APA Symposium on
"Project TALENT — Analysis of Student and School Characteristics"

Tentative Outline of Marion F. Shaycoft's Paper

I. The TALENT Battery (with particular reference to the aptitude and achievement tests)
   A. Description of the battery
      (Very brief description: not more than about two minutes.)
   B. Reliability
      1. TALENT population, 4% sample
         KR-21's for boys, girls, Grades 9 and 12
      2. AF samples
         C-33*
         C-36* (Info. Tests, Parts I, II)
      3. Reliability for the TALENT population (Grade 12 boys) inferred from AF
         reliabilities corrected for range
         C-33
         C-36 (Info. Test, Part I only)
   C. Validity
      1. Evidence from correlation matrices
         a. Subsample 0 correlation matrices (60 variables); by grade (9 and 12)
            and sex; complete cases only
         b. Matrices II, III, and IV (incomplete cases included); these matrices
            are not available yet, but it is hoped they will be completed in August.
            The 3 matrices are to be based on 10% subsample of subsample 0, and
            contain 50 variables each (selected aptitude and achievement tests and
            selected SIB items; two SA1 scales are also included: Culture and
            Mature Personality). It is hoped that some of the SIB items will con-
            tribute evidence of test validity.
      2. Plans for validating against first follow-up questionnaire items will be
         mentioned. Brief description of the follow-up items that are explicitly
         intended for use as test criteria.
      3. C-33 validities
         Multiple R's and multiple regression weights against AFOQT, AQE, AFQT
         criteria. (Multiple R's on the whole are high.)

*C-33 is a project in which about 2500 airmen were given the entire TALENT battery,
so that AF test scores could be equated to TALENT scores.
C-36 is a project in which scores on the TALENT Info. Test were obtained for 1000
airmen.
4. Nominations of students with special talents (in school questionnaires); relation to test score profiles

D. Reliability of residual scores on Info. Test
   1. Presentation of results (C-36 data)
   2. Interpretation
      (These results in one sense are evidence of construct validity of Information test.)

II. Some technical considerations involving psychometric problems

A. Evidence concerning level of "noise" in data
   1. In general, noise level appears low.
   2. Specific examples - e.g., Boys who plan to be "housewives"; boys who say they don't expect to be drafted because they are girls

B. Utility of part-scores on tests (with particular reference to the C-36 data)
   1. Reliability of part-score residuals after effect of a perfectly reliable measure of the function measured by total score has been removed from the part-score
      a. Interpretation
         1) With and without correction for inclusion of part-score itself in total
         2) Importance of correcting correlations for attenuation due to lack of perfect reliability of total score
            a) Logic of correcting
            b) Partial r may be spurious measure of reliability if no correction for attenuation is used.
      b. Formula
         Note that formula is identical with partial r, providing variable partialed out has been corrected for attenuation
      c. C-36 results (previously presented; see paragraph I-D above)
      d. Distinction between use of this measure (reliability of residual) and use of Flanagan uniqueness procedure
         1) Latter used in test development to help determine organization of final battery - what tests contribute unique variance, etc.
2) Reliability of residual not useful for this purpose (since the total score partialled out is an unweighted total) except perhaps as a preliminary screening device. (If none of the scores turns out to have any residual reliability, there is no point in scoring the test for anything except total score.)

3) The existence of residual reliability for some or all of the part-scores is evidence that raw unweighted total does not yield all the info. to be yielded. (However, even for some of the part-scores possessing residual reliability, a uniqueness analysis might still demonstrate that the part-score contributes no unique variance when correlated with an optimally weighted composite of the other parts.)

4) Thus the two statistical procedures (uniqueness analysis and reliability of residuals) are complementary, the former playing the leading role in development of a battery and the residual playing an important role after the battery is in final form, in decisions as to what scores besides the total are worth obtaining.

C. Stability of multiple regression weights

1. Betas don’t bounce as much as they have been believed to.

a. When the N is substantial and the tests adequately reliable, regression weights are surprisingly stable; and in the few instances where they bounce much, a special reason is usually apparent.

b. C-33 data as evidence
Regression weights obtained independently on two samples

1) Correlation between beta-weights for the two samples

2) Correlation between b-weights for the two samples

3) Point out that correlation between b's (and perhaps between betas as well) shouldn’t be taken at face value except as evidence of $r > 0$. Statistical artifact affects the exact value, which can readily be made higher or lower by legitimate algebraic manipulation of the variables (e.g., change scale, "rotate axes", etc.)

D. Interpretation of betas

1. A sensible interpretation can usually be found even for betas that look strange, if they are large enough to be significant.

a. Examples from the C-33 data

1) Reading Comp. Attempts score has positive beta for Officer Biographical Inventory criterion, while Reading Comp. Rights score wasn’t pulled as predictor. (Rdg. Comp. is largely a power test. Interpret results as weighting reading ability plus tendency to guess whenever don’t know the answer. Tendency to guess is a personality factor. Attempts score provides operational measure of it.)
2) "Screening" rights score: has negative weight for verbal criterion. (This can be interpreted as essentially a weight for Screening Attempts since there are almost no errors. Negative weight apparently has effect of "correcting" for unusually slow response rate which might have had effect of pulling down some of the other scores unduly.)

3) Speeded tests

4) Other

E. Relation between Rights and Attempts scores

1. Can incorporate two scores per test, if available, e.g., Rights and Attempts (or perhaps "corrected" score and "omits" into multiple regression analysis. This yields optimal weights for specific criterion; thus optimal scoring formula.

2. Or, with Rights and Attempts in multiple regression analysis, can see what the weights would be for score corrected for chance in conventional way \( X = R - \frac{W}{n-1} \) and for number of items presumably guessed \( G = \frac{n}{n-1}W = A-X \). (This procedure is mathematically akin to rotation of correlated factors.) Apply procedure to C-33 results and interpret.

