Personality and Individual Differences 85 (2015) 66–68

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Short Communication

Interpreting the NLSY79 empirical data on "IQ" and "achievement": A comment on Borghans et al.'s "Identification problems in personality psychology"

David Salkever *

Department of Public Policy, University of Maryland, Baltimore County (UMBC), Public Policy Bldg. Rm. 418, 1000 Hilltop Circle, Baltimore, MD 21250, United States

ARTICLE INFO

Article history:
Received 13 March 2015
Accepted 21 April 2015

Keywords:
Achievement
Intelligence
Ability
NLSY79
IQ tests
AFQT

ABSTRACT

In an otherwise interesting and enlightening article, Borghans, Golsteyn, Heckman, and Humphries (2011) analyzed evidence from the 1979 National Longitudinal Survey of Youth (NLSY79) to support their contention that "achievement" tests have greater power than "IQ" tests in predicting "a variety of life outcomes". A key point in their argument is their contention that scores on the Armed Forces Qualifications Test (AFQT) represent "achievement" scores and that the AFQT is qualitatively different from purported true "IQ" score data also available in the NLSY79. This contention is based on both conceptual argument and empirical analysis of NLSY79 data. This comment disputes their contention on both grounds. First, it argues that their conceptual distinction is contradicted in the educational testing literature and is based on erroneous assumptions about the nature of the purported true "IQ" test data in the NLSY79. Second, it presents evidence that their empirical findings flow from problems in true "IQ" score imputation and large gaps in calendar time between the purported true "IQ" tests and AFQT and personality test data in the NLSY79 data set.

1. Introduction

In an otherwise interesting and enlightening article, Borghans, Golsteyn, Heckman, and Humphries (2011) provide an arguably misleading analysis of evidence from the NLSY79 to support their contention that "achievement" tests have greater power than "IQ" tests in predicting "a variety of life outcomes". Their analysis flows from (1) a presumption that the "IQ" test data from the NLSY79 that they use primarily reflects "fluid intelligence" rather than the "acquired knowledge as captured by achievement tests", (2) the hypothesis that the most important "achievement test" measure in the NLSY79 data (the AFQT) is more affected by "personality traits" (as reported in the NLSY79) than are the particular test scores in the NLSY79 that they view as true "IQ" tests, and (3) the empirical result that when we control for true "IQ" scores, an important share of the variance in the NLSY79 "achievement test" scores is still explained by data on "personality traits". [1]

In this note, we first suggest that the sharp conceptual distinction that they pose between "achievement" tests versus the *specific* "IQ" tests that they study is not supported by authoritative literature on educational testing. We then report revised empirical results that are much less supportive of their proposed "IQ" vs. achievement test distinction. The results presented suggest their original empirical finding primarily reflects imprecision introduced by use of imputed data, and substantial time lags between some reported "IQ" scores and the dates of the AFQT testing.

2. "IQ" tests in the NLSY79 data

The "IQ" test data used by Borghans et al. are a composite based on scores from 7 different tests reported in NLSY79 school transcript data. The 7 different "IQ" tests, and the numbers of respondents reporting each test, were as follows: Otis–Lennon Mental Ability Test (1191), California Test of Mental Maturity (599), Lorge–Thorndike Intelligence Test (691), Henmon-Nelson Test of Mental Maturity (201), Kuhlmann–Anderson Intelligence Test (176), Wechsler Intelligence Scale for Children (120), and the Stanford–Binet Intelligence Scale (101). Since school transcript data for the 12,000+ NLSY79 respondents did not routinely include scores for any specific one of these tests, Borghans et al. created a composite "IQ" variable equal to the percentile score for whichever

* Tel.: +1 410 455 8459; fax: +1 410 455 8066.
E-mail address: Salkever@umbc.edu

[1] On the distinction between "IQ" and achievement tests, Borghans et al. also invoke Almlund, Duckworth, Heckman, and Kautz (2011), who suggest it is "useful to reserve the term 'intelligence tests' for tests that primarily measure fluid intelligence and the term 'achievement tests' for tests that primarily measure crystallized intelligence". In what follows, I suggest that the "IQ" test data examined by Borghans et al. do not meet this "intelligence test" criterion.

