THE VALIDITY OF ARITHMETICAL-REASONING TESTS

R. V. Hunkins and F. S. Breed, University of Chicago

The primary purpose of this investigation was to throw light on the relative validity of several arithmetical-reasoning tests now in use in the public schools. A secondary purpose was to explain in some measure the different degrees of merit shown by the different tests.

The following tests were used in the experiment:

(1) Daniel Starch: Arithmetical Scale A.
(2) C. W. Stone: Reasoning Test.
(3) W. S. Monroe: Standardized Reasoning Test in Arithmetic, Form 1.
(4) B. R. Buckingham: Scale for Problems in Arithmetic, Form 1.
(5) E. H. Chapple, S. A. Courtis, F. R. Matthews: Arithmetic Tests—Reasoning.
(6) R. M. Yerkes, M. E. Haggerty, L. M. Terman, E. L. Thorndike, G. M. Whipple: National Intelligence Tests, Scale A, Form 1, Test 1.
(7) M. E. Haggerty: Intelligence Examination, Delta 2, Exercise 2.
(8) A. S. Otis: Group Intelligence Scale, Advanced Examination, Form B, Test 5.
(9) L. M. Terman: Group Test of Mental Ability, Form A, Test 5.
(10) A. S. Otis: Group Intelligence Scale, Advanced Examination, Form B (complete).

For special reasons the results from tests (5) and (9) were not included for further study.

The subjects to whom the tests were administered were pupils in grades five to eight, inclusive, of a Hot Springs, S. D., public school. Complete data on all tests were secured from 127 subjects.

In advance of experimentation, a plan of administering the tests was devised to control such variables as time of day, day of week, and order of administration. All tests were administered and scored personally by Mr. Hunkins.

Each reasoning test provided a single measure of the arithmetical-reasoning ability of each pupil. On the assumption that the average of several expert attempts to measure an object or reaction is more reliable than a single attempt, the average score of a pupil from several tests was regarded as approximating more closely the true measure of the ability in question than the measurement derived from a single test.

The average score was obtained after each test series had been transmuted into values on a percentile scale and thus made comparable. The method of transmutation was shown to have satisfactory reliability by the high correlation between each original series of scores and the corresponding derived series. The average coefficient (product-moment) was .99. The five tests which have but one set of problems for the four grades in which the experiment was conducted were employed in deriving the average or composite score.

The average coefficient of correlation of each test series with the composite in each grade was found, and the several tests ranked according to closeness of agreement with the composite.

For the purpose of verifying the results obtained by the use of the composite scores, all the possible inter-test correlation coefficients were computed for the sixth and seventh grades. The method of composite scores showed which test agreed most closely with the average result of five of the tests. The method of inter-test correlation showed which test on the average agreed most closely with the rest of the tests. The results by the two methods should be closely similar.

Table I presents the ranking of the tests according to the two different methods of estimating their validity.

TABLE I

RANKING OF THE TESTS

Test	Method of Composite Scores	Method of Inter-test Correlation	Combined Ranks	Final Rank
National	5	4	9	4
Haggerty	2	3	5	2.5
Otis	3	2	5	2.5
Starch	4	6.5	10.5	5
Stone	1	1	2	1
Monroe	6	5	11	6
Buckingham	7	6.5	13.5	7

It should be remembered that in determining the composite score only five of the tests were used, Monroe and Buckingham being omitted. In determining the inter-test correlations, all the tests were used. It will be noted that the results by the two methods differ but slightly. It is important, however, to consider the inter-test values side by side with the other series, primarily because of the above omissions in the computation of the composite. Inclusion of a test in deriving the composite naturally tends to improve the rating of the test by the composite

The two series of ranks appearing in Table I were combined to form a final series representing the net result of the investigation. This final series is shown in the column designated "Final Rank." It will be observed that the Stone test occupies first position, the Haggerty and Otis tests divide honors for second place, the National ranks fourth, Starch fifth, Monroe sixth, and Buckingham seventh.

Analysis of the method and content of the tests revealed marked variations with respect to (1) length of time allowance, (2) selection of problems for different grades, (3) space for computation, (4) weighting of problems, (5) method of scoring, (6) use of preliminary exercises, (7) kind of problems. Most of these variables need further investigation in order to lay the foundation for more accurate tests of arithmetical-reasoning ability.