III. The American high school student

A. National norms (means only): Grades 9–12

B. Shape of growth curves; Grades 9–12
   (Point out inaccuracies due to dropouts)

C. Replication for 10 subsamples

D. Intercorrelations
The program says I am going to talk about the characteristics of American youth. However on the assumption that the title "Tests and Techniques" is at least slightly more appropriate to what I am going to talk about, that is the title that appears on the handouts. I am going to focus first on what evidence we have that the TALENT tests can provide us with reliable and valid information concerning the characteristics of American youth, and second, on some statistical techniques that are or will be involved in the analyses of the raw data provided by the tests.

I. The TALENT Battery

A. Description of the battery.

The cognitive tests portion of the TALENT battery consists of a very long information test that can yield scores on amounts of information in over 20 areas, an English achievement test, a math achievement test, and the following aptitude tests:

- Memory for Sentences
- Memory for Words
- Disguised Words
- Word Functions in Sentences
- Reading Comprehension
- Creativity
- Mechanical Reasoning
- Visualization in 2 dimensions
- Visualization in 3 dimensions
- Abstract Reasoning
- Arithmetic Computation
- Table Reading
- Clerical Checking
- Object Inspection
B. Reliability of the tests

How reliable are the scores we yielded by these tests? For most of the tests we now have quite a bit of evidence on that point. Before taking up specific results, let's consider, for a moment, the purposes of the battery. In developing the battery, our primary goal was prediction of success in a very wide variety of activities. Therefore we aimed at broad coverage with a great many relatively short tests, rather than a smaller number of very long, very reliable tests. Extreme reliability of individual tests didn't seem essential since in almost all cases the best predictors would turn out to be composite scores—and also since any analyses where we were concerned with individual tests would probably be of such a nature that we would be studying groups, not individuals.

All this is a preamble to the fact that the reliability coefficients for most of our tests are considerably below the .90 or .95 that some old-fashioned textbooks on testing say is the minimum for an acceptable test, but that we nevertheless consider the results quite satisfactory.

Table 1 shows KR21 reliabilities for Grades 9 and 12, separately for boys and girls.

We do have some round-about split-half reliabilities. As you heard if you were at the Division 19 symposium yesterday, the TALENT battery has been given to some military personnel, including several thousand airmen. We obtained split-half reliability coefficients on all the tests for which this approach seemed defensible. (This excluded
our six highly speeded tests and the two memory tests, each of which involved a single learning task. The resultant split-half reliability coefficients are shown in Tables 2 and 3, in the columns headed $r_{tt}$.

Using these Air Force reliabilities and applying the standard procedure for correcting reliabilities for range, we obtained reliability coefficients for 12th grade boys. (For this purpose we used the standard deviation obtained for a subsample consisting of 10% of the TALENT schools. These reliability estimates for Grade 12 boys are shown in Table 4.

As you will note from Tables 1, 2, 3, and 4, we have evidence on the reliability of most of the tests from several sources, and the results on the whole are fairly consistent.

C. Validity of the tests

On validity of the tests, of course we haven't nearly as much data as on their reliability. This is the old story—reliability can be determined almost immediately, while validity requires criterion data—and criterion data for predictive validity requires time! Five years from now we will know far more about what tests are valid for what purposes than we do now. As a matter of fact we will have a substantial amount of information along these lines in a matter of months when we start analyzing the data from our first follow-up, on which we are now receiving returns. This follow-up is a 56-item questionnaire. Most of these questions will have some bearing—direct or indirect—on test validity. We will find out, for instance, which of the students went to college, their college grades, their plans for a major, their plans for advanced degrees, if any, and their career plans. For the students who didn't go to college, among other things, we are finding out what kinds of jobs they have, if
any, what their career plans are, and their reasons for not going to college.

Here are some of the questions in the questionnaire:

(Read a few questions)

Responses to these questions will take care of a large proportion of our validation activities in the immediate future.
But we are in a position to make a few definite statements concerning certain kinds of validity for the battery. For instance, the Air Force study that I have already mentioned provides some evidence. We know, for one thing, that the TALENT battery can reproduce with a high degree of accuracy the validated predictive composite scores in the Air Force Officer Qualification Test. Using these measures as criteria and the TALENT scores as predictors, we obtained multiple R's ranging from .82 to .91. More specifically, the multiple R's were:

- Pilot composite: R = .82
- Navigator-technical composite: R = .91
- Officer quality composite: R = .86
- Verbal composite: R = .87
- Quantitative composite: R = .85

Of course these multiple R's just indicate the degree to which a TALENT composite can predict the Air Force predictors—not the degree to which it predicts an Air Force performance criterion. But with multiple R's as high as the ones I just cited, it seems reasonable to suppose that the criteria are also predicted reasonably well.

There is one more kind of evidence concerning validity that I want to mention. This kind applies to the subscores of the Information Test only. It would probably be a little more accurate to refer to it as evidence concerning reliability since it is only indirectly an indication of potential validity, at that.

This is evidence concerning the reliability of residual scores when true score on the total Information test is partialled out. One of the purposes of the Information test—its main purpose, as a
matter of fact—is to measure differential information and thus differential interests.

There are over 25 information areas in which there are at least six items each. But separate scores in these areas would be pointless if they didn’t provide some information qualitatively different from that provided by the total score. Conversely, if it can be shown that the scores in the separate areas are qualitatively different from the total score, this suggests strongly—though it doesn’t prove—that the separate scores are doing what it was hoped they would do—indicating differential interests rather than just general information.

Thus the partial correlation between scores on parallel forms of a part score with the hypothetical true total score partialed out may be regarded as an indication of the utility of the part score. I am referring to this statistic as a “coefficient of part-score utility—Type A”—because it seems to need a label of some kind, and especially because it seems to need a label that will differentiate it from the “coefficients of part-score utility, Type B.” More about that later.

The formula for the Type A coefficient is given on page 1 of your handout, where it is identified as partial correlation $r_{t\cdot g}$. It can be proved that this partial correlation equals the reliability of the residual scores.
The type B part-score residual coefficients are a slight refinement of the Type A ones. In the Type B coefficient a correction is made for inclusion of the score itself in the total. In other words, the variable partialed out is the true score on the remainder of the test. This refinement doesn't make much difference when there are numerous parts in the total test, so that each part has a relatively minor effect on the total.

