Reanalysis of the 1980 AFQT Data from the NLSY79[1]

Pamela Ing
Carole A. Lunney
Randall J. Olsen
Center for Human Resource Research, Ohio State University

PART I. FACTOR ANALYSIS

Motivation: One of the more frequently pursued issues in explaining the behavior and life-course outcomes for individuals is the importance of education. However, many analysts worry that part of the estimated effect of education is really the effect of cognitive ability or mastery of certain cognitive skills rather than the effect of years of education, *per se*. Few data sets attempt to measure cognitive ability or skills, leading to concern that the effects of education may be overstated due to misattribution bias. The NLSY79 and NLSY97 are unique in that virtually all respondents took the Armed Forces Vocational Aptitude Battery (ASVAB), with the results for the NLSY79 being used to norm the test.[2] In this memo, we discuss the psychometrics of the Armed Forces Qualifying Test (AFQT) administered in 1980 with attention to the implications the analysis has for the efficient administration of those tests.

Background: The ASVAB was administered to 11,914 of the 12,686 respondents in 1980 to generate national norms for this battery of tests. The test had ten sections, General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, Numerical Operations, Coding Speed, Auto and Shop Knowledge, Mathematics Knowledge, Mechanical Comprehension and Electronics Information. The test questions were in a booklet and the respondents indicated their answers by filling in the circles on an answer sheet. Prior to 2013, the NLSY79 public use data set only contained the number of questions the respondent got right on each section plus a scale score, which was a standard normal variable with two implied decimal places and an estimated standard error of the scale score with three implied decimal places. There was no indication of what the questions were nor which questions the respondent got right or wrong.

A commonly used score is the AFQT, which was used by the military to determine how readily a recruit could be trained in various military occupational specialties (Army and Marines), ratings (Navy) or Air Force specialty codes (Air Force). Users of the NLSY79 data frequently use the AFQT score to measure ability, although the evidence suggests the AFQT measures, to a significant extent, achievement. Three AFQT percentile scores, an AFQT-1, AFQT-2 and an AFQT-3, were created for each Profiles respondent and are described below.

AFQT-1: To construct AFQT-1, the raw scores from the following four sections of the *ASVAB* are summed:

[1] This research was supported by grant 1RC1HD063405-01from the Eunice Kennedy Shriver Institute of Child Health and Human Development.

[2] The 1997 version of the ASVAB was normed on different respondents as the NLSY97 cohort was much younger than recruits sitting for the ASVAB in order to enlist in the Armed Forces and hence inappropriate for generating norms for the latter group.

- Section 2 (arithmetic reasoning),
- Section 3 (word knowledge),
- Section 4 (paragraph comprehension),
- and one half of the score from Section 5 (numerical operations).

AFQT-2: Beginning in January 1989, DOD began using a new calculation procedure. The numerical operations section of the AFQT-1 had a design inconsistency resulting in respondents getting tests that differed slightly and resulted in slight completion rate differences.

Creation of this revised percentile score, called AFQT-2, involves:

- computing a verbal composite score by summing word knowledge and paragraph comprehension raw scores;
- converting subtest raw scores for verbal, math knowledge, and arithmetic reasoning;
- multiplying the verbal standard score by two;
- summing the standard scores for verbal, math knowledge, and arithmetic reasoning;
- converting the summed standard score to a percentile.

AFQT-3: In 2006 CHRR renormed the AFQT-2 scores, controlling for age because there was a clear age gradient in the AFQT scores described above, which implied the scores confounded education and achievement with native ability. CHRR staff recommend using the AFQT-3. Although the formula is similar to the AFQT score generated by DOD for the NLSY79 cohort, this variable reflects work done by NLS program staff and is neither generated nor endorsed by DOD.

To calculate the AFQT-3, NLS Program staff first grouped respondents into three-month age groups. That is, the oldest cohort included those born from January through March of 1957, while the youngest were born from October through December 1964, a total of 32 cohorts, with an average of about 350 respondents per cohort (there was one unusually small cohort: the youngest cohort has only 145 respondents). The revised dates of birth from the 1981 survey (R0410100 and R0410300) were used whenever these disagreed with the information from the 1979 survey. With the revised birth dates, a few respondents were born outside the 1957-1964 sampling space of the survey.

Those born before 1957 were assigned to the oldest cohort, while those born after 1964 were assigned to the youngest cohort. ASVAB sampling weights from the Profiles section were used (R0614700). Within each three-month age group and using the sampling weights, staff assigned percentiles for the raw scores for the tests on Mathematical Knowledge (MK), Arithmetic Reasoning (AR), Word Knowledge (WK), and Paragraph Comprehension (PC) based on the weighted number of respondents scoring below each score (ties are given half weight). Staff added the percentile scores for WK and PC to get an aggregate Verbal score (V) for which an aggregated intra-group, internally normed, percentile was then computed. NLS Program staff then added the percentile scores for MK, AR and two times the aggregated percentile for V. Finally, within each group we computed a percentile score, using the weights, on this aggregate score, yielding a final value between zero and 100. The work above was all done using the

numerical scores on the various component tests inasmuch as the detailed data on the tests were not available.

Data Recovery: In 2010, the Bureau of Labor Statistics mandated that all paper materials for the NLSY79 stored in Chicago be destroyed in the near future. Preparing for this, CHRR requested that those paper materials be searched so that the documents from Round 1, Round 10 and, if possible, the ASVAB could be found and sent to Columbus to gather additional information on the respondents. The Round 1 documents would be helpful in resolving some long-running problems on the location of the respondents at Round 1; Round 10 had severe quality control problems and the hope was we could recover some of the missing data from Round 10, and finally, the importance of the ASVAB and AFQT data argued for an attempt to find the original score sheets. NORC was able to find the ASVAB score sheets. They did data entry from those sheets and then forwarded the sheets and data to CHRR.

Based on a search of archival files and an examination of the score sheets, CHRR was able to determine that a test form in its files corresponded to the 1980 ASVAB and that all respondents received the same test. CHRR developed an answer key and then cross-checked the scores in the public use data with the scores derived from the NORC files generated in 2010. There was a significant number of cases where the two sets of answers disagreed, and in those cases we returned to the hard copy for adjudication and constructed revised correct/incorrect indicators based on the answer key and hard copy. This was done for the tests used for the AFQT – Paragraph Comprehension, Word Knowledge, Arithmetic Reasoning and Math Knowledge. These discrepancies suggested data entry errors. The prevalence of discrepancies was higher than would be expected from repeating the same error when 100% verification was done. In some cases, runs of errors suggested machine scoring with a template that was physically misaligned. Consequently, the variables in the newly released AFQT answer file will, for some respondents, show a different number of correct answers than in the original scoring for these tests. The results for MK, AR, WK and PC in the newly-released answer files are, we believe, correct.

Item Response Theory: There are two major approaches to testing: Classical Testing Theory and Item Response Theory (IRT). IRT is the newer approach with origins that can be traced to Frederic Lord[3], Georg Rasch[4] and Paul Lazarsfeld[5]. The IRT approach is computationally demanding and started to become ascendant in the late 1970's and 1980's with the proliferation of computing power. Briefly, Classical Testing Theory looks at tests (scales) in their entirety, for example the number of questions out of 25 answered correctly, or, when there are Likert Scale or other ordinal responses, assigning zero to n points, depending on the number of categories[6], to the various responses, summing these up and norming the total.[7] In contrast, IRT looks at

[3] Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

[4] Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

[5] Lazarsfeld P.F, & Henry N.W. (1968). *Latent Structure Analysis.* Boston: Houghton Mifflin.

[6] Sometimes a different point assignment algorithm is used.

[7] If there are missing items due to respondent error or non-response, one must decide at what point to reject the test results for this person.

individual questions as separate items that differ in their innate difficulty as well as their ability to discriminate among respondents at various points in the underlying scaled continuum. Respondents likewise have innate positions on the scaled continuum or have a "true score". IRT jointly estimates item scores *and* respondent scores (the latter usually referred to as theta scores) via fixed effect estimation. Thus there are question, or "item", fixed effects and respondent fixed effects, with the former being "nuisance parameters"[8] given our primary interest in generating respondent scaled scores, and the latter fixed effects being the scores of interest.