http://dx.doi.org/10.1016/j.paid.2015.04.042
0191-8869/

of these 7 tests were found in each respondent's data.[2] They also explain, in the web appendix to their paper, that they used reported percentile scores rather than "IQ" scores to construct their composite variable because the percentile scores were presumed to be more consistent across the 7 different tests in the data.[3]

While Borghans et al. drew a sharp conceptual distinction between "IQ" tests and achievement tests (as noted above), eminent educational testing experts have argued that conceptual distinctions between tests of intelligence, aptitude, and achievement are often unclear. Since at least the 1950's, it has been widely acknowledged that the achievement test criterion proffered by Borghans et al. (i.e., that such tests measure "acquired knowledge") applies to all three types of tests (e.g., Wesman AG. Aptitude & achievement. The Psychological Corporation, 1956). Lennon, author of the predominant "IQ" test used by Borghans et al. in their NLSY79 analysis, observed (Lennon, 1980) that his own (Otis–Lennon) test and several similar "true" IQ tests are best interpreted as testing "scholastic aptitude" rather than "intelligence" per se. He notes that all such tests include question domains (e.g., vocabulary, arithmetic reasoning) presumed to be "effective in predicting achievement in scholastic areas". He views these tests as similar to "achievement tests" since they sample "learned behaviors, developed abilities" and "in this sense *are* achievement measures." (Emphasis in the original.)[4]

Similarly, commenting on the aptitude vs achievement distinction, Anastasi (1984) observes that

> "tests of developed abilities do not fall into sharply differentiated categories (but) along a continuum. Both aptitude and achievement tests vary ... among themselves; ... those near the center of the continuum overlap to such a degree as to be nearly indistinguishable ... (T)ests of developed ability differ in the degree of precision versus vagueness with which the relevant domain of antecedent experience is defined ... (T)he experiential pool upon which the test constructor draws ... is defined with considerable clarity ... in constructing ... an achievement test in solid geometry, or medieval history ... At the other extreme ... (for) a test like the Stanford–Binet ... the definition specifies little beyond growing up in America in the twentieth century. (For) broadly oriented educational achievement batteries ... the domain of antecedent experience could be defined as growing up and going to school in America in the twentieth century."

A clearer and more recent restatement of Anastasi's perspective on aptitude and achievement tests is in her authoritative text (Anastasi & Urbina, 1997):

> "... the difference between achievement and aptitude tests is a difference in degree of uniformity of relevant antecedent experience. Thus, achievement tests measure the effects of relatively standardized sets of experience, such as a course in elementary French, trigonometry, or computer programming. In contrast, *aptitude* test performance reflects the cumulative influence of a multiplicity of experiences in daily living ... aptitude tests measure the effect of learning under relatively uncontrolled and unknown conditions, whereas achievements measure the effects of learning that occurred under partially known and controlled conditions ... Any cognitive test, regardless of what it has been called traditionally, provides a sample of ... what the individual knows ... and measures the level of development attained in one or more abilities" (p. 475).

In sum, agreement among a number of educational testing experts has emerged that many tests commonly referred to as testing "IQ" or "intelligence", including those from the NLSY79 data examined by Borghans et al., are in fact fundamentally similar to "aptitude" and "achievement" tests in that all are tests of *developed abilities* that are influenced by individuals' learning and training. Admittedly, some distinctions among tests have recently been noted (Nisbett et al., 2012), including the view that a few specific "IQ" tests (such as the Raven's Progressive Matrices test) are relatively more indicative of the "fluid intelligence" component of "general intelligence" (or "g"); however, these recent assessments have not argued that the specific "IQ" tests on which Borghans et al. relied are in fact primarily tests of "fluid intelligence". (It has also been noted (Nisbett et al., 2012) that even "fluid intelligence" is also influenced by learning experiences.)