Table 2 shows both kinds of part-score utility coefficients for each of the 25 scales of the information test, for one Air Force Sample, and Table 3 shows those for the 15 scales of interest in Part I of the Information Test, for two other Air Force samples.

Let's look at these results. It doesn't make much difference in this case whether you look at the Type A coefficients (the ones with $C_{ii}$ partialed out) or the Type B coefficients (the ones with $C'$ partialed out). The results are very similar since the Information test is so long and has so many scales.

Whichever column we look at, we draw the conclusion that the Information test has many scales that are useful in their own right as well as contributing to the total Information score.

The four Information scales in Part I that are most clearly exceptions to this generalization (at least for an airman population) are Vocabulary, Biological Sciences, Scientific Attitude, Aeronautics and Space, and Home Economics. Most of these exceptions are not surprising.
Vocabulary, for instance:

Having a large vocabulary is one sign of having a large fund of general information, since knowledge of word meanings is, after all, basically a kind of information, and apparently it is a kind that is not qualitatively different from other kinds.

Scientific Attitude was the second scale mentioned as apparently not being different from general information. This scale consists of items interspersed throughout the test which are not actually information items, but are designed to find out whether the individual tends to be superstitious and generally unscientific, or whether, on the other hand, his general approach to events and phenomena is rational and logical. Just to give some idea of what this scale was driving at, here are a couple of sample items:

**Item 1.** Professor Rogers wished to find out whether any of the 950 students in Central High School could demonstrate the power of "mind over matter." When ten pennies are tossed, the chances that all ten of them will fall "heads up" are about one in a thousand. Rogers had each student in turn toss ten pennies. He instructed them to try, by thinking very hard about it, to make all ten pennies fall "heads up." But when one of the boys, Joe Thompson, tossed the coins they all fell "tails up." What does this suggest about Joe?

A. Joe was purposely trying to get all tails.
B. Joe became confused.
C. Joe didn’t have faith in the power of mind over matter.
D. Joe is unlucky.
E. Nothing.
Item 2. Jim Wilson has entered many golf tournaments but has never won one. He usually does very well until the finals. But in the finals when the score is close, he tends to make a few wild shots, which result in his defeat. The best explanation of Wilson's failure to win a tournament is that

A. he is just naturally unlucky.
B. he "goes to pieces" under pressure.
C. he hasn't practiced.
D. he doesn't really care whether he wins.
E. his opponents are just naturally lucky.

Anyhow, for at least the military population, it apparently works out as another kind of general intelligence test, highly correlated with general information. This is not too surprising.

The third scale in this low-utility-coefficient category is home economics.

In view of the fact that the sample is all-male, this outcome is hardly surprising. If boys know much about home economics, it is most likely to be because they have a lot about everything. In other words they have a big fund of general information.

The fourth scale with a lower utility coefficient than we had hoped for is the Biological Science Scale. For this one reason is not so clear. Why is Biological Science information so much less a separate entity, distinct from general information than is Physical science?

I don't have any explanation to offer that convinces me. It may be an artifact due to specific test content, it may have something to do with the fact that biology is more likely to be a required subject in high school than physics or chemistry and that even if it isn't required it is more likely to be chosen on an optional basis, or it may be inherent in the nature of the field itself.
So much for the scales on which the part-score utility coefficients are low. For most of the other information scales in Part I, they are very substantial—even for the screening scale, which was designed to help identify illiterates, non-English-speaking students, students marking answers at random, any students who might deliberately mark wrong answers, etc. Because of its purpose the Screening Scale consists of items so easy that anyone who can learn to read and write and who understands English would be able to answer.

For instance:

**How many days are there in a week?**

A. 2  
B. 3  
C. 5  
D. 7  
E. 10

On the basis of the utility coefficients it would appear that the Screening scale is performing some kind of function in this military population. Further exploration is necessary to find out just how it is operating, but it is operating.

Math Information turned out extremely well, and so did Mechanical Information and Electrical and Electronics Information. So did Sports Information. I might mention in this connection that the outcomes of the part-score utility analysis seem to validate the uniqueness analysis procedure used in developing the test. This procedure which had been developed by Dr. Flanagan, involves determining how much reliable-unique variance is contributed to a battery by a test. This procedure was applied to the Information Test and the kinds of information that seemed to provide considerable unique variance in the analysis of the experimental tryout results were put in Part I, which was earmarked for scoring in scales while the scales which apparently had
little or no uniqueness were put in Part II, which was earmarked for punching of individual responses. It was thought that the Part II items might later be recombined in different ways for predicting specific criteria.

If you compare the utility coefficients for the Parts I scales and the Part II scales you will find that most of the high utility coefficients are in Part I. The chief exception is the "Hunting and Fishing" scale in Part II. This scale, which landed in Part II because it consisted of items that were added after the experimental tryout, turned out extremely well.

Knowledge about hunting and fishing presumably represents special interest in the area, rather than general information. Here is one of the items:

**Bait** would be a good bait for catching

A. carp  
B. small-mouth bass  
C. large-mouth bass  
D. trout  
E. pike
In a sense, then, these part-score utility coefficients provide some evidence of construct validity.

Having decided that most of the separate scales on the Information test are well worth while, and that having just a total information score wouldn’t begin to serve the same purpose, we next directed our attention to an analogous question about the two other tests with part-scores (the Math test and the English test). Again using the Air Force samples we performed the same sort of analysis. The resultant utility coefficients are shown in Table 3.

In this case, whether the type A coefficients or the type B coefficients are used makes a real difference, because the tests have fewer parts, and therefore each part has a greater effect on the total. Type A is systematically lower than Type B because with Type A, more is partialled out. Type B seems to have a slight advantage over Type A, not only on the basis of magnitude but also on theoretical grounds.

Where the two types differ, the interpretation of the Type B coefficient may be of little clearer than Type A and also it probably would be simpler to derive a standard error for Type B, although this hasn’t been done yet.