Within the class of IRT models, there are different approaches to parameterizing the item scores – namely one, two and three parameter models. In the simplest one parameter model, only a single parameter is estimated that indicates where, in the distribution of traits, the item has power. For example, in the simple testing model, respondent i has an innate capability θ_i which one may think of as having been drawn from a standard normal distribution. Item j has a difficulty level b_j and the probability that respondent i will answer item j correctly is $F[a(\theta_i - b_j)]$, where F is the logistic cumulative distribution function. The parameter b_j is referred to as the "threshold" or "difficulty" as it is the point at which persons with larger values of θ_i are more likely to endorse the item than not. This specification is referred to as the Rasch model[9] and, because the value of a is set to one for all items, the sum of correct answers is a sufficient statistic for θ_i. Thus the difference in scoring with CTT and IRT when using the one-parameter IRT model with no missing data will likely be minor. When the value of a differs across items, however, a two parameter model is required. The two parameter model has no sufficient statistics, and hence the results from IRT and CTT are more likely to differ.[10]

The item scores, a_j and b_j, measure different facets of the *ability of a question* to place a person on a linear continuum. The value of b_j measures how difficult it is for the respondent to select the right answer. When b_j is large and positive, the respondent must have a substantial degree of true, underlying intellectual acuity to have a 50-50 or better chance of getting the answer correct. The numerical magnitude of a_j, on the other hand, measures the ability of the item to discriminate between varying degrees of intellectual acuity. A large value of a_j indicates the item is a powerful predictor of whether the respondent's acuity is above or below the threshold value. The estimated value of θ for a person scores them in terms of their underlying acuity, where the distribution of θ in the population is assumed standard normal.[11]

---

[8] Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons.

[9] Rasch, G. (1961). On general laws and the meaning of measurement in psychology, pp. 321-334 in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, IV. Berkeley: University of Chicago Press, 1980.

[10] The three parameter model adds a parameter to account for guessing in multiple choice questions for which answers are either correct or incorrect. In psycho-social scales we do not generally think in terms of guessing the correct answer – as there isn't one - hence three parameter models are not applicable. The ASVAB instructs the test taker to answer all questions as they are scored based on the number correct. There is no subtraction for incorrect answers.

[11] The assumptions that θ has a mean of zero and variance equal to one are "normalizing assumptions" to avoid a lack of identification (or observational equivalence) if one adds the same value to b and θ or re-scales a, b and θ.

IRT is highly advantageous for survey applications because scoring depends only on the items that are asked. If the individual items are legitimate indicators of the underlying scale we measure for respondents – which is a maintained hypothesis for both CTT and IRT– then one can solve for θ_i using whichever questions are asked. If a respondent skips an item or respondents are asked different items, it makes no difference. Scoring is direct, albeit requiring iterative estimation in the case of the two parameter model.[12] For CTT, replacing an item in a scale with another having a different value of b_j or dropping the item entirely will change the population distribution of the total score. With IRT a change in items does not have that effect and the scoring of shortened scales is more readily accomplished once the item scores for the full scale have been estimated.

Substantively, we find the rank order correlations between the CTT and IRT scores to be very high, suggesting existing inferences based on CTT scores will likely be little changed by using IRT scores and percentiles. For both methods, percentile scores are generated from the weighted non-parametric distribution of the raw scores.

Unidimensionality: In order to apply IRT to a psychometric scale, a key prerequisite for most analysis is that the various items in the scale be unidimensional, that is, that each item measures a single latent trait, and that latent trait is the same for all items in the scale.[13] The starting point for the analysis is to do factor analysis on the scale items to determine whether there appears to be a single factor. This analysis is usually summarized using a Cattell "scree plot" of the sort in Figure 1 below. These plots show the eigenvalues for the first, second, third, and so forth, factors. This method looks for an "elbow" in the plot where the eigenvalues stop dropping sharply and level out. While this criterion of looking for an "elbow" can be subjective, in the plots below we see a sharp decline, for every test, after the first factor, at which point the plot levels out. This pattern suggests a single underlying trait. We also show the factor loadings for the one and two factor solutions, which are the only plausible solutions revealed by the scree plots. We follow a conventional technique of performing exploratory analysis on part the sample and then confirmatory factor analysis on the suggested model(s) with another part.

The confirmatory factor analyses (CFAs) were conducted in LISREL[14] using polychoric correlations and diagonally weighted least squares estimation, which are appropriate for the analysis of categorical data and result in accurate fit indices[15]. While there are numerous measures of fit, we evaluated model fit by comparing CFA results to the recommended cutoff values of four selected fit

[12] IRT also allows for computer adaptive testing (CAT). This testing approach uses an algorithm to determine which questions should be asked of each respondent to generate the greatest efficiency in ascertaining the estimate of θ_i given their previous responses. In common sense terms, when a respondent gets most of a set of questions correct, one should proceed to ask items that are more difficult, not easier. This is the general approach used with high-stakes educational testing for college admission as well as the method chosen by the U.S. Department of Defense to test potential entrants to the armed forces.

[13] Unidimensionality is an assumption of unidimensional IRT, however, there are also multidimensional and bi-factor IRT models as well, which do not assume unidimensionality. For the analysis here, we focus on the issue of unidimensionality for reasons that will become clear.

[14] Jöreskog, K. G., & Sörbom, D. (2004). LISREL (Version 8.8) [Computer software]. Lincolnwood, Illinois: Scientific Software International.

[15] Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12,* 58-79.

indices: the root mean square error of approximation (RMSEA), the root mean square residual (RMSR), the Tucker-Lewis Index (TLI), and the Comparative Fit Index (CFI).

Values of the RMSEA less than .05 indicate close fit, while values between .05 and .08 suggest fair fit, and values greater than .10 indicate poor fit[16]. Additionally, the cutoff value for the RMSR should be close to .08 or less to imply good fit[17]. Finally, values of .95 or higher indicate good model fit with the TLI and the CFI. We looked for consistency across the four indices to provide stronger evidence for model fit than could be achieved through any single measure of fit.

The item response theory (IRT) analyses[18] were conducted in flexMIRT[19], using unidimensional[20] three parameter logistic (3PL) models for each AFQT component. The reliability plots illustrate the reliability of the IRT scores across the range of each latent construct (i.e., Math Knowledge, Arithmetic Reasoning, Word Knowledge, and Paragraph Comprehension).

If unidimensionality failed and there were (at least) two latent factors for each test, the factor analysis would show some questions loading heavily on a distinct, additional factor. However, if the second factor is highly correlated with the first, that suggests the two factors are very nearly measuring the same latent trait. We start with the two verbal scales – Word Knowledge (WK) and Paragraph Comprehension (PC) and then discuss the two quantitative scales – Arithmetic Reasoning (AR) and Math Knowledge (MK).

We find that WK, PC, AR and MK are all unidimensional, hence the prerequisite for IRT is met for all four scales. We then go further and explore whether all four tests were really necessary to obtain an AFQT score for the NLSY79 respondents. We find that when we combine the questions in Word Knowledge and Paragraph Comprehension, those pooled items are unidimensional. The Paragraph Comprehension items do not have notably high discriminant power, so they are not inherently superior to the simpler Word Knowledge items. These unidimensionality findings suggest when using the AFQT to test respondents for verbal acuity, the most efficient strategy is to drop the Paragraph Comprehension test and use whatever respondent time one is willing to devote to measuring verbal acuity to administering Word Knowledge items.

In the mathematical acuity domain, we similarly find Arithmetic Reasoning and Math Knowledge items are, when combined, unidimensional. As with the verbal scales, there is no need to ask sufficient questions to generate reliable scores on these two mathematics tests separately inasmuch as they measure the same thing. The two sets of items are involve similar time requirements, so there is no reason to drop either test in favor of the other nor is there a compelling reason to retain both. The

---

[16] Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods, and Research, 21*, 230-258; Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

[17] Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

[18] The 2012 release of the item characteristics for the four scales analyzed here (WK, PC, AR and MK) show a "correct/incorrect" indicator for each question-respondent pair as well as the IRT item parameters for each question based on the three parameter model. Thus there is a discriminant parameter, a difficulty parameter and a "guessing" parameter. We also present the theta score for each of the four tests for each respondent.

[19] Cai, L. (2012). flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.