3. Re-analysis of "IQ" vs. "achievement" scores in the NLSY79 data

The various "IQ" tests and the AFQT "achievement" tests in the NLSY79 examined by Borghans et al. may fall at various points along Anastasia's "continuum", so the validity of their conclusions about "IQ" vs. AFQT test should correspond to differences among these specific tests. While in principle these differences might be documented directly by delineating the differences in "relevant antecedent experience" among these tests, Borghans et al. have chosen the more practical course of basing their contentions on correlational analyses. As noted above, their main contentions are: (1) that relatively low correlations between respondents' "IQ" scores and their AFQT "achievement" score support the view that the "IQ" and AFQT tests are different in kind and (2) that the contribution of personality factors to explained variance in AFQT scores, beyond that which is explained by the "IQ" scores, confirms this distinction between the "IQ" and AFQT tests.

Several concerns may, however, be raised about these analyses. First, while the AFQT scores and the scores on the two NLSY79 personality factors (self-esteem and locus of control) were all obtained at roughly the same time (1980) in the NLSY79 data-gathering process, "IQ" scores from the school transcript data typically pertained to tests taken at least 5 and up to 15 years earlier. Thus, relatively weak correlations between "IQ" and AFQT scores could simply reflect increases (or even decreases) in respondents' "developed abilities" as they gain more education or acquire additional life experiences.

Second, there is at least a possibility that weak correlations could arise because of measurement errors from the imputation process used by Borghans et al. for those respondents with "IQ" test scores in the school transcript data that were not reported as percentile scores. In their web appendix, Borghans et al. explain that percentile scores were imputed for these respondents to make the scores comparable across the 7 different "IQ" tests used to construct their composite "IQ" variable. While the fraction of cases in which imputation was performed was not reported by Borghans et al., our tabulations of the NLSY79 data suggest that this fraction is probably about two-thirds of all respondents in their analysis.

[2] The majority of respondents did not have any of these scores reported and were therefore not included in the Borghans et al. analyses. While a small number of respondents actually had scores from two or more of these 7 tests in their school transcript data, Borghans et al. do not explain for these respondents how they selected the one test score included in their analysis.

[3] in cases where the school transcript data only included an "IQ" score for one of these 7 tests rather than a percentile score, they converted the "IQ" test score to an imputed percentile score based on data from other respondents who had *both* an "IQ"

[4] While Lennon's analysis focused specifically on 4 tests, including two (Otis–Lennon and Henmon-Nelson) that account for substantial portions of the NLSY79 "IQ" scores used by Borghans et al., he also clearly views these 4 tests as commonly-used examples of the general category of group-administered intelligence tests. Other tests in this category (e.g., California Test of Mental Maturity, Kuhlmann–Anderson Test, Lorge–Thorndike Test) also comprise much of the NLSY79 "IQ" data used by Borghans et al.

Table 1
Comparison of Borghans et al. and Re-analysis Correlational Results.

	Statistic	A Borghans[a]	B Re-analysis[b]	C Re-analysis: IQ tests post-1974[c]	D Re-analysis[b] for age < 18 in round 1	E Re-analysis for age < 18 in round 1: IQ tests post-1974[c]
1	Correlation AFQT w. IQ	0.65	0.753	0.824	0.773	0.839
2	AFQT R-squared w. IQ only	0.43	0.567	0.678	0.598	0.704
3	AFQT R-squared w. IQ and personality factors[d]	0.48	0.592	0.705	0.616	0.716
4	Increase in AFQT R-squared w. only Addition of personality factors[d]	0.05	0.025 (n = 1412)	0.027 (n = 539)	0.018 (n = 678)	0.012 (n = 401)

[a] All figures in this column are reproduced from Borghans et al. (2011), Table 1 and Fig. 2(A).
[b] Restricted to NLSY79 respondents with a percentile scores for at least one of the 7 "IQ" tests (in Borghans et al.) and an AFQT score.
[c] Restricted to NLSY79 respondents with a post-1974 percentile score for at least one of the 7 "IQ" tests (in Borghans et al.) and an AFQT score.
[d] Personality factors in cols. B–E included the Rotter Locus of Control and the Rosenberg Self-Esteem Scale item response score.