Let’s look at Table 3 again, where the utility coefficients are for English and math. It would appear on the basis of these data that part scores are serving a
purpose in both these areas. All three parts of the Math test have substantial utility coefficients. Part I is arithmetic reasoning. Part II is any other mathematics generally taught in the ninth grade or earlier. It is largely but not wholly algebra. Part III is Math of the kinds covered in Grades 10-12 of an academic curriculum.

Of the five parts of the English test, all except English Usage have substantial utility coefficients. Why this one part doesn't when all the others do, isn't entirely clear. The English usage subtest consists of items of the following sort:

Ed and ______ planning to go

A. myself was
B. me was
C. I was
D. myself were
E. I were

* E. I were

The statistical results seem to say that the ability to answer this type of item correctly is almost identical with knowledge of correct spelling, capitalization and punctuation.

II. Some technical considerations

A. Part-score utility coefficients

Before we move on to the next topic there are a couple of additional points I want to make concerning the part-score utility coefficient procedure. First, there is
out true scores. Wouldn't obtained scores do as well? The answer is that if the total score is extremely reliable as it is in the case of our Information Test, it doesn't actually make much difference, but that for shorter tests the results are likely to be quite misleading if obtained scores were partialed out instead of true scores. Even with part scores factorially identical to total scores, one would be quite likely to wind up with substantial positive partial r's due merely to the unreliability of the total.

Another reason that it seems desirable to use true scores for the total is that with true scores the partial r is mathematically equivalent to the reliability of the residuals, and with obtained scores it isn't. Incidentally, the reliability of residuals with obtained scores partialed out is also shown in Table 2 for the Information Test. These values are very close to the Type A and Type B utility coefficients, as you will note, but this is only because the total has such a high reliability coefficient.

One last question that may have occurred to some of you concerning utility coefficients. Having used the uniqueness analysis procedure to develop a test, why don't we also use it instead of part-score utility coefficients to analyze the results later? The answer is that the two procedures are complementary. The uniqueness coefficient is used in test development to help determine the organization of the final battery by determining what tests contribute unique variance. The utility coefficient would be useless for this purpose,
except conceivably as a preliminary screening device, since the total score partialed out is an unweighted total. The function of the utility coefficients is to determine whether any (or all) of the part scores have any residual reliability. If they do, this is adequate evidence that the raw unweighted total does not yield all the information to be yielded, and even that any conceivable weighted total wouldn't. Thus the utility coefficients help decide what scores beside the total are worth obtaining.

B. Analysis of Scoring Formulas, etc.

Analysis of the data has led to some interesting sidelong concerning the use of number right, number wrong, number attempted, and other components of various scoring formulas.

Table 5 summarizes some of our findings, again for one of the Air Force samples. In this table and table 6, A, R and W represent number attempted, number right and number wrong respectively; X represents the conventional correction-for-chance score; G is the number of items assumed to have been answered by guessing or as the result of some other chance agency. The correlation between R and X are extremely high, most of them being over .99. The correlation between number right and number attempted, however, is quite low for most of the tests, confirming that rights and attempts can be considered in effect to operate as two separate predictors. However, if you're going to have two separate scores
Sometimes the use of two linear functions of these variables gives results that are easier to interpret. Table 6 has some bearing on this point. In this table we show the multiple regression weights of 44 predictors, namely, the rights and attempts scores on 22 Talent tests, for 17 Air Force criteria. These 44 predictors are not the complete set that were involved in this multiple regression analysis; actually there were 74, but there were only 22 tests for which both Rights and Attempts scores were available.

In addition to the multiple regression weights for the Rights and Attempts scores, Table 6 also shows what the weights would have been if the predictors had been the \( X \) and \( G \) scores instead --- in other words, the corrected-for-chance score and the number "guessed".

In some cases the regression weights for \( X \) and \( G \) seem a little easier to interpret or more meaningful than those for \( R \) and \( A \).

For instance, let's look at the weights for Predictor Variable 250, Reading Comprehension, for Criterion 17, Officer Biographical Inventory. There is a significant regression weight for number attempted on this test, while the weight is essential zero for number right. This seems strange, until we look at the weights for variables \( X \) and \( G \), and observe that both would have substantial positive weights. Apparently their Reading Comprehension ability, as measured by the \( X \) variable, is a positively weighted predictor for Officer B.I. score. Likewise, a personality factor, willingness to\( x \) guess when one isn't sure of the correct answer, is also positively related to the Officer B.I. score.
Here's another case in point. On the screening scale, Variable 101, the X weight is -.35 and the G weight is + .39. We can assume that anyone who gets into the Air Force knows there are seven days in a week, and that they can read and write. Therefore we rule out feebleminded and illiteracy as the prime explanatory factors for low screening scores and assume such scores are more likely due to clerical errors and carelessness. Then we can interpret the negative weight for X as indicating that the corrected score is a suppressor variable that corrects for careless errors on other predictor tests, that are positively related to the "verbal" criterion. A somewhat similar interpretation applies to the positive weight for the G scores, which are directly proportional to the number of wrong responses, and thus to the number of careless errors. These are just two examples of the way the X and G weights can be used to clarify interpretation of the A and R weights. In interpreting weights for the speeded tests, as in the case of the screening score, the G score is related more to something that might be called "carelessness" or "inaccuracy" than to something labeled "willingness to guess". Another point to remember is that for the speeded tests, the A score is directly related to speed of response.

A word of caution may be in order concerning interrelation of these regression weights. To avoid pushing the interpretation too far it is well to bear in mind that many of the lower weights are not significant. John Dailey di-
cussed this point yesterday in the Division 19 symposium, when he told about our stepwise and asymptotic multiple regression procedures, which help us distinguish between the weights that are significant and those that are not.

Before closing, I would like to mention a couple of pieces of data we have secured that throw some light on the question of whether the student's responses are badly contaminated by clerical errors and other careless errors. Apparently not too badly.