[20] Unidimensionality is an assumption of unidimensional IRT, however, there are also multidimensional and bi-factor IRT models as well, which do not assume unidimensionality.

real payoff to unidimensionality for the mathematics tests is that when pooled they offer many potential items of varying difficulty.[21]

Word Knowledge (n=3710)

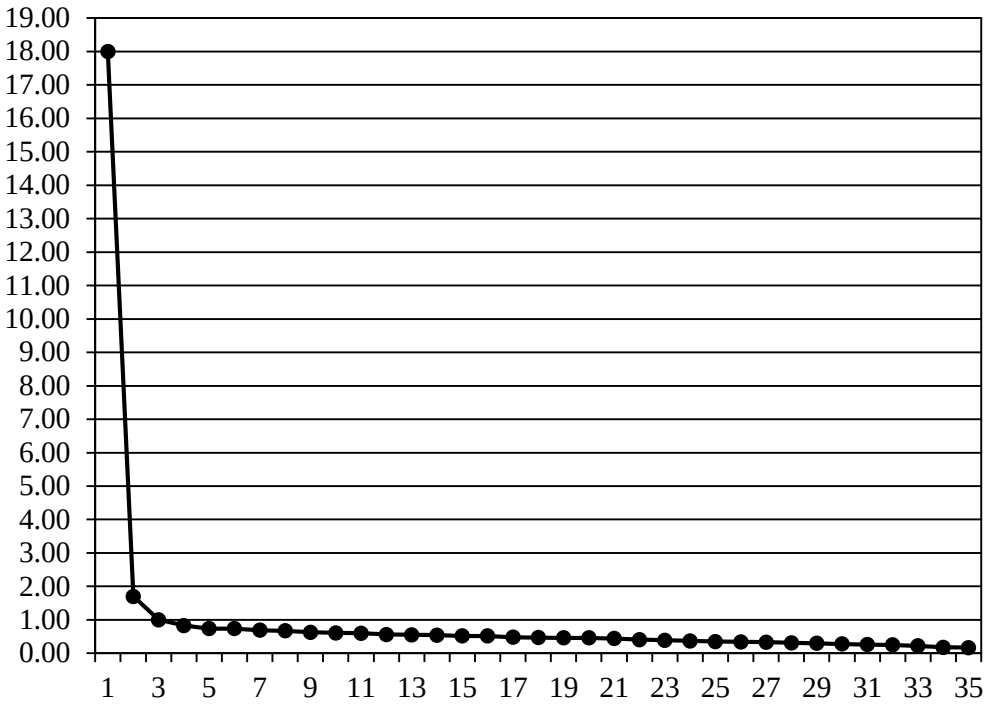

Figure 1 – Scree Plot for Word Knowledge Eigenvalues

Factor Loadings

Table 1 - One Factor Solution for Word Knowledge

Item	Factor Loading
wk1	0.71
wk2	0.70
wk3	0.75
wk4	0.64
wk5	0.69
wk6	0.82
wk7	0.60
wk8	0.78
wk9	0.65
wk10	0.81
wk11	0.78
wk12	0.69
wk13	0.77

[21] For Computer Adaptive Testing, having high variability in item difficulty supports the construction of efficient tests.

wk14	0.69
wk15	0.83
wk16	0.69
wk17	0.79
wk18	0.70
wk19	0.64
wk20	0.77
wk21	0.66
wk22	0.73
wk23	0.70
wk24	0.50
wk25	0.62
wk26	0.75
wk27	0.64
wk28	0.68
wk29	0.52
wk30	0.57
wk31	0.86
wk32	0.52
wk33	0.52
wk34	0.73
wk35	0.79

Table 2 - Two Factor Solution for Word Knowledge

Item	Factor 1	Factor 2
wk1	**0.75**	-0.17
wk2	**0.77**	**-0.34**
wk3	**0.78**	-0.15
wk4	**0.69**	-0.27
wk5	**0.75**	-0.29
wk6	**0.83**	-0.08
wk7	**0.60**	-0.01
wk8	**0.77**	0.02
wk9	**0.66**	-0.04
wk10	**0.83**	-0.13
wk11	**0.80**	-0.09
wk12	**0.72**	-0.19
wk13	**0.77**	-0.01
wk14	**0.68**	0.09
wk15	**0.84**	-0.03
wk16	**0.69**	-0.02
wk17	**0.77**	0.10
wk18	**0.68**	0.10
wk19	**0.62**	0.10
wk20	**0.77**	0.01
wk21	**0.63**	0.14
wk22	**0.69**	0.23
wk23	**0.66**	0.21

wk24	**0.46**	0.24
wk25	**0.58**	0.20
wk26	**0.71**	0.26
wk27	**0.61**	0.14
wk28	**0.66**	0.10
wk29	**0.47**	**0.33**
wk30	**0.53**	0.23
wk31	**0.86**	0.00
wk32	**0.48**	0.22
wk33	**0.45**	**0.43**
wk34	**0.69**	0.20
wk35	**0.73**	**0.32**

Note. Inter-factor correlation is .12. Statistically significant loadings for the two factor model are shown in bold.

Here, the one factor solution seems to best fit the data. The loadings on the first factor are large and significant with only a few items having a significant loading on the second factor with the largest loading for the second factor being smaller than the smallest loading for the first factor. This reflects the sharp elbow in the scree plot after the first factor.

Paragraph Comprehension (n=3625)

Scree Plot of Eigenvalues

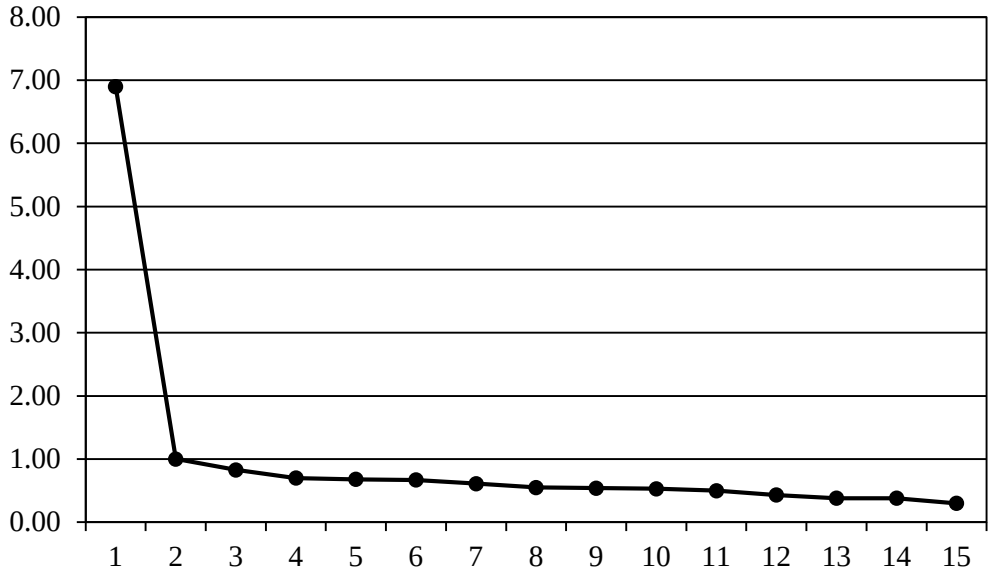

Figure 2 – Scree Plot for Paragraph Comprehension Eigenvalues

Factor Loadings

Table 3 - One Factor Solution for Paragraph Comprehension

Item	Factor Loading
pc1	0.67
pc2	0.68
pc3	0.84
pc4	0.58
pc5	0.75
pc6	0.57
pc7	0.69
pc8	0.58
pc9	0.72
pc10	0.65
pc11	0.52
pc12	0.60
pc13	0.77
pc14	0.67
pc15	0.25

Table 4 –Two Factor Solution for Paragraph Comprehension

Item	Factor 1	Factor 2
pc1	**0.70**	-0.10
pc2	**0.70**	-0.08
pc3	**0.84**	0.01
pc4	**0.61**	-0.08
pc5	**0.73**	0.06
pc6	**0.56**	0.03
pc7	**0.71**	-0.06
pc8	**0.63**	-0.14
pc9	**0.70**	0.06
pc10	**0.68**	-0.07
pc11	**0.47**	0.13
pc12	**0.58**	0.07
pc13	**0.71**	0.20
pc14	**0.63**	0.11
pc15	0.08	**0.55**

Note. Inter-factor correlation is .29.

The one factor solution fits the data best. The weak loading on item 15 as well as the fact that for the two factor model, the second factor is primarily item 15 suggests a problem with this item. This is the only item in the Paragraph Comprehension test for which the pick list offers three alternative meanings for the paragraph with the fourth alternative being “all of the above”. In addition, the discriminant power of item 15 is quite weak. This suggests that when scoring this test, 15 should be dropped. Several staff members with strong verbal abilities reviewed item 15

in the process of generating an answer key. Those lively discussions as to the correct answer reinforced the conclusion this item was problematic.

