

DEVELOPMENT AND NORMING OF READING TESTS FOR AIR FORCE USE

I. INTRODUCTION

Many Air Force organizations have been administering various commercially published reading tests to military personnel. These tests have been used for assignment to remedial training programs, as aids in counseling students, or for description of reading levels of airmen in various occupational specialties. The Tests of Adult Basic Education (TABE) (CTB/McGraw-Hill, 1976) is the reading test instrument most frequently used in the Air Force.

A study on service applicants (Mathews, Valentine, & Sellman, 1978) found considerably divergent reading grade levels (RGL) from different commercial tests for subjects of the same Armed Services Vocational Aptitude Battery (ASVAB) ability level. The ASVAB General (called General-Technical for some services) composite correlated as highly with some reading tests as those reading tests correlated with each other. This suggests that ASVAB can be used to estimate reading ability of groups. Some problems occur in using ASVAB composites to measure reading ability of individuals, however. These composites contain several short subtests covering different ability factors. The General composite includes Arithmetic Reasoning (AR), in addition to the verbal subtests of Word Knowledge (WK) and Paragraph Comprehension (PC). While most women perform slightly better than men on verbal tests, they generally do somewhat less well than men on AR. When the General composite is used to gauge reading ability of women, underestimation

will result in the majority of cases. For individual measurement, therefore, a more content specific and reliable measure of reading than that based on ASVAB appears warranted.

The use of commercial tests has several drawbacks, including high testing material costs and RGL norms of unknown appropriateness for military personnel. The objective of this report is to provide a description of the development and norming of forms of the Air Force Reading Abilities Test (AFRAT) which have been designed as replacements for TABE and other reading tests for Air Force use.

II. METHOD

Development of AFRAT Forms

The following general goals were pursued in developing reading tests:

- a. Vocabulary and comprehension sections as found in TABE and most other reading tests used by the Air Force were to be designed.
- b. Comprehension passages were written with expository prose.
- c. Comprehension questions covered factual matter which was unlikely to be answered correctly based solely on prior knowledge.
- d. Vocabulary words were selected which might likely be encountered in a work environment. Esoteric adverbs and adjectives were avoided to keep the test from being overly academic in nature.

e. The test was designed to be as reliable as possible but require less than an hour of testing time.

Reading Measurement Instruments

The following reading tests were used in this study.

AFRAT Form X. An experimental form of AFRAT was constructed based on available items from obsolete Air Force classification tests. This test was used to obtain initial estimation of the construct and predictive validity of the item types. Due to the limited pool of items, the difficulty of AFRAT X items varied considerably from very easy to very hard.

AFRAT Forms A and B. Two parallel AFRAT forms were developed. The second form allows for retesting after remedial training. Items were selected from a pool assembled specifically to specifications for AFRAT. The tests were targeted at an 8th reading grade level (RGL) as measured by the Adult Basic Learning Examination (see the following paragraphs). While AFRAT A and B were to be peaked at a difficulty level corresponding to the 8th RGL, norms were desired which would span from the 5th through the 12th RGL.

Gates-MacGinitie Reading Tests (Survey D). This test gives vocabulary and comprehension RGLs from 1 to 11.9 (Gates & MacGinitie, 1972).

Tests of Adult Basic Education (TABE) Level D. TABE yields Vocabulary and Comprehension RGLs from 5 to 12.9 (CTB/McGraw-Hill, 1976).

Adult Basic Learning Examination (ABLE). Vocabulary, reading, and problem solving sections were used to calibrate the ASVAB General Composite to ABLE in a 1980 unpublished Army study. ABLE gives RGLs from 3 to 12.9 (Karlsen, Madden, & Gardner, 1971).

Samples

A total of 12,983 subjects tested from May to July 1981, except as noted, formed the following samples:

1. 625 Air Force trainees administered AFRAT A and B.
2. 820 Air Force trainees given ABLE II and AFRAT A (N = 413) or B (N = 407).
3. 946 Air Force trainees given Gates-MacGinitie and AFRAT A (N = 454) or B (N = 492).
4. 883 Air Force trainees given AFRAT X and AFRAT A (N = 459) or B (N = 424).
5. 3,274 Air Force trainees given TABE and AFRAT A (N = 1951, composed of subjects from samples 1-4) or AFRAT B (N = 1948, composed of subjects from samples 1-4 less 625 subjects given both forms).
6. 1,049 Army trainees given AFRAT A (N = 500) or AFRAT B (N = 549).

7. 2,253 Air Force trainees given AFRAT X in 1978.

In addition, data based on about 1,100 Army trainees given ABLE I, II, or III in 1980 and 2,033 service applicants given Gates-MacGinitie in 1978 (Mathews, Valentine, & Sellman, 1978) were used in developing norms. These two tests and the TABE are widely used by the armed services.