In response to the Student Information Blank question -- "What do you expect to do about military service?", only 1.1% of the 9th grade boys and 0.5% of 12th grade boys choose the option "Never serve, because I am a girl." Some of these responses are probably flippant and the rest probably due to careless errors. Likewise 0.5% of the 9th grade boys and 0.1% of the 12th grade boys mark "housewife" as their career choices. These percentages are probably fairly good indications of the degree of error in other responses, where error is not so easily identified.

In conclusion then, we have found out some encouraging things about the reliability and validity of our tests, the usefulness of the scores in giving differential measures, the kinds of scoring procedures that are useful, and the small degree to which the test and inventory data are apparently contaminated by careless errors.

Everything we have found out along these lines gives
us confidence in the results of the analyses to date of
data about schools and students, and the many more analyses
that we expect to complete in the future.
Symposium: "Project TALENT--Analysis of Student and School Characteristics"

Tables to accompany paper on
"Tests and Techniques"

--Marion F. Shaycoft
American Institute for Research
and
University of Pittsburgh

In the attached tables, the following notation is used throughout, except where otherwise indicated:

\[ N = \text{no of cases in sample} \]
\[ n = \text{no of items in test} \]
\[ n' = \text{no of options per item (or mean no of options per item)} \]
\[ r_{tt} = \text{split-half reliability corrected by Angoff formula \#16 unless otherwise indicated} \]

Formulas:

Angoff formula \#16* is

\[
 r_{tt} = \frac{r_{ab} \sigma_{t}^2}{(\frac{1}{\sigma_a^2} + r_{ab} \frac{1}{\sigma_b^2}) (\frac{1}{\sigma_b^2} + r_{ab} \frac{1}{\sigma_a^2})} 
\]

where \( a \) and \( b \) are half-test scores and \( t = \text{total score} = a + b \)

Coefficient of part-score utility: Type A

\[
 r_{tt} \cdot c_\infty = 1 - \frac{r_{cc} \sigma_c^2 (1 - r_{tt})}{r_{cc} \sigma_c^2 - (r_{tt} \sigma_t + r_{tc} \sigma_c - \sigma_t)^2} 
\]

where: \( t_i = \text{part-score (on part } i) = \text{one of the components of } c \)

\[ c = \text{total score} = \sum_{i=1}^{t} t_i \]

\[ c_\infty = \text{hypothetical "true score" on variable } c \]

Coefficient of part-score utility: Type B

\[
r_{tt} \cdot c_{\infty} = - \frac{r_{c'c'} \sigma_{c'}^2 (1 - r_{tt})}{r_{c'c'} \sigma_{c'}^2 - (r_{tc} \sigma_{c} - \sigma_{t})^2}
\]

where \( r_{c'c'} \sigma_{c'}^2 = r_{cc} \sigma_{c}^2 - r_{tt} \sigma_{t}^2 - 2 \sigma_{t} (r_{tc} \sigma_{c} - \sigma_{t}) \)

\( c = \text{total score} = \sum_{i} t_{i} \)

\( t_{i} = \text{part-score (on part i) = one of the components of c} \)

\( c'_{i} = \text{remainder of test} = c - t_{i} \)

\( c'_{\infty} = \text{hypothetical "true score" on variable c'} \)

Reliability of residuals

It can be proved that \( r_{tt} \cdot c_{\infty} \) equals the reliability of the residuals of \( t \) when hypothetical true score \( c_{\infty} \) is partialled out.

It can also be proved that \( r_{tt} \cdot c_{\infty} \) equals the reliability of the residuals of \( t \) when hypothetical true score \( c'_{\infty} \) is partialled out.

The reliability of the residuals of \( t \) when obtained total score \( c \) is partialled out is given in the following formula, where it is designated as \( r' \).

\[
r' = \frac{r_{tt} - 2 r_{tc}^2 + r_{cc} r_{tc}^2}{1 - r_{tc}^2} + \frac{2 \sigma_{t} r_{tc} (1 - r_{tt})}{\sigma_{c} (1 - r_{tc}^2)}
\]

In the above formula, \( t \) and \( c \) have the same meaning as for the \( r_{tt} \cdot c_{\infty} \) formula.
Notation and formulas for Tables 5 and 6:

Scores

\[ R = \text{rights score} \]
\[ A = \text{no. of items attempted} \]
\[ W = \text{no. of wrong responses} = A - R \]
\[ X = \text{score corrected for chance by conventional formula} \]
\[ X = R - \frac{W}{n' - 1} = \frac{n'R - A}{n' - 1} \]
\[ G = \text{no. of responses assumed to have been obtained by guessing or by chance} \]
\[ G = \frac{n'W}{n' - 1} \]

Means and variances

\[ \bar{W} = \bar{A} - \bar{R} \]
\[ \bar{X} = \frac{n'R - \bar{A}}{n' - 1} \]
\[ \sigma_W^2 = \sigma_A^2 + \sigma_R^2 - 2r_{AR} \sigma_A \sigma_R \]
\[ \sigma_X^2 = \frac{1}{(n' - 1)^2} (n^2 \sigma_R^2 + \sigma_A^2 - 2nr_{AR} \sigma_A \sigma_R) \]

Correlations

\[ r_{RW} = \frac{r_{AR} \sigma_A - \sigma_R}{\sigma_W} \]
\[ r_{AW} = \frac{r_{AR} \sigma_A - \sigma_R}{\sigma_W} \]
\[ r_{XG} = \frac{(n' + 1) r_{AR} \sigma_A - \sigma_R^2 - n' r_{AR} \sigma_A^2}{(n' - 1) \sigma_X \sigma_W} \]
\[ r_{AX} = \frac{n'r_{AR} \sigma_A - \sigma_A}{(n' - 1) \sigma_X} \]
\[ r_{RX} = \frac{n' r_{AR} \sigma_A}{(n' - 1) \sigma_X} \]
Regression weights

\[ b_X = \frac{c + \beta}{\sigma^2_A} \]