Arithmetic Reasoning (n=3710)

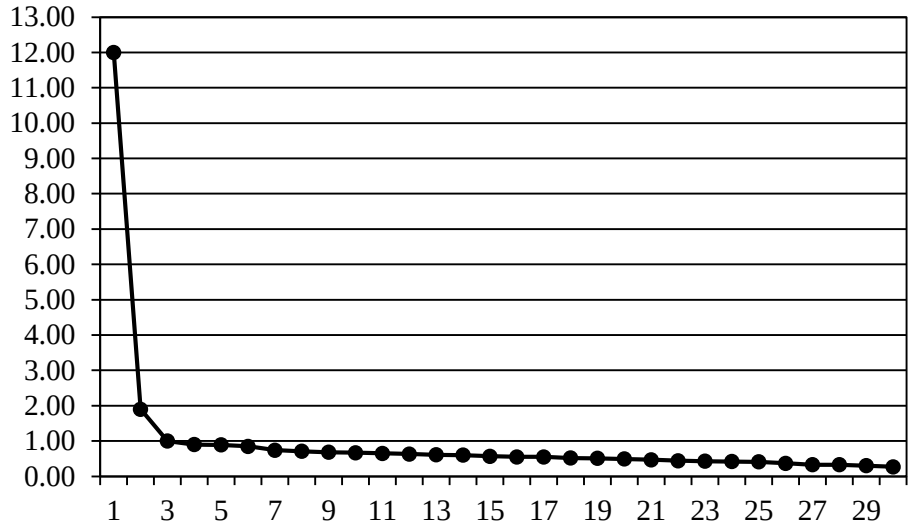

Figure 3 – Scree Plot for Arithmetic Reasoning Eigenvalues

Factor Loadings

Table 5 - One Factor Solution for Arithmetic Reasoning

Item	Factor Loading
ar1	0.39
ar2	0.48
ar3	0.66
ar4	0.76
ar5	0.65
ar6	0.65
ar7	0.51
ar8	0.62
ar9	0.67
ar10	0.77
ar11	0.80
ar12	0.67
ar13	0.61
ar14	0.70
ar15	0.68
ar16	0.56
ar17	0.50
ar18	0.54
ar19	0.59
ar20	0.68

ar21	0.56
ar22	0.62
ar23	0.54
ar24	0.57
ar25	0.65
ar26	0.62
ar27	0.77
ar28	0.57
ar29	0.52
ar30	0.58

Table 6 - Two Factor Solution for Arithmetic Reasoning

Item	Factor 1	Factor 2
ar1	**0.58**	-0.18
ar2	**0.71**	-0.21
ar3	**0.64**	0.06
ar4	**0.83**	-0.03
ar5	**0.54**	0.15
ar6	**0.78**	-0.09
ar7	**0.54**	0.00
ar8	**0.60**	0.06
ar9	**0.63**	0.08
ar10	**0.70**	0.12
ar11	**0.75**	0.09
ar12	**0.54**	0.18
ar13	**0.46**	0.19
ar14	**0.49**	0.26
ar15	**0.52**	0.20
ar16	0.21	**0.39**
ar17	0.20	**0.34**
ar18	-0.03	**0.63**
ar19	**0.37**	0.26
ar20	0.15	**0.60**
ar21	-0.07	**0.70**
ar22	0.08	**0.60**
ar23	0.09	**0.49**
ar24	-0.02	**0.65**
ar25	0.11	**0.60**
ar26	0.20	**0.47**
ar27	**0.31**	**0.53**
ar28	-0.09	**0.73**
ar29	0.01	**0.57**
ar30	0.07	**0.57**

Note. Inter-factor correlation is .72.

Although the pattern of factor loadings for the two factor solution appears to provide evidence for the plausibility of the arithmetic reasoning items measuring two separate factors, the high inter-factor correlation of .72, and scree plot of eigenvalues lead us to believe there is a single

underlying factor. The questions loading on the second factor involve rates and respondents with stronger algebra skills will be at an advantage.

Math Knowledge (n=3629)

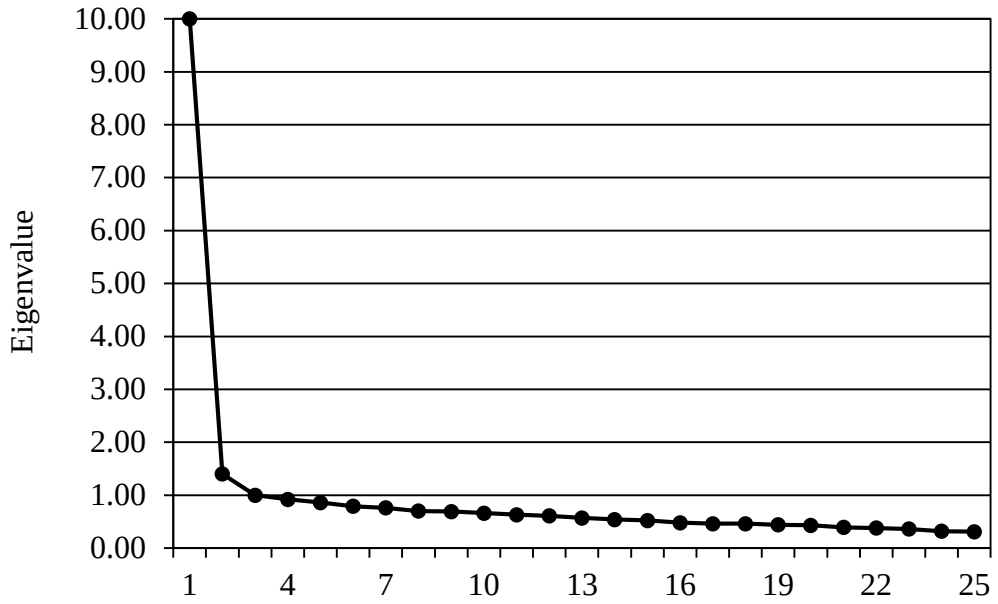

Figure 4 – Scree Plot for Math Knowledge Eigenvalues

Table 7 - One Factor Solution for Math Knowledge

Item	Factor Loading
MK1	0.60
MK2	0.48
MK3	0.68
MK4	0.60
MK5	0.51
MK6	0.72
MK7	0.62
MK8	0.62
MK9	0.65
MK10	0.62
MK11	0.56
MK12	0.69
MK13	0.75
MK14	0.77
MK15	0.30
MK16	0.64
MK17	0.76
MK18	0.72
MK19	0.54

MK20	0.52
MK21	0.53
MK22	0.60
MK23	0.68
MK24	0.57
MK25	0.64

Table 8 - Two Factor Solution for Math Knowledge

Item	Factor 1	Factor 2
MK1	-0.12	0.83
MK2	0.01	0.54
MK3	0.09	0.68
MK4	0.34	0.31
MK5	0.06	0.52
MK6	0.33	0.46
MK7	0.35	0.33
MK8	0.18	0.51
MK9	0.30	0.41
MK10	0.29	0.40
MK11	0.36	0.24
MK12	0.42	0.33
MK13	0.47	0.35
MK14	0.48	0.37
MK15	0.36	-0.05
MK16	0.57	0.11
MK17	0.60	0.22
MK18	0.60	0.18
MK19	0.34	0.26
MK20	0.43	0.13
MK21	0.64	-0.07
MK22	0.75	-0.11
MK23	0.76	-0.04
MK24	0.56	0.05
MK25	0.80	-0.12

Note. Inter-factor correlation is .67.

The items that load on the second factor articulate the problem in verbal terms, with the items loading on the first factor requiring relatively less verbal acuity. Whether an item taps into algebra or geometry does not appear to explain the factor loading pattern. The large inter-factor correlation suggests the two factors are measuring the same domain. The scree plot shows the second factor explains comparatively little of the variance. A one factor model seems the best explanation for MK as well.

Pooling Verbal Acuity Measures: Next, we test whether pooling the Word Knowledge and Paragraph Comprehension items yields a single unidimensional battery. The results suggest that is the case, with item 15 in PC being discrepant. In a two factor solution, we find each item in

both WK and PC load more strongly on the first factor. The apparent fact that the two tests measure the same underlying factor suggests an efficient alternative for computing an AFQT is to simply drop the difficult-to-administer PC items and focus on the WK test. With 35 items in the WK battery, a computer adaptive test that asks progressively harder items the better the respondent does, and easier items to respondents missing more of the initial items, should be able to achieve good precision over a range of aptitudes while asking fewer items than the 50 items in the combined PC and WK batteries. Contemporaneous research[22] on the Paragraph Comprehension test showed it was time intensive. Figure 5 shows a sharp elbow in the scree plot at the second eigenvalue with item 15 of PC again showing the smallest factor loading of all verbal items. When we estimate a three factor model, the third factor is primarily item 15 on the PC test.

