Constructing AFQT Scores that are Comparable Across the NLSY79 and the NLSY97

Joseph G. Altonji
Prashant Bharadwaj
Fabian Lange

August 2009

Introduction

Social and behavioral scientists routinely use and analyze cognitive ability measures. Examples include studies that consider the relation between cognitive ability measures and labor market outcomes, crime, health, marriage, and savings and portfolio behavior. Other studies use cognitive ability measures to control for pre-existing ability differences. For example, this is a common practice in the literature on the returns to post secondary education, studies of the union wage premium, and studies of race and ethnic differences in labor market outcomes.

Unfortunately, however, the test instruments used to measure cognitive ability typically vary across data sets. This generates problems in comparing empirical findings across data sets, regardless of whether the coefficient on the cognitive test score is the primary object of interest or whether the scores are simply used as controls or proxies for unobserved ability. This problem also routinely prevents researchers from using data on cognitive test scores to examine how cognitive ability distributions vary across populations or time.

An example of this problem is the Armed Forces Qualification Test (AFQT) score reported in both the NLSY79 and NLSY97. The ASVAB test battery on which the AFQT is based was administered at different ages to the respondents of both surveys.[1] Furthermore, the test format has changed from a paper and pencil test (P&P) in 1979 to a computer administered test format (CAT) in 1997. For both reasons, the tests are not directly comparable. Consequently, one cannot make direct use of the tests to identify changes in the distribution of cognitive ability across the cohorts or to determine whether the relationship between cognitive ability and labor market outcomes differs between cohorts.

We confronted the problem of comparing AFQT scores across the NLSY cohorts in Altonji, Bharadwaj, and Lange (2009). In that study, we compared the distribution of labor market relevant skills, including cognitive skills, across the cohorts surveyed in NLSY79 and NLSY97. In this note, we explain the procedures used to make the AFQT-scores in both surveys comparable. The code and the data required to replicate this procedure are available at http://www.econ.yale.edu/~fl88[2]

We relied on two percentile mappings. The first transforms the CAT scores of the NLSY97 cohort into P&P scores and was performed for us by Daniel Segall. It is based on a study of a sample of test takers who were randomly assigned to take either the CAT or the P&P format of the ASVAB.

[1] The Armed Services Vocational Aptitude Battery (ASVAB) consists of 10 components. The results from a subset of these components are used to generate the AFQT-score. The AFQT80 reported by the NLSY79 is equal to the sum of the scores on the arithmetic reasoning, word knowledge, paragraph comprehension components plus one half times the score from the numerical comprehension component of the ASVAB.

[2] The program and data file reflects some modifications to our procedures that we made after completing the May 2009 draft of Altonji, Bharadwaj, and Lange (2009). The modified data will be used in the next draft, although this will make little difference in our results.

The second mapping addresses the problem that we observe individuals who were tested at different ages.

Mapping the CAT into the Paper and Pencil Test Scores.

The military screens recruits using the AFQT. The switch from the P&P format of the AFQT to the CAT format has raised concerns that the switch in the testing format might affect the consistency over time of selection criteria into the military.[3] To address this concern, Segall (1997) performed an equating study based on a sample of test takers who were randomly assigned to either the P&P or the CAT format of the test. Based on the study, he developed a mapping of the component scores for the CAT into the corresponding component scores of the P&P ASVAB.

The details of the study leading to this equating procedure are in Segall (1997). As mentioned above, the equating is based on randomly assigning individuals to either take the P&P or the CAT. Data collection for the equating procedure took place in 6 different geographic regions across the US in 1988 (N=8,040) and between 1990 and 1992 (N=10,379). The sample is drawn from military applicants. Females composed 20% of the sample and blacks composed 27%. Based on the assignment to either the P&P or the CAT, Segall estimated empirical cumulative distribution functions (CDFs) of the test score components. These form the basis for the equating of scores across percentiles of the CDF in the P&P and in the CAT. Instead of using the empirical CDFs directly, Segall smoothed both the CDFs and the equating transformations to reduce the sampling variation induced by estimating the empirical CDFs.

Segall (1997) examines the validity of this procedure. He found no evidence that the distribution of composite scores constructed after equating the component scores varies by test-taking format, suggesting that the reliability of the component tests and their cross-correlations are similar across formats. Furthermore, differences across demographic groups in the distribution of test scores are similar across CAT and P&P once the equating procedure was applied.

In Altonji, Bharadwaj, and Lange (2009), we were not able to obtain the original data from Segall's random assignment study, nor the actual mapping used by Segall. Instead, we provided Segall with the component scores of CAT-ASVAB from NLSY97, and he very kindly provided us with a set of equated scores for the P&P-ASVAB components. We then added the scores for the Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension and half the Numerical Operations component scores to construct AFQT80 for the NLSY97 sample that are comparable to the AFQT80 scores obtained for the NSLY79 sample.

Equating AFQT scores across the Test-taking Ages.

The above equating procedure addresses differences in the test formats, but not in the age at which individuals took the tests. For this purpose we rely on another set of equipercentile mappings. The individuals sampled in the NLSY79 and NLSY97 differed in the age at which they took the AFQT. The test was administered in the second half of 1979 for NLSY79 and in the second half of 1997 and the first months of 1998 for NLSY97. Table 1 shows the frequency distribution of the age when individuals took the test for both samples.