Statistics

An item analysis program (Koplyay, 1981) was used to compute internal AFRAT and test summary statistics. These include difficulty (proportion answering each item correctly), item biserial (correlation of item with test scale), internal consistency reliability (Kuder-Richardson Formula 20), test mean, and standard deviation. Means for Army samples were also adjusted for sampling differences. This was accomplished by using the regression equations (Guilford & Fruchter, 1978) for predicting AFRAT scores based on the relationship of AFRAT forms with the General Composite.

Construct and predictive validities of AFRAT forms were assessed through Pearson correlation coefficients (r 's), which were computed among tests and between AFRAT Form X scores and technical training grades for subsamples. Fisher's r to z transformations were used to average r 's, across combined samples (Guilford & Fruchter, 1978). The technical training validation was only a preliminary analysis as a more comprehensive study will be done on AFRAT forms A and B when sufficient data are available.

Percentile norms were obtained for AFRAT forms, and AFRAT Forms A and B were placed on the same scale through equipercentile equating (Angoff,

1971). This same procedure was used to equate AFRAT to TABE Reading Grade Levels (RGL). AFRAT Forms A and B were also equated to ABLE and Gates-MacGinitie RGL scales through the use of the ASVAB General composite as an anchor test.

III. RESULTS AND DISCUSSION

AFRAT Internal Analyses

The AFRAT consists of vocabulary items in a synonym format and comprehension items consisting of one or several paragraphs followed by one or more questions. The comprehension items require either paraphrasing or making inferences from the passages. AFRAT A and B each contain 45 vocabulary and 40 comprehension items with a total test limit of 50 minutes (see Table 1). All items are multiple choice with four alternatives.

Table 2 gives the item difficulties for AFRAT A and B based on Air Force trainees given both tests (Sample 1). These alternate forms appear to be of parallel difficulty, with fairly similar means and distributions. The bulk of the items are quite easy with means around .85 (not corrected for guessing). In comparison, the TABE items had an average difficulty of .84 for the same sample (N = 625).

The item-test biserial correlations (r_{bis}) are moderate-to-high for virtually all items, with means and medians of the r 's of about .60 and a range of .29 to .89. Again, the AFRAT forms seem parallel (see Table 3).

An estimate of mean AFRAT item performance for subjects equal in average ability to that of the normative population for ASVAB can be obtained from the data collected on Army trainees (Sample 6). Army samples given AFRAT had an average ASVAB General composite score of about 50 percentile, 50.6 for AFRAT A sample and 49.7 for AFRAT B sample. Mean AFRAT difficulty (\underline{p}) for these subjects (sample 6) are given in Table 4. Because the lowest ability subjects are excluded from service, the distribution of scores would differ in an applicant or normative sample. The average \underline{p} was .69 for the Army samples compared to the \underline{p} of .85 for Air Force samples. Since ASVAB selection tests have \underline{p} 's of about .70 (Ree, Mullins, Mathews, & Massey, 1981), AFRAT seems to be comparable in mean difficulty to these tests.

AFRAT internal consistency reliability coefficients are shown in Table 5 for subgroups of Air Force samples. These data are based on all female and all Black trainees, and representative subsamples of male and Caucasian trainees. The average reliabilities were .92 for AFRAT A and .91 for AFRAT B. These values are quite high considering that reliability is maximized when moderate item difficulties maximize variance.

Reliabilities were not as high for female samples, .89 for AFRAT A and .87 for AFRAT B. This is most likely due to significantly lower score variance for women compared to men ($\underline{F} = 1.6, \underline{p} < .01$ on AFRAT A and $\underline{F} = 2.3, \underline{p} < .01$ on AFRAT B). At least two plausible explanations for the gender difference in score variance exist. First, the mean AFRAT scores were 2.5 points higher for women than men, thus restricting the range. Second, some previous studies of aptitude/achievement tests have revealed

higher male variance on a number of tests (Jensen, 1980).

Test Intercorrelations

Table 6 shows the intercorrelations for tests given to Air Force subjects in sample 5. These r 's have not been corrected for restriction in range from selection on the ASVAB since it is doubtful that assumptions required to make corrections can be met. Despite the attenuation, the alternate AFRAT forms correlated .73. The degree of restriction in these r 's is illustrated by comparing the r between G-M and ASVAB General of .57 to the r of .76 obtained between the same two measures in a study using service applicants (Mathews, Valentine, & Sellman, 1978). The average r with other tests was .65 for AFRAT A and .63 for AFRAT B. These AFRAT forms correlated somewhat more highly with other reading tests than did the TABE. Average r 's for AFRAT and TABE were .65 and .57 with G-M, respectively, and .62 and .50 with ABLE, respectively. The two AFRAT forms correlated to the same degree with TABE as they did with the G-M and ABLE (average $r = .64$) with both AFRAT and TABE.

Table 7 gives intercorrelations of similar subtests across reading tests. Among vocabulary subtests, the highest r , .68, was between the two AFRAT forms. For comprehension subtests, the r between AFRAT A and B, .62, was again the highest. Correlations among comprehension tests were generally lower than r 's among vocabulary tests. This would be indicative of more unique variance within the different tests.