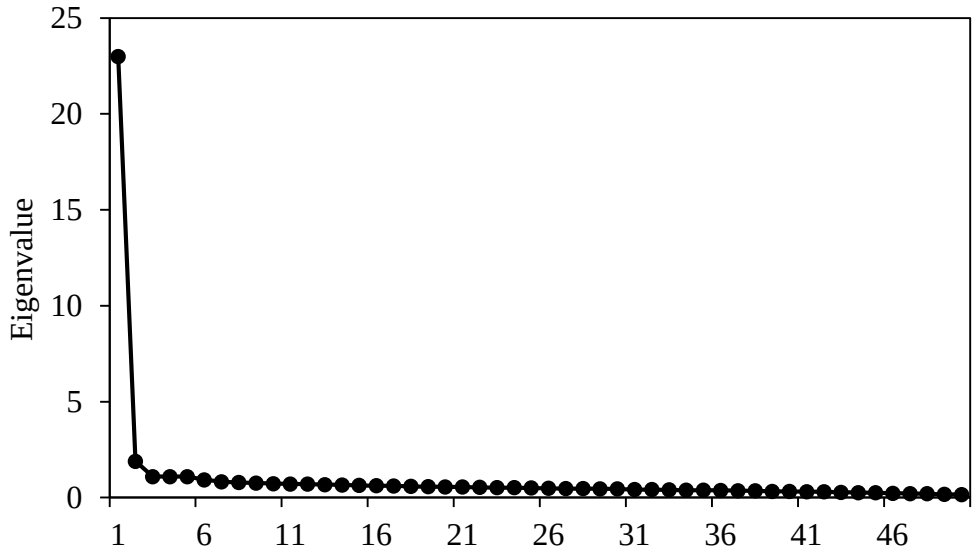

Figure 5 – Eigenvalues for Combined Word Knowledge and Paragraph Comprehension Items

Table 9 - Factor Analysis Summary for Combined Word Knowledge and Paragraph Comprehension Items

Item	Lambda
W1	0.73
W2	0.73
W3	0.76
W4	0.66
W5	0.71
W6	0.80
W7	0.63
W8	0.77
W9	0.64
W10	0.80

[22] Segall, Daniel O., Moreno, Kathleen E., Kieckhaefer, William F., Vicino, Frank L. and McBride, James R. "Validation of the Experimental CAT-ASVAB System" in *Computerized Adaptive Testing – From Inquiry to Operation*, William A. Sands, Brian K. Waters and James R. McBride, eds. American Psychological Association, 1997, pp. 103-114.

W11	0.77
W12	0.73
W13	0.78
W14	0.68
W15	0.80
W16	0.71
W17	0.79
W18	0.70
W19	0.64
W20	0.77
W21	0.63
W22	0.71
W23	0.69
W24	0.48
W25	0.62
W26	0.74
W27	0.62
W28	0.67
W29	0.50
W30	0.58
W31	0.85
W32	0.50
W33	0.54
W34	0.70
W35	0.78
P1	0.66
P2	0.68
P3	0.80
P4	0.56
P5	0.71
P6	0.56
P7	0.65
P8	0.51
P9	0.66
P10	0.63
P11	0.47
P12	0.54
P13	0.75
P14	0.64
P15	0.26

Note. $n = 3650$

Table 10 – Factor Loading for a Two Factor Model of the Word Knowledge and Paragraph Comprehension Items

Item	Factor 1	Factor 2
W1	0.74	-0.23
W2	0.74	-0.36
W3	0.76	-0.20
W4	0.67	-0.27
W5	0.72	-0.38
W6	0.80	-0.07
W7	0.63	-0.07
W8	0.77	0.00
W9	0.64	0.03
W10	0.81	-0.20
W11	0.77	0.00

W12	0.74	-0.19
W13	0.78	-0.03
W14	0.68	0.11
W15	0.80	0.05
W16	0.71	0.03
W17	0.79	0.10
W18	0.70	0.05
W19	0.64	0.14
W20	0.77	-0.02
W21	0.63	0.09
W22	0.70	0.25
W23	0.69	0.24
W24	0.48	0.24
W25	0.62	0.21
W26	0.73	0.24
W27	0.62	0.15
W28	0.67	0.03
W29	0.49	0.29
W30	0.58	0.18
W31	0.85	-0.05
W32	0.50	0.14
W33	0.54	0.38
W34	0.70	0.16
W35	0.78	0.30
P1	0.66	-0.04
P2	0.68	-0.12
P3	0.80	-0.04
P4	0.56	-0.03
P5	0.71	0.00
P6	0.56	0.13
P7	0.65	-0.04
P8	0.51	-0.27
P9	0.66	-0.07
P10	0.63	0.13
P11	0.47	-0.17
P12	0.54	0.07
P13	0.75	-0.02
P14	0.64	-0.01
P15	0.26	-0.02

Note. n = 3650 and r = -.01

Table 11 – Factor Loadings from a Three Factor Solution of the Combined Word Knowledge and Paragraph Comprehension Items

Item	Factor 1	Factor 2	Factor 3
W1	0.66	-0.19	0.18
W2	0.65	-0.30	0.21
W3	0.71	-0.18	0.12
W4	0.61	-0.24	0.14
W5	0.62	-0.31	0.23
W6	0.78	-0.07	0.05
W7	0.60	-0.05	0.09
W8	0.78	-0.05	-0.05
W9	0.66	-0.01	-0.05
W10	0.75	-0.17	0.14
W11	0.78	-0.05	-0.04

W12	0.70	-0.19	0.07
W13	0.78	-0.07	-0.02
W14	0.71	0.05	-0.08
W15	0.81	0.01	-0.03
W16	0.70	0.02	0.03
W17	0.81	0.05	-0.06
W18	0.70	0.03	0.01
W19	0.68	0.08	-0.10
W20	0.74	-0.02	0.06
W21	0.61	0.10	0.07
W22	0.73	0.21	-0.05
W23	0.73	0.18	-0.10
W24	0.51	0.20	-0.06
W25	0.63	0.19	-0.01
W26	0.75	0.23	0.00
W27	0.65	0.10	-0.07
W28	0.64	0.06	0.10
W29	0.50	0.29	0.01
W30	0.56	0.21	0.08
W31	0.76	0.02	0.24
W32	0.47	0.18	0.11
W33	0.55	0.38	0.02
W34	0.71	0.14	0.00
W35	0.81	0.27	-0.05
P1	0.70	-0.12	-0.13
P2	0.69	-0.16	-0.04
P3	0.79	-0.07	0.01
P4	0.59	-0.08	-0.07
P5	0.71	-0.03	-0.01
P6	0.60	0.08	-0.08
P7	0.69	-0.11	-0.10
P8	0.54	-0.35	-0.11
P9	0.65	-0.09	0.02
P10	0.68	0.06	-0.13
P11	0.42	-0.13	0.13
P12	0.55	0.04	-0.03
P13	0.71	0.00	0.11
P14	0.60	0.01	0.10
P15	0.06	0.26	0.64

Note. n = 3650, r_{12} = -.03, r_{13} = .32, and r_{23} = -.20.

Table 12 – Confirmatory Analyses of One Factor Model for Verbal Acuity

Item	Lambda
W1	0.73
W2	0.74
W3	0.78
W4	0.63
W5	0.72
W6	0.82
W7	0.61
W8	0.77
W9	0.65
W10	0.79
W11	0.77
W12	0.67
W13	0.78

W14	0.69
W15	0.81
W16	0.66
W17	0.79
W18	0.69
W19	0.63
W20	0.75
W21	0.65
W22	0.72
W23	0.66
W24	0.51
W25	0.62
W26	0.75
W27	0.63
W28	0.68
W29	0.50
W30	0.56
W31	0.85
W32	0.52
W33	0.54
W34	0.74
W35	0.80
P1	0.68
P2	0.66
P3	0.79
P4	0.55
P5	0.74
P6	0.53
P7	0.65
P8	0.53
P9	0.70
P10	0.61
P11	0.49
P12	0.57
P13	0.71
P14	0.68
P15	0.25

Note. n = 3684

Indices of Fit for a One-Factor Confirmatory Factor Analysis Model for the Combined Word Knowledge and Paragraph Comprehension Data

Index	Value
Root Mean Square Error of Approximation (RMSEA)	.03
90% Confidence Interval for RMSEA	.03-.03
Tucker-Lewis Index	1.00
Comparative Fit Index	1.00
Root Mean Square Residual	.04

Pooling Quantitative Measures: While the MK and AR test items are not as complex or lengthy to administer as the PC items, the two tests contain 55 items and these tests were designed to generate reliable results on each scale within the classical testing framework. The scree plot resulting when we pool the two sets of items suggests a single factor. In a two factor model 9 of 25 items in MK load more heavily on the first factor and 16 of 30 items in AR load more heavily on the first factor. A two factor model on the pooled data might have been expected to show the two factor loadings corresponding to the two separate tests, which is not the case. However, the estimated correlation between the two latent factors is high, suggesting there is really a single factor at work for both sets of items. Viewing the two batteries as measuring the same latent construct appears consistent with the evidence, arguing in favor of using the combined battery of items as a resource to measure a latent "mathematical acuity" factor.