[3] The ASVAB is the main screening test for applicants to the U.S. military.

Frequency of Test-taking Age in Both Surveys

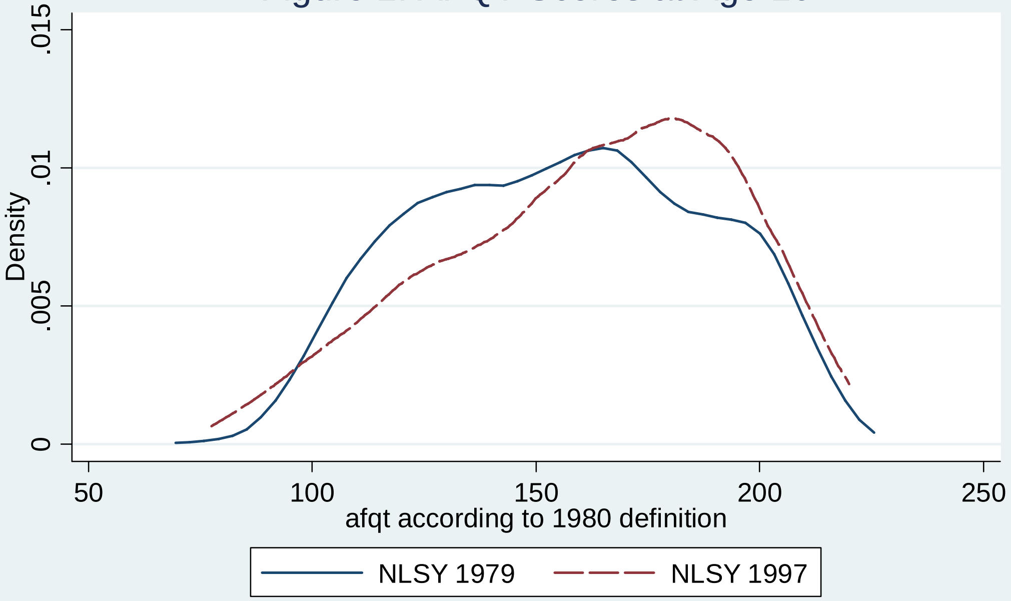
Age at Test	*NLSY 1979*	*NLSY 1997*
12	0	944
13	0	1,387
14	0	1,460
15	962	1,478
16	1,511	1,303
17	1,488	427
18	1,432	3
19	1,502	0
20	1,558	0
21	1,539	0
22	1,529	0
23	357	0
Total	11,878	7,002

Reported are (unweighted) frequency counts of test-taking age in NLSY79 and NLSY97. The age at test is obtained using the survey responses for the years the test was administered. For 202 respondents in NLSY79, the age at test is constructed using (age = 1980 minus birth-year).

Clearly, comparisons of the AFQT-distribution across cohorts will be biased if we do not adjust for the fact that the distribution of test-taking age differs significantly across samples. We base our adjustment on the observed overlap in the distribution of test-taking age. In particular, we exploit the fact that both surveys have a large group that took the test at age 16.

Figure 1 shows the NLSY79 and NLSY97 distributions of equated test-scores for respondents who were 16 years old when they took the test. (The scores reported for the NLSY97 are those obtained from mapping the CAT-ASVAB into the P&P.) The distributions in Figure 1 are directly comparable, because they are based on populations who took the test at age 16. These two distributions are quite similar, but there is some suggestion that the NLSY97 cohort is a bit stronger in cognitive ability than the NLSY79 cohort. The NLSY79 distribution has a mean of 155.93 and a standard deviation of 31.48. The NLSY97 distribution by contrast has both a higher mean (161.25) and a higher standard deviation (32.45).

Figure 1: AFQT Scores at Age 16



The NLSY79-scores are the P&P scores reported by the NLSY79.
The NLSY-97 scores are based on the CAT scores from NLSY97 and the equation by Segal (1997).
Both populations are weighted to be population representative.

In order to assign equivalent test-scores to the remainder of the surveys, we applied an equipercentile mapping across age-groups within sample. That is, for each individual, we determined the percentile in the AFQT distribution within sample and age.[4] We then assigned to these individuals the corresponding AFQT-score from the same percentile in the age 16 distribution from the same survey.[5] For example, if an individual was observed at the qth percentile in the distribution of age 13 in the NLSY97, then we assigned to this individual the score of the qth percentile of the age 16 distribution from the NLSY97.

We therefore mapped the distributions of scores from all ages into the distributions of scores presented in Figure 1, thus achieving comparable scores across both surveys.

References:

Altonji, J., Bharadwaj, P. & Lange, F. "Changes in the Characteristics of American Youth - Implications for Adult Outcomes" revised May 2009.

[4] Our procedure requires sufficient observations within each age and sample. We therefore aggregate ages 22 and 23 in NLSY79 and ages 17 and 18 in NLSY97.

[5] To achieve population representative samples, we use the custom weights provided by the NLSY for the years (1979 and 1997) during which the ASVAB was adminstered.

Segall, D. O. (1997). "Equating the CAT-ASVAB". In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized adaptive testing: From inquiry to operation (pp. 181-198). Washington, DC: American Psychological Association.