AFRAT Norming

Descriptive statistics for forms A and B are listed for Sample 1 in Table 8. AFRAT means and standard deviations for Army samples are given in Table 9. Adjusted means are also shown based on regression to compensate for ability differences on the ASVAB General composite. These differences noted earlier are caused by sample fluctuations. These means, 58.6 for AFRAT A and 58.1 for AFRAT B, should be representative since these samples had the same average ability as the normative population. However, as previously mentioned, the distribution of scores in the general population would differ.

AFRAT A and B were equated using the equipercentile method with Air Force samples. Because the forms appear parallel, the raw scores were combined to compute percentiles and to give a single, more stable conversion table (see Table 10). At every percentile point, AFRAT A and B raw scores are within one point of the average raw score.

The AFRAT is negatively skewed, which is appropriate for a test designed to identify low-performing subjects. The median AFRAT score (50th percentile) was 72, compared to a mean of about 69 (from Table 8). A higher median than mean is characteristic of negatively skewed tests.

AFRAT percentiles for Army samples are listed in Table 11. The median score was about 62, compared to a mean of 58.

Table 12 contains an equipercentile calibration of AFRAT scores to ASVAB General (or General-Technical) composite percentiles based on combined Air Force and Army subjects (Samples 5 and 6). The General composite is the ASVAB measure which has been found to correlate highest with reading tests (Mathews et al., 1978).

Equipercentile calibrations of other reading tests to ASVAB general percentiles are shown in Table 13. The data on ABLE and G-M are based on previous studies (see "Samples" subsection), and the TABE data are from Air Force Sample 5 in this study.

It is apparent that there are substantial differences in grade level norms among the commercial reading tests. At some specific levels, one grade or more separates each of the reading tests from another. Without substantial evidence as to which test yields the most accurate RGL conversions, a good estimate should be obtained by averaging the RGLs across the commercial tests. The column on the right side of Table 13 gives this average.

Equipercentile conversions of average RGL for each AFRAT total raw score point are shown in Table 14. Separate RGL conversions for AFRAT Vocabulary and Comprehension subscores are listed in Table 15.

Technical Training Validation

In order to get an initial estimate of the predictive validity of the item types in AFRAT, Form X was administered to about 3,000 airmen.

Technical training grades were subsequently obtained for those in common Air Force Specialty Code (AFSC) groups. Validities for AFRAT X in 19 AFSC groups (total N = 2,253) are listed in Table 15. The median r with training grades was .40. Validities were generally higher for Comprehension than Vocabulary. This is to be expected due to selection on the ASVAB General Composite which has more vocabulary than reading comprehension content. This would severely restrict r 's involving a vocabulary test given after qualifying on ASVAB. A more complete validation study involving AFRAT forms A and B will be accomplished when criterion data are obtained for sufficiently large samples.

IV. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Two parallel forms of the AFRAT have been developed and calibrated to three commonly used reading tests. AFRAT forms appear to adequately meet administrative and psychometric specifications. All items correlate positively with total test score and are in an appropriate range of difficulty (from average to very easy) for use in detecting reading deficiency.

The AFRAT appears to be a highly reliable instrument (internal consistency coefficients of .92 for Form A and .91 for Form B). The two AFRAT forms appear parallel based on similar distributions of item difficulty and criterion correlation values and equal means and variances. AFRAT correlated .60 or higher with each of the three commercial tests.

Interpretation of AFRAT score is provided by percentile norms and calibration to an average RGL based on the commercial tests. A calibration is also presented with General percentile scores. A preliminary analysis indicated the AFRAT would be a valid predictor of technical training performance.

It is recommended that AFRAT Forms A and B replace the TABE and other reading tests for use in screening enlistees for marginal or inadequate reading ability.

REFERENCES

- Angoff, W. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971, 508-600.
- Gates, A. I., & MacGinitie, W. H. Gates-MacGinitie Reading Tests Technical Manual. New York: Teachers College Press, 1972.
- Guilford, J. P., & Fruchter, B. Fundamental statistics in psychology and education. New York: McGraw-Hill (6th ed.), 1978).
- Jensen, A. R. Bias in mental testing. London: Methuen, 1980.
- Karlsen, B., Madden, R., & Gardner, E. ABLE Handbook. New York: Harcourt Brace Jovanovich, 1971.
- Koplyay, J. B. Item analysis program (IAP) for achievement tests. AFHRL-TP-81-22. Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division, October 1981.
- Mathews, J. J., Valentine, L. D., & Sellman, W. S. Prediction of reading grade levels of service applicants from Armed Services Vocational Aptitude Battery (ASVAB). AFHRL-TR-78-82. Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, Air Force Systems Command, December 1978.

Ree, M. J., Mullins, C. J., Mathews, J. J., & Massey, R. H. Armed Services Vocational Aptitude Battery: Item and factor analyses of forms 8, 9, and 10. AFHRL-TR-81-55. Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division, 1981, in press.

Tests of Adult Basic Education Examiner's Manual: Level D. Monterey, CA: CTB/McGraw-Hill, 1976.