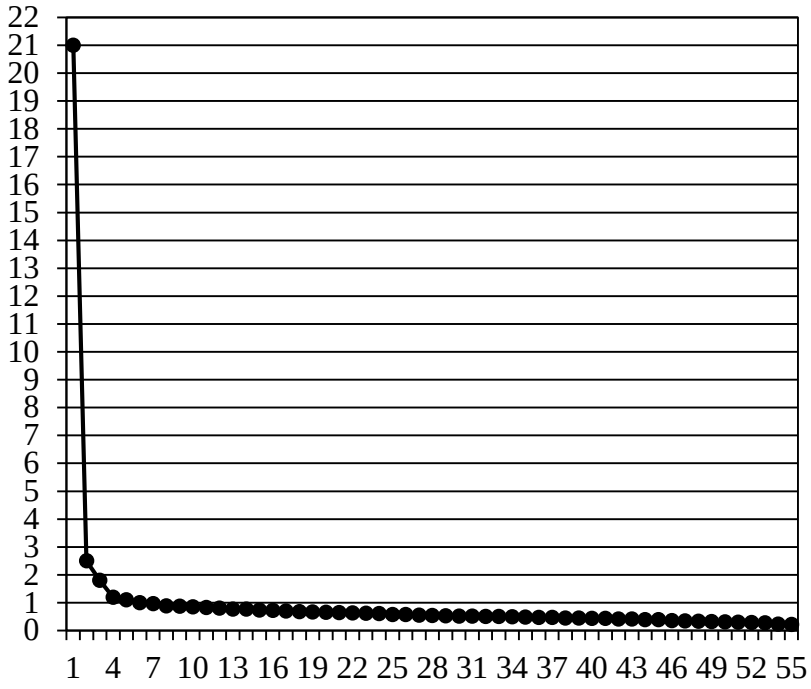

Figure 6 - Scree Plot of Eigenvalues from an Exploratory Factor Analysis of the Combined Math Knowledge and Arithmetic Reasoning Data

Table 13 - Factor Loadings from Three Exploratory Factor Analysis Models for the Combined Math Knowledge and Arithmetic Reasoning Data

	One-Factor	Two-Factor Model		Three-Factor Model		
Item	Model	Factor 1	Factor 2	Factor1	Factor 2	Factor 3
MK1	0.66	0.85	-0.13	0.81	-0.15	0.07
MK2	0.51	0.53	0.01	0.51	-0.03	0.09
MK3	0.75	0.68	0.13	0.65	0.11	0.08
MK4	0.55	0.31	0.29	0.30	-0.05	0.41
MK5	0.50	0.58	-0.04	0.56	-0.13	0.13
MK6	0.69	0.46	0.29	0.45	-0.04	0.40
MK7	0.62	0.31	0.35	0.30	0.03	0.39

MK8	0.63	0.46	0.21	0.44	0.12	0.15
MK9	0.61	0.36	0.30	0.35	0.00	0.37
MK10	0.58	0.37	0.25	0.36	-0.04	0.36
MK11	0.52	0.22	0.35	0.21	0.02	0.39
MK12	0.70	0.29	0.47	0.29	0.18	0.36
MK13	0.68	0.27	0.47	0.27	0.06	0.50
MK14	0.77	0.29	0.55	0.28	0.40	0.22
MK15	0.31	0.02	0.32	0.03	0.06	0.30
MK16	0.63	0.16	0.53	0.16	0.22	0.38
M17	0.72	0.24	0.54	0.23	0.07	0.57
MK18	0.66	0.11	0.60	0.11	0.12	0.57
MK19	0.55	0.14	0.45	0.14	0.21	0.31
MK20	0.50	0.00	0.53	0.01	0.31	0.28
MK21	0.47	-0.08	0.59	-0.07	0.13	0.54
MK22	0.54	-0.06	0.65	-0.07	0.04	0.73
MK23	0.63	-0.14	0.83	-0.12	0.52	0.38
MK24	0.56	-0.03	0.64	-0.02	0.45	0.25
MK25	0.60	-0.09	0.75	-0.07	0.37	0.46
AR1	0.38	0.51	-0.10	0.48	-0.07	0.00
AR2	0.45	0.69	-0.20	0.65	-0.09	-0.08
AR3	0.65	0.64	0.06	0.61	0.08	0.03
AR4	0.73	0.81	-0.02	0.78	0.08	-0.04
AR5	0.63	0.47	0.20	0.46	0.22	0.04
AR6	0.65	0.72	-0.02	0.69	-0.07	0.10
AR7	0.47	0.55	-0.04	0.52	0.09	-0.11
AR8	0.59	0.63	0.00	0.61	0.12	-0.07
AR9	0.66	0.64	0.08	0.61	0.06	0.07
AR10	0.74	0.69	0.10	0.66	0.14	0.02
AR11	0.79	0.78	0.07	0.75	0.09	0.05
AR12	0.64	0.59	0.11	0.56	0.20	-0.04
AR13	0.60	0.43	0.21	0.42	0.24	0.02
AR14	0.71	0.44	0.32	0.43	0.26	0.12
AR15	0.64	0.52	0.16	0.50	0.18	0.04
AR16	0.54	0.28	0.31	0.27	0.32	0.03
AR17	0.46	0.24	0.26	0.23	0.42	-0.13
AR18	0.53	0.03	0.55	0.04	0.46	0.14
AR19	0.59	0.41	0.22	0.40	0.24	0.03
AR20	0.68	0.21	0.52	0.21	0.51	0.07
AR21	0.53	0.07	0.50	0.07	0.56	-0.01
AR22	0.61	0.16	0.50	0.16	0.54	0.01
AR23	0.52	0.08	0.49	0.09	0.43	0.11
AR24	0.54	-0.05	0.64	-0.04	0.60	0.09
AR25	0.64	0.13	0.57	0.13	0.58	0.05
AR26	0.60	0.29	0.36	0.28	0.39	0.02
AR27	0.73	0.38	0.41	0.37	0.48	-0.02
AR28	0.56	-0.02	0.63	-0.02	0.65	0.05

AR29	0.51	-0.02	0.57	-0.01	0.49	0.14
AR30	0.55	0.13	0.46	0.13	0.54	-0.03

Note. For the two-factor model, r_{12} = .71 and for the three-factor model r_{12} = .62 , r_{13} = .57 and r_{23} = .59

Table 14 - Factor Loadings from a One-Factor Confirmatory Factor Analysis of the Combined Math Knowledge and Arithmetic Reasoning Data

Item	Lambda
MK1	0.67
MK2	0.50
MK3	0.75
MK4	0.58
MK5	0.51
MK6	0.68
MK7	0.59
MK8	0.64
MK9	0.60
MK10	0.57
MK11	0.52
MK12	0.68
MK13	0.69
MK14	0.76
MK15	0.32
MK16	0.63
MK17	0.73
MK18	0.63
MK19	0.53
MK20	0.54
MK21	0.48
MK22	0.58
MK23	0.62
MK24	0.57
MK25	0.61
AR1	0.42
AR2	0.48
AR3	0.66
AR4	0.75
AR5	0.62
AR6	0.65
AR7	0.48
AR8	0.60
AR9	0.68
AR10	0.75
AR11	0.80
AR12	0.67

AR13	0.60
AR14	0.70
AR15	0.66
AR16	0.56
AR17	0.48
AR18	0.52
AR19	0.60
AR20	0.66
AR21	0.60
AR22	0.62
AR23	0.53
AR24	0.60
AR25	0.65
AR26	0.59
AR27	0.74
AR28	0.59
AR29	0.53
AR30	0.55

Note. n=3732

Indices of Fit for a One-Factor Confirmatory Factor Analysis Model for the Combined Math Knowledge and Arithmetic Reasoning Data

Index	Value
Root Mean Square Error of Approximation (RMSEA)	.04
90% Confidence Interval for RMSEA	.04-.04
Tucker-Lewis Index	.99
Comparative Fit Index	.99
Root Mean Square Residual	.05

Item Parameters: In this section we estimate the item scores for the four AFQT components. While the fit for the two and three parameter models is not appreciably different we estimate the three parameter model, which allows for guessing, in order to generate difficulty scores for the test items that match the scaling of the theta scores for respondents. We estimate item parameters for the full sample. The *a* parameter is discriminant power, *b* is the difficulty parameter, and *g* is the "guessing" parameter from the three parameter logit (3PL) model.

The item scores show few items that are difficult. The reliability of the scales is weak at the top of the distribution, but recruits scoring near the top will qualify for enlistment, making reliability in the upper tail moot. However, for quantitative acuity, we see MK has relatively better resolution at the upper end of the distribution with AR relatively better at the lower end. This confirms the advantage of pooling the two scales and using adaptive testing to administer more items at the end of the distribution relevant for a particular respondent.