\[ b_G = \frac{n'b_A + b_R}{n'} \]
Table 1: KR-21 Reliabilities for Talent Tests

Based on 4% sample of the answer sheets returned by May 1960

<table>
<thead>
<tr>
<th>Test</th>
<th>No. of Items</th>
<th>9th Grade</th>
<th></th>
<th></th>
<th>12th Grade</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Boys</td>
<td>Girls</td>
<td>Boys</td>
<td>Girls</td>
<td>Boys</td>
<td>Girls</td>
</tr>
<tr>
<td>Information I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>101. Screening</td>
<td>12</td>
<td>.732</td>
<td>.629</td>
<td>.630</td>
<td>.623</td>
<td></td>
<td></td>
</tr>
<tr>
<td>102. Vocabulary</td>
<td>21</td>
<td>.707</td>
<td>.697</td>
<td>.707</td>
<td>.728</td>
<td></td>
<td></td>
</tr>
<tr>
<td>103. Literature</td>
<td>24</td>
<td>.689</td>
<td>.607</td>
<td>.760</td>
<td>.754</td>
<td></td>
<td></td>
</tr>
<tr>
<td>104. Music</td>
<td>13</td>
<td>.632</td>
<td>.617</td>
<td>.705</td>
<td>.691</td>
<td></td>
<td></td>
</tr>
<tr>
<td>105. Soc. Studies</td>
<td>24</td>
<td>.850</td>
<td>.790</td>
<td>.843</td>
<td>.785</td>
<td></td>
<td></td>
</tr>
<tr>
<td>106. Math</td>
<td>23</td>
<td>.711</td>
<td>.636</td>
<td>.892</td>
<td>.856</td>
<td></td>
<td></td>
</tr>
<tr>
<td>108. Biol. Sci.</td>
<td>11</td>
<td>.570</td>
<td>.445</td>
<td>.555</td>
<td>.528</td>
<td></td>
<td></td>
</tr>
<tr>
<td>110. Aero &amp; Space</td>
<td>10</td>
<td>.619</td>
<td>.216</td>
<td>.671</td>
<td>.382</td>
<td></td>
<td></td>
</tr>
<tr>
<td>111. Elec.</td>
<td>20</td>
<td>.731</td>
<td>.459</td>
<td>.804</td>
<td>.488</td>
<td></td>
<td></td>
</tr>
<tr>
<td>112. Mech</td>
<td>19</td>
<td>.678</td>
<td>.455</td>
<td>.647</td>
<td>.523</td>
<td></td>
<td></td>
</tr>
<tr>
<td>113. Farming</td>
<td>12</td>
<td>.621</td>
<td>.622</td>
<td>.496</td>
<td>.570</td>
<td></td>
<td></td>
</tr>
<tr>
<td>114. Home Ec.</td>
<td>21</td>
<td>.421</td>
<td>.522</td>
<td>.414</td>
<td>.567</td>
<td></td>
<td></td>
</tr>
<tr>
<td>190. Total</td>
<td>252</td>
<td>.957</td>
<td>.937</td>
<td>.962</td>
<td>.949</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>231. Spelling</td>
<td>16</td>
<td>.575</td>
<td>.499</td>
<td>.596</td>
<td>.542</td>
<td></td>
<td></td>
</tr>
<tr>
<td>232. Cap.</td>
<td>33</td>
<td>.837</td>
<td>.811</td>
<td>.805</td>
<td>.738</td>
<td></td>
<td></td>
</tr>
<tr>
<td>233. Punct.</td>
<td>27</td>
<td>.700</td>
<td>.702</td>
<td>.730</td>
<td>.713</td>
<td></td>
<td></td>
</tr>
<tr>
<td>234. Usage</td>
<td>25</td>
<td>.510</td>
<td>.454</td>
<td>.567</td>
<td>.468</td>
<td></td>
<td></td>
</tr>
<tr>
<td>235. Expression</td>
<td>12</td>
<td>.633</td>
<td>.523</td>
<td>.599</td>
<td>.538</td>
<td></td>
<td></td>
</tr>
<tr>
<td>236. Total</td>
<td>113</td>
<td>.883</td>
<td>.866</td>
<td>.881</td>
<td>.855</td>
<td></td>
<td></td>
</tr>
<tr>
<td>240. Wd. Punct.</td>
<td>24</td>
<td>.780</td>
<td>.805</td>
<td>.832</td>
<td>.817</td>
<td></td>
<td></td>
</tr>
<tr>
<td>240. Creativity</td>
<td>20</td>
<td>.717</td>
<td>.646</td>
<td>.741</td>
<td>.708</td>
<td></td>
<td></td>
</tr>
<tr>
<td>282. Vis. in 3 Dim.</td>
<td>16</td>
<td>.670</td>
<td>.563</td>
<td>.704</td>
<td>.612</td>
<td></td>
<td></td>
</tr>
<tr>
<td>290. Abst. Reas.</td>
<td>15</td>
<td>.676</td>
<td>.687</td>
<td>.627</td>
<td>.627</td>
<td></td>
<td></td>
</tr>
<tr>
<td>311. Math I</td>
<td>16</td>
<td>.704</td>
<td>.684</td>
<td>.757</td>
<td>.735</td>
<td></td>
<td></td>
</tr>
<tr>
<td>312. Math II</td>
<td>24</td>
<td>.691</td>
<td>.630</td>
<td>.850</td>
<td>.786</td>
<td></td>
<td></td>
</tr>
<tr>
<td>320. Math I + II</td>
<td>40</td>
<td>.812</td>
<td>.781</td>
<td>.889</td>
<td>.851</td>
<td></td>
<td></td>
</tr>
<tr>
<td>340. I + II + III</td>
<td>54</td>
<td>.750</td>
<td>.756</td>
<td>.910</td>
<td>.867</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

N                  | 1152         | 990      | 1153  | 1250 |

NOTE: The following tests are omitted from this table because KR-21 coefficients would have been spuriously high.