We tested the sensitivity of the empirical results from Borghans et al. to these potential concerns via re-analysis of NLSY79 data. This re-analysis eliminated possible measurement errors due to imputation by only including respondents for whom at least one "IQ" percentile score was reported in the data. We used two alternative strategies to constrain the number of years in the time span between the "IQ" test and the AFQT. First, since NLSY79 respondents ranged in age at the initial survey date from 14 to 22 years, we obtained correlation results based only on respondents who were under age 18 in the first survey round. This reduced the mean time span between the tests from 9.25 years to about 6.9 years. Second, we restricted some analyses to respondents whose "IQ" tests were obtained after 1974, which reduced the time span from the "IQ" test to the AFQT to about 3.1 years on average. Analyses using both restrictions reduced the mean time span between the tests to about 2.8 years.

Table 1 compares the empirical results from Borghans et al. (in column A) with the results from our re-analysis (in columns B through E). Results in row 1 show that the AFQT-"IQ" correlations are considerably higher in our re-analysis compared with column A. The increase from col. A to col. B is presumably due to the exclusion of imputed "IQ" values. The fact that the correlations are highest for cols. C and E suggests that reducing the time span between the date of the "IQ" test and the date of the AFQT is also important. The same pattern of results is, of course, also observed in row 2 (since the figures in row 2 are simply the squares of the figures in row 1).

The remaining two rows of the table show that the increment in the fraction of explained variation in the AFQT when personality factors are added, over and above the variation explained by "IQ", is substantially reduced when imputations are excluded (from 0.05 to 0.025 in col. B and to 0.027 in col. C), and declines even further (to 0.012) when pre-1975 "IQ" data is excluded and when respondents over age 17 (at the initial survey date) are excluded.

4. Conclusion

Our re-analysis suggests the two empirical findings, used by Borghans et al. to differentiate the AFQT from the "IQ" tests in the NLSY79 data are substantially diminished when imputation is eliminated and when one reduces the time differential between the date of the AFQT test and the date of the "IQ" test. Our results suggest that as an empirical matter, the AFQT test and the group "IQ" tests in the NLSY79 data behave in a similar fashion. Given the murky conceptual distinctions between the AFQT and these "IQ" tests, these empirical results are not surprising.

It may be that one needs data on other types of "IQ" tests more heavily weighted toward fluid intelligence to demonstrate clear empirical distinctions between these other types of tests vs. the AFQT *or* vs. other "IQ" test scores. Such data are not available in the NLSY79 survey data files.

Appendix A. Data Appendix

Correlations in columns B through E of Table 1 use data for NLSY79 respondents excluding those in the military subsamples. All respondents had a percentile score reported for at least one of the 7 "IQ" tests used for the composite "IQ" variable used by Borghans et al., and had a valid AFQT-3 percentile code. The "IQ" percentile code for each respondent was the valid percentile code reported for whichever one of the 7 tests appeared in their NLSY79 data. For respondents with valid percentiles for more than one test, we used the code of the test that appeared least frequently in the data set. Personality variables were summed item response scores for the locus of control and for self-esteem. While 3 different versions of the self-esteem variable are available on the NLSY79 data, and Borghans et al. did not specify which of these variables they used; we therefore used the summed item response since this variable produced results that were closest to the Borghans et al. results. As an alternative, use of the percentile of the standardized item response score for self-esteem would have reduced the R-squared increments from adding personality factors even further (ranging from 0.021 to 0.010).

References

Almlund, M., Duckworth, A. L., Heckman, J., & Kautz, T. (2011). Personality psychology and economics. In E. A. Hanushek et al. (Eds.). *Handbook of the economics of education* (Vol. 4). Amsterdam: North-Holland.
Anastasi, A. (1984). Aptitude and achievement tests: The curious case of the indestructible straw person. In B. S. Plake (Ed.), *Social and technical issues in testing: Implications for test construction and usage*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, N.J.: Prentice-Hall.
Borghans, L., Golsteyn, B. H. H., Heckman, J., & Humphries, J. E. (2011). Identification problems in personality psychology. *Personality and Individual Differences, 51*, 315–320.
Lennon RT. The anatomy of a scholastic aptitude test. National Council on Measurement in Education. Measurement in Education 11(2) (Winter 1980).
Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., et al. (2012). Intelligence: New findings and theoretical developments. *American Psychologist, 67*(2), 130–159.
Wesman AG. Aptitude, intelligence, and achievement. The Psychological Corporation, Test Service Bulletin No. 51 (Dec. 1956).