	.94	.92	.86	.011	.015	.025	49
Group II.							
Sub-Group X.	.84	.80	.67	.032	.037	.059	40
Sub-Group Y.	.89	.83	.85	.021	.032	.028	44
Sub-Group Z.	.84	.78	.82	.030	.040	.034	43
Ar. Mean (unweighted) ...	.89	.85	.82				

TABLE III.

CORRELATIONS WITH CRITERION.

(Content identical; equivalent groups; time and order of taking tests identical.)

	Type B Word or Phrase Ans.	Type A Multiple Choice	Type C True-False
Group I.			
Form 1.....	.94	.92	.88
Form 1.....	.92	.91	.86
Group II.			
Form 1.....	.89	.80	.82
Form 2.....	.84	.78	.85
Form 3.....	.84	.83	.67

NOTE.—The data in Table III are, of course, the same as those in Table II. The separate comparisons made are, however, different.

case except one the highest validity coefficient. The multiple choice test gives a higher coefficient than the true-false test in six out of the ten comparisons and a lower validity in the other four. The average (Table II.) is in favor of the multiple choice test. The differences revealed by the study are not great, whether considered separately or in the average. The data are not of such nature as to make possible the mathematical determina-

tion of the significance of these differences. Since, however, the word or phrase answer test is superior to the other two in every comparison except one, and practically equal there (Table III.), the average difference of .04 between it and the multiple choice test and of .07 between it and the true-false test is more likely to be a *real* difference than that of .03 between the multiple choice test and the true-false test.

The data may be interpreted, therefore, to indicate a slight but real advantage in validity of the word or phrase answer test over the other two and no essential difference in validity between the multiple choice and the true-false tests. These conclusions based on a study of *reading tests* are in agreement with the conclusions of a more thorough study of types of tests in the field of *history* made by the writer[1].

[1]Brinkley, S. G., *Values of New Type Examinations in the High School, with Special Reference to History.* Teachers College, Columbia University, Contributions to Education, No. 161.