211. Memory for Sentences 212. Memory for Words
(.memory tests with a single learning task)

250. Reading Comprehension
(Items not experimentally independent)

220. Disguised Words
281. Visualization in 2 Dimensions
410. Arithmetic Computation
420. Table Reading
430. Clerical Checking
440. Object Inspection

Highly speeded tests
Table 2: TALENT Information Test: Part-Score Utility Coefficients, Reliability Coefficients, and Other Statistics - Air Force Sample II: N = 466 Basic Airmen

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>M</th>
<th>σ</th>
<th>r_{tt}</th>
<th>r_{tc}</th>
<th>r_{tt,c}</th>
<th>r_{tt,c}</th>
<th>r'</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PART I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>101 Screening</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>102 Vocabulary</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>102a Vocab.</td>
<td></td>
<td></td>
<td></td>
<td>1.364*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>103 Literature</td>
<td>24</td>
<td>12.15</td>
<td>4.37</td>
<td></td>
<td>0.701</td>
<td>-0.223</td>
<td>-0.207</td>
<td>-0.214</td>
</tr>
<tr>
<td>104 Music</td>
<td>12</td>
<td>5.32</td>
<td>2.68</td>
<td></td>
<td>0.760</td>
<td>0.349</td>
<td>0.312</td>
<td>0.328</td>
</tr>
<tr>
<td>105 Social Studies</td>
<td>24</td>
<td>14.38</td>
<td>5.26</td>
<td></td>
<td>0.817</td>
<td>0.527</td>
<td>0.474</td>
<td>0.487</td>
</tr>
<tr>
<td>106 Math</td>
<td>23</td>
<td>7.58</td>
<td>4.82</td>
<td></td>
<td>0.728</td>
<td>0.690</td>
<td>0.652</td>
<td>0.607</td>
</tr>
<tr>
<td>107 Physical Science</td>
<td>18</td>
<td>8.56</td>
<td>4.10</td>
<td></td>
<td>0.789</td>
<td>0.554</td>
<td>0.515</td>
<td>0.518</td>
</tr>
<tr>
<td>108 Biological Science</td>
<td>11</td>
<td>5.61</td>
<td>2.01</td>
<td></td>
<td>0.689</td>
<td>0.216</td>
<td>0.015</td>
<td>0.015</td>
</tr>
<tr>
<td>109 Scientific Attitude</td>
<td>10</td>
<td>5.78</td>
<td>1.84</td>
<td></td>
<td>0.577</td>
<td>0.165</td>
<td>0.158</td>
<td>0.162</td>
</tr>
<tr>
<td>110 Aeronautics, Space</td>
<td>10</td>
<td>4.53</td>
<td>2.36</td>
<td></td>
<td>0.718</td>
<td>0.284</td>
<td>0.268</td>
<td>0.272</td>
</tr>
<tr>
<td>111 Electricity, Electronics</td>
<td>20</td>
<td>9.26</td>
<td>4.27</td>
<td></td>
<td>0.692</td>
<td>0.645</td>
<td>0.612</td>
<td>0.622</td>
</tr>
<tr>
<td>112 Mechanics</td>
<td>19</td>
<td>12.74</td>
<td>3.24</td>
<td></td>
<td>0.656</td>
<td>0.451</td>
<td>0.426</td>
<td>0.438</td>
</tr>
<tr>
<td>113 Farming</td>
<td>12</td>
<td>7.62</td>
<td>2.23</td>
<td></td>
<td>0.600</td>
<td>0.380</td>
<td>0.365</td>
<td>0.371</td>
</tr>
<tr>
<td>114 Home Economics</td>
<td>21</td>
<td>8.33</td>
<td>2.83</td>
<td></td>
<td>0.512</td>
<td>0.287</td>
<td>0.273</td>
<td>0.283</td>
</tr>
<tr>
<td>115 Sports</td>
<td>14</td>
<td>7.65</td>
<td>2.81</td>
<td></td>
<td>0.576</td>
<td>0.560</td>
<td>0.541</td>
<td>0.549</td>
</tr>
</tbody>
</table>

PART II

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>M</th>
<th>σ</th>
<th>r_{tt}</th>
<th>r_{tc}</th>
<th>r_{tt,c}</th>
<th>r_{tt,c}</th>
<th>r'</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Art</td>
<td>12</td>
<td>5.73</td>
<td>2.38</td>
<td></td>
<td>0.603</td>
<td></td>
<td>0.706</td>
<td>0.238</td>
</tr>
<tr>
<td>Law</td>
<td>9</td>
<td>4.89</td>
<td>1.71</td>
<td></td>
<td>0.647</td>
<td></td>
<td>0.249</td>
<td>0.239</td>
</tr>
<tr>
<td>Health</td>
<td>9</td>
<td>5.54</td>
<td>1.92</td>
<td></td>
<td>0.699</td>
<td></td>
<td>0.215</td>
<td>0.204</td>
</tr>
<tr>
<td>Engineer</td>
<td>6</td>
<td>3.19</td>
<td>1.21</td>
<td></td>
<td>0.575</td>
<td></td>
<td>0.101</td>
<td>0.098</td>
</tr>
<tr>
<td>Architecture</td>
<td>6</td>
<td>2.52</td>
<td>1.33</td>
<td></td>
<td>0.616</td>
<td></td>
<td>0.084</td>
<td>0.081</td>
</tr>
<tr>
<td>Military</td>
<td>7</td>
<td>3.00</td>
<td>1.38</td>
<td></td>
<td>0.596</td>
<td></td>
<td>0.236</td>
<td>0.229</td>
</tr>
<tr>
<td>Business</td>
<td>10</td>
<td>4.76</td>
<td>1.94</td>
<td></td>
<td>0.657</td>
<td></td>
<td>0.142</td>
<td>0.134</td>
</tr>
<tr>
<td>Outdoor activities</td>
<td>9</td>
<td>4.62</td>
<td>1.72</td>
<td></td>
<td>0.631</td>
<td></td>
<td>0.123</td>
<td>0.118</td>
</tr>
<tr>
<td>Hunting, fishing</td>
<td>10</td>
<td>3.98</td>
<td>2.10</td>
<td></td>
<td>0.372</td>
<td></td>
<td>0.571</td>
<td>0.562</td>
</tr>
<tr>
<td>Theater, ballet</td>
<td>8</td>
<td>3.73</td>
<td>1.59</td>
<td></td>
<td>0.620</td>
<td></td>
<td>0.263</td>
<td>0.254</td>
</tr>
<tr>
<td>Foreign travel</td>
<td>5</td>
<td>2.60</td>
<td>1.28</td>
<td></td>
<td>0.648</td>
<td></td>
<td>0.110</td>
<td>0.106</td>
</tr>
<tr>
<td>Other</td>
<td>52</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C. Total (102a - 126) | 320 | 160.35 | 42.625 | 0.9579 |