Table 15 - Summary of Item Response Theory Results for Math Knowledge (n=11,153)

Item	*a*	*b*	*g*
MK1	2.24	-0.99	0.06
MK2	1.19	-0.89	0.09
MK3	1.88	-0.15	0.03
MK4	3.32	0.22	0.40
MK5	1.17	-0.71	0.03
MK6	2.29	0.09	0.10
MK7	3.36	0.27	0.34
MK8	1.62	0.11	0.12
MK9	2.91	0.24	0.32
MK10	2.04	0.25	0.21
MK11	2.02	0.58	0.24
MK12	2.61	0.51	0.14
MK13	3.89	0.42	0.19
MK14	3.13	0.61	0.11
MK15	1.81	1.51	0.28
MK16	4.06	0.65	0.26
MK17	3.29	0.60	0.11
MK18	4.42	0.63	0.21
MK19	2.22	0.69	0.26
MK20	2.66	0.97	0.25
MK21	3.39	1.06	0.23
MK22	4.15	1.00	0.17
MK23	4.46	1.06	0.13
MK24	2.64	1.18	0.13
MK25	4.20	1.10	0.12

Note. The rank order correlation between summed scores and θ scores is $r = .97$.

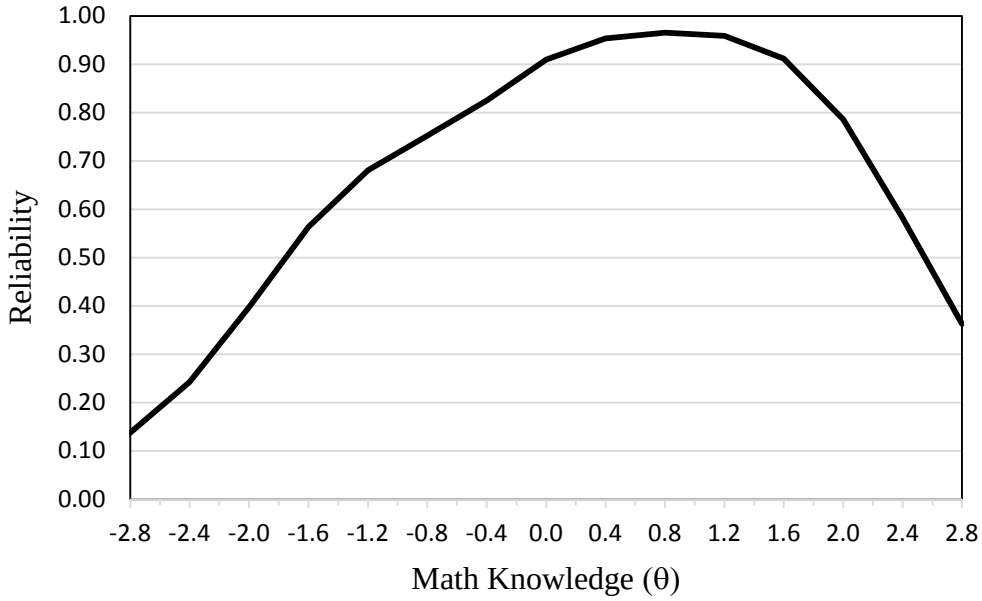

Figure 7 – Reliability Plot for Math Knowledge

Table 16 - Summary of Item Response Theory Results for Arithmetic Reasoning (n = 11,171)

Item	*a*	*b*	*g*
AR1	1.01	-2.45	0.08
AR2	1.36	-1.96	0.08
AR3	2.50	-0.31	0.28
AR4	2.91	-0.27	0.11
AR5	2.16	0.05	0.24
AR6	1.72	-0.53	0.06
AR7	1.25	-0.63	0.22
AR8	2.07	-0.25	0.27
AR9	2.18	-0.15	0.17
AR10	3.10	-0.09	0.15
AR11	2.96	-0.09	0.06
AR12	2.24	0.11	0.16
AR13	2.37	0.34	0.27
AR14	2.90	0.28	0.17
AR15	2.32	0.06	0.19
AR16	2.74	0.52	0.34
AR17	2.18	0.77	0.29
AR18	4.85	0.82	0.28
AR19	1.68	0.33	0.14
AR20	3.13	0.90	0.12
AR21	3.48	0.92	0.21
AR22	3.47	0.65	0.24

AR23	3.24	0.84	0.27
AR24	5.34	0.85	0.23
AR25	4.07	0.78	0.18
AR26	2.33	0.76	0.18
AR27	2.26	0.75	0.05
AR28	3.56	1.16	0.14
AR29	3.50	1.26	0.15
AR30	2.58	1.00	0.18

Note. The rank order correlation between summed scores and IRT θ scores is r = .98.

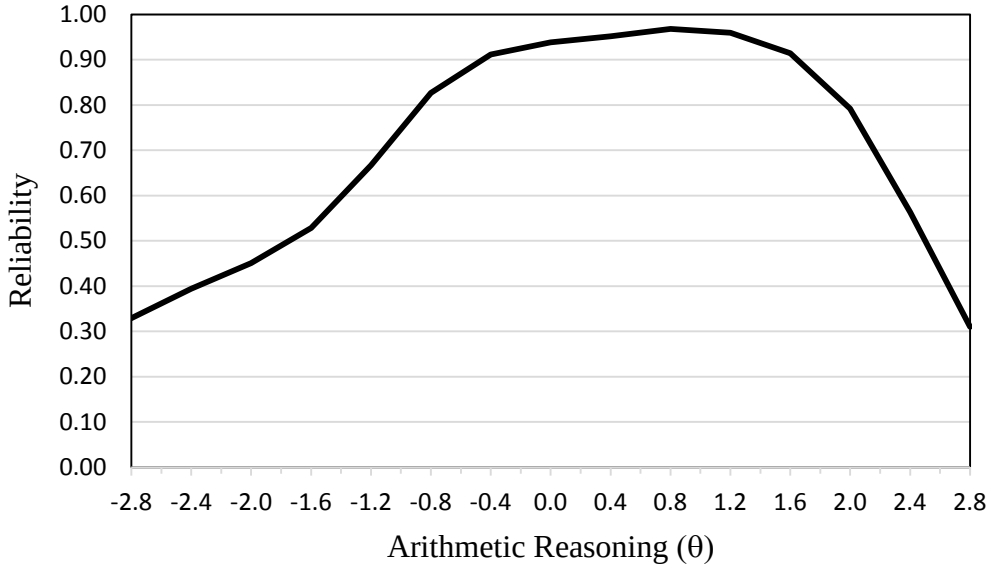

Figure 8 – Reliability Plot for Arithmetic Reasoning

Table 17 - Summary of Item Response Theory Results for Word Knowledge (n = 11,160)

Item	*a*	*b*	*g*
WK1	2.12	-1.72	0.15
WK2	2.03	-1.81	0.02
WK3	2.41	-1.26	0.18
WK4	1.42	-1.71	0.02
WK5	1.43	-1.97	0.00
WK6	3.50	-0.68	0.23
WK7	1.81	-1.00	0.36
WK8	3.82	-0.57	0.32
WK9	1.55	-0.33	0.08
WK10	2.74	-1.15	0.17
WK11	2.49	-0.64	0.13
WK12	1.65	-1.13	0.03
WK13	3.29	-0.63	0.27

WK14	2.52	-0.26	0.24
WK15	2.93	-0.44	0.11
WK16	2.25	-0.56	0.30
WK17	3.72	-0.22	0.19
WK18	2.72	-0.30	0.28
WK19	1.92	-0.12	0.20
WK20	2.57	-0.57	0.19
WK21	2.17	-0.27	0.28
WK22	4.18	0.04	0.26
WK23	2.51	0.06	0.19
WK24	2.34	0.52	0.31
WK25	1.90	0.29	0.16
WK26	3.64	0.12	0.18
WK27	1.74	0.19	0.13
WK28	1.60	-0.26	0.06
WK29	2.03	0.65	0.21
WK30	1.66	0.36	0.14
WK31	2.72	-0.72	0.03
WK32	1.44	0.71	0.12
WK33	3.56	0.74	0.22
WK34	2.02	0.22	0.07
WK35	3.59	0.29	0.08

Note. The rank order correlation between summed scores and IRT θ scores is r = .99.