* .364 = KR-21 reliability for 10 vocabulary items
** .546 = reliability for 21-item Vocabulary Scale; it was computed by applying the Spearman-Brown formula to the 10-item reliability (.364 ).
Table 3: Means, standard deviations, reliability coefficients, and other statistics, for TALENT Tests  
Based on Air Force cases (two subsamples of Basic Airmen)

<table>
<thead>
<tr>
<th>Test*</th>
<th>Subsample IA: N=1247</th>
<th>Subsample IB: N=1242</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inf.</td>
<td>Part I</td>
<td></td>
</tr>
<tr>
<td>101. Screening</td>
<td>11.30 1.27 .640 .423 .571 .564</td>
<td>11.18 1.37 .645 .462 .558 .549</td>
</tr>
<tr>
<td>102. Vocab.</td>
<td>13.20 3.55 .771 .863 .141 .120</td>
<td>13.13 3.55 .754 .850 .134 .114</td>
</tr>
<tr>
<td>103. Lit.</td>
<td>12.15 4.57 .788 .789 .513 .460</td>
<td>12.25 4.53 .763 .769 .533 .481</td>
</tr>
<tr>
<td>104. Music</td>
<td>5.60 2.99 .751 .733 .500 .469</td>
<td>5.58 2.86 .693 .702 .431 .402</td>
</tr>
<tr>
<td>108. Biol.Sci.</td>
<td>5.80 2.04 .541 .675 .192 .180</td>
<td>5.82 2.03 .594 .665 .295 .278</td>
</tr>
<tr>
<td>109. Sci.Att.</td>
<td>6.10 1.86 .478 .606 .206 .196</td>
<td>6.12 1.84 .491 .586 .251 .239</td>
</tr>
<tr>
<td>110. Aero&amp;Space</td>
<td>4.82 2.48 .674 .735 .282 .261</td>
<td>4.78 2.36 .633 .726 .251 .232</td>
</tr>
<tr>
<td>111. Elec.</td>
<td>10.23 4.60 .830 .751 .662 .516</td>
<td>10.14 4.54 .813 .728 .562 .506</td>
</tr>
<tr>
<td>113. Farming</td>
<td>7.96 2.18 .605 .556 .455 .439</td>
<td>8.04 2.15 .575 .566 .401 .384</td>
</tr>
<tr>
<td>114. Home Ec.</td>
<td>8.76 2.98 .517 .616 .281 .261</td>
<td>8.64 2.95 .472 .591 .245 .227</td>
</tr>
<tr>
<td>190. Total **</td>
<td>36.04 .956</td>
<td>34.88 .941</td>
</tr>
</tbody>
</table>

English

| 231. Spell. | 8.72 2.81 .646 .685 .490 .402 | 8.62 3.02 .679 .709 .513 .416 |
| 232. Csp. | 28.93 3.23 .758 .673 .678 .584 | 28.92 3.69 .815 .760 .763 .675 |
| 234. Usage | 16.06 3.11 .454 .746 .121 .083 | 16.05 3.12 .473 .710 .220 .162 |
| 230. Total** | 78.9111.66 .687 | 78.7811.49 .685 |

Misc.

| 240. Wd.Funct. | 8.36 4.50 .775 | 8.53 4.42 .769 |
| 282. Vis.in 3 Dim. | 9.72 3.27 .755 | 9.74 3.20 .746 |
| 290. Abst. Reas. | 8.91 3.07 .735 | 9.03 2.94 .706 |

Math

| 311. Math I Arith Reas | 8.73 3.53 .760 .874 .522 .300 | 8.82 3.53 .754 .872 .523 .297 |
| 312. II Introd. | 9.78 4.95 .857 .938 .671 .265 | 9.78 4.81 .850 .933 .666 .268 |
| 320. "I+II SubTot" | 18.51 7.73 .889 | 18.60 7.57 .882 |
| 339. "I+II SubTot" | 18.51 7.73 .889 .981 .760 .158 | 18.60 7.57 .882 .979 .752 .137 |
| 333. III Adv. | 3.16 2.47 .692 .794 .403 .304 | 3.09 2.36 .660 .759 .412 .318 |

NOTE: The following tests are omitted from this Table because reliability coefficients could not be obtained. (Split-half coefficients would have been spuriously high.)

211. Memory for Sentences  
212. Memory for Words  
220. Disguised Words  
281. Visualization in Two Dimensions  
410. Arithmetic Computation  
420. Table Reading  
430. Clerical Checking  
440. Object Inspection

Memory Tests with a single learning task
Highly speeded tests

* The data in this table refer to "rights" scores  
** Used as "Variable C" in computing rtt.∞ and rtt.c∞
Based on an AFO sample of N = 1247 basic ability

Test

Table 5. Means, Standard Deviations, and Correlations of Various Kinds of Scores

(On T and AFO tests for which both TTPS scores and attempt scores are available)
<table>
<thead>
<tr>
<th>Letter</th>
<th>Officer Quality</th>
<th>Written-Technical</th>
<th>Pilot</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>q</td>
<td>q</td>
<td>q</td>
<td>q</td>
</tr>
<tr>
<td>B</td>
<td>q</td>
<td>q</td>
<td>q</td>
<td>q</td>
</tr>
<tr>
<td>C</td>
<td>q</td>
<td>q</td>
<td>q</td>
<td>q</td>
</tr>
</tbody>
</table>

*Note: The table represents a score prediction for various officers based on the predictor test.*

Based on Air Force Sample IV: N=1247 Basic Altimetry

Parent score predictions for air force criteria: T=7

Table 6: Regression weights for various kinds of scores where both kinds of scores and attempts scores are available.