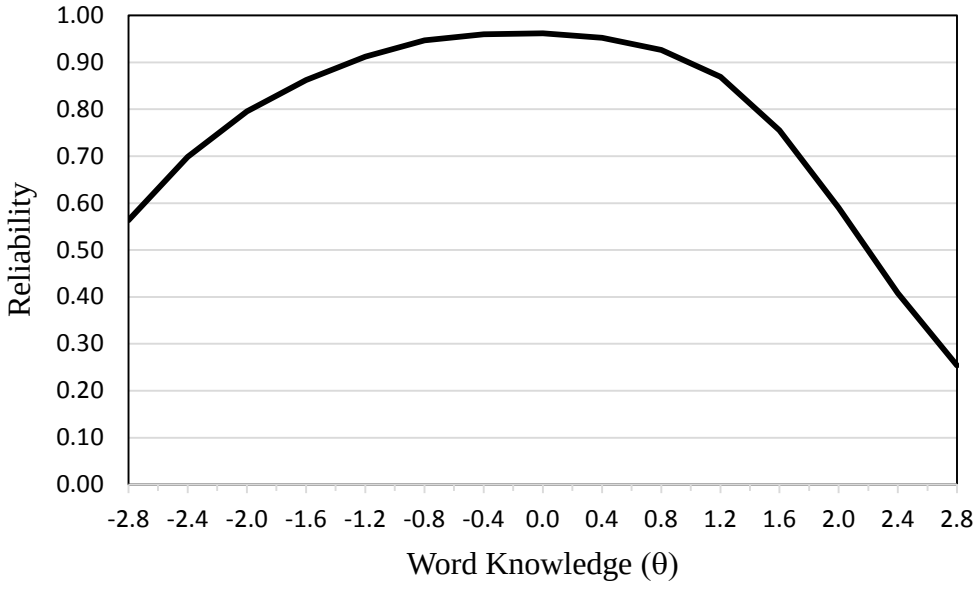

Figure 9 – Reliability Plot for Word Knowledge

Table 18 - Summary of Item Response Theory Results for Paragraph Comprehension (n = 11,151)

Item	*a*	*b*	*g*
PC1	2.12	-0.49	0.23
PC2	2.23	-0.83	0.35
PC3	5.16	-0.49	0.27
PC4	1.66	-0.32	0.25
PC5	2.28	-0.36	0.10
PC6	1.72	-0.05	0.32
PC7	1.69	-0.50	0.12
PC8	1.09	-1.10	0.01
PC9	1.84	-0.65	0.08
PC10	1.82	0.34	0.10
PC11	1.00	-0.97	0.02
PC12	1.50	0.09	0.13
PC13	2.20	-0.63	0.12
PC14	1.50	-0.18	0.03
PC15	0.40	0.49	0.04

Note. The rank order correlation between summed scores and IRT θ scores is r = .99.

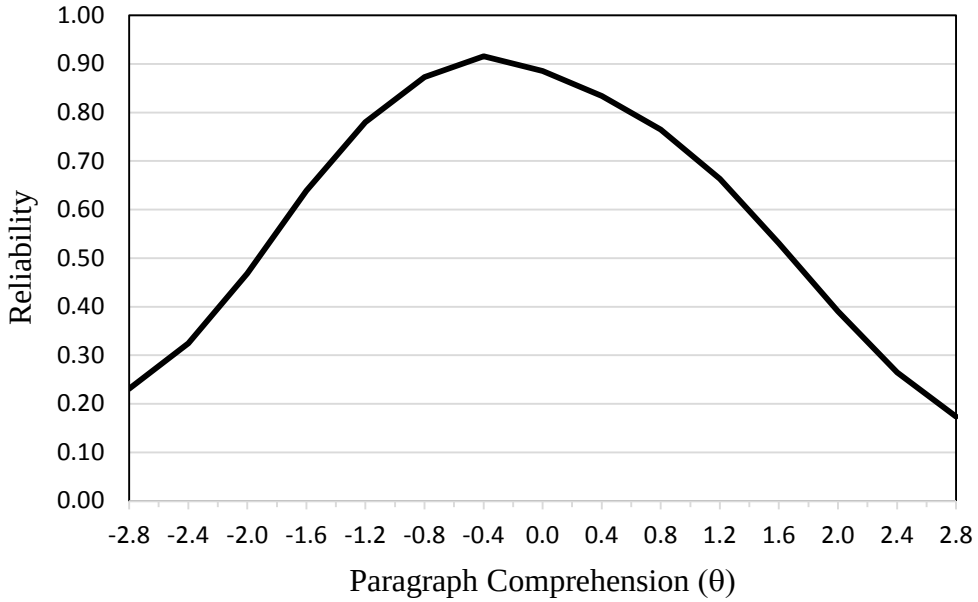

Figure 10 – Reliability Plot for Paragraph Comprehension

Summary: The psychometric situation with the AFQT is not what we expected to find. While there may have been other reasons for keeping PC and WK as separate scales as well as keeping AR and MK separate, there is a strong case that the two verbal and two quantitative tests measure the same latent trait. A two factor solution for both the verbal and quantitative tests would not yield factors that coincide with the respective test boundaries.

PART II – NORMING

One issue identified in the factor analysis was the questionable value of item 15 in Paragraph Comprehension. Above, we showed the factor loadings assigned a much smaller weight to item 15. For persons getting fourteen items correct on Paragraph Comprehension, the most frequently missed question was number 15. Despite this evidence that item was questionable, when norming Paragraph Comprehension, we did not discard the results for item 15. That was the only item among the four current AFQT scales (Arithmetic Reasoning, Math Knowledge, Paragraph Comprehension and Word Knowledge) that we considered discarding.

We used the software package flexMIRT to estimate both two and three parameter logistic models (abbreviated 2PL and 3PL). The two parameter logistic model posits a slope parameter as well as a difficulty parameter for each question. The three parameter logistic model adds a third parameter for the probability the respondent guesses the correct answer when they don't know the correct answer. The central focus for most users is on the normed score for respondents for the various tests. The actual questions will not be released, nor will the set of potential answers. Given interest in how respondents' mental acuity in their late teens translates into labor market success and other outcomes, our focus is on scoring each respondent's performance on the various tests.

In norming these tests, we follow recent practice of taking respondent age into account. Respondents were sampled to fall into the birth years 1957-64, so there was an eight year age spread.[23] When administered in 1980, some respondents might not have entered high school and others will have graduated from college. This spread in potential education would be highly likely to influence testing outcomes. Before the item data were recovered and only raw scores on each test were available, analysis showed a pronounced age gradient in Armed Forces Qualifying Test (AFQT) scores. To correct for this, we grouped the data into four-month age intervals for each birth year (January-April, May-August, September-December) and normed scores with these four-month intervals. The estimated theta scores were ranked within each interval, ASVAB sampling weights attached, and a non-parametric distribution function calculated. These estimated probabilities were transformed to standard normal *z* scores. In some age cohorts, many respondents got all the questions right. To deal with these tied scores, we computed the expected value of a standard normal variable given the probability in the upper tail of the distribution.[24]

By breaking the norming step into 24 cohorts, we avoid the risk that some might conclude younger respondents have less native ability than the cohorts that are older. Unlike the Armed Services that test candidates of reasonably uniform age and education, the NLSY79 respondents are not uniform in either. Users should be aware that the ASVAB tests were normed to meet the needs of the Department of Defense in screening potential recruits who are in their late teens, whereas users may have other analytic needs.

---

[23] The date of birth questions were re-asked in 1981 and five persons gave dates of birth outside the designed range. These cases were grouped with the oldest or youngest cohorts, as appropriate.

[24] When more than seven respondents in an age cohort had the same theta score, we computed the expected value of a standard normal variate given the fractile in which those respondents were located. This was most common for Paragraph Comprehension, which has only 15 items. Ties were more frequent in older cohorts for which a higher fraction of respondents had very high scores.

In examining the norms generated by the 2PL, the 3PL and raw scores, the 2PL norms had a slightly higher correlation with the raw scores (see correlations between Arithmetic Reasoning 2PL and 3PL norms below). In addition, the 3PL generated instances of respondents being ranked higher even though they got three or four fewer questions right. Such "reversals" were less common and smaller for the 2PL.Norms were generated using the 3PL theta scores. Users wishing to generate alternative norms can avail themselves of the item data.

Table 19 – Rank-Order Correlations Between Raw Score and 2PL and 3PL Norms for Arithmetic Reasoning (n = 11,171)

	2PL	3PL
Raw Score	0.98	0.97
2PL		0.99

In working with the data, the score sheets revealed that respondents often skipped items within a battery. The three parameter model assumes respondents guess on items they don't know. The original raw scores for the tests counted the number of items correct without regard for whether the respondent may have missed an item because they gave the incorrect answer or whether they provided no answer for the item. For the IRT analyses used to generate the 3PL theta scores, both skipped and incorrect items were coded as '0'. In the individual item variables provided for users, skipped items are coded as '-4', so that users have the option to recode skipped scores as missing instead of as incorrect.

One of the interesting results that emerged from the factor analysis was the apparent unidimensionality of pooled Paragraph Comprehension and Word Knowledge, as well as Arithmetic Reasoning and Math Knowledge. We generated two new scores – a verbal and a quantitative score, respectively, based on this pooling. These pooled tests should give better resolution of these two domains. Alternatively, users may want to use one of the paired tests as an instrumental variable for the other as a method of dealing with attenuation bias resulting from the fact these tests measure the latent construct with error. The item parameters are little affected whether the generated individually test by test or when pooling pairs of tests (i.e. Arithmetic Reasoning and Math Knowledge).