6-1-2014

Measurement Error In The AFQT In The NLSY79

Lynne Steuerle Schofield
Swarthmore College, lschofi1@swarthmore.edu

Recommended Citation

Lynne Steuerle Schofield. (2014). "Measurement Error In The AFQT In The NLSY79". *Economics Letters*. Volume 123, Issue 3. 262-265.
http://works.swarthmore.edu/fac-math-stat/80

Published in final edited form as:
Econ Lett. 2014 June 1; 123(3): 262–. doi:10.1016/j.econlet.2014.02.026.

Measurement Error in the AFQT in the NLSY79*

Lynne Steuerle Schofield
Department of Mathematics and Statistics Swarthmore College 500 College Avenue Swarthmore, PA 19081 lschofi1@swarthmore.edu 610-328-7896

Abstract

Many promising efforts in the social sciences aim to measure future outcomes (such as wages or health outcomes) given some base level of human capital or ability. They typically fail to recognize the proxies for human capital are all measured with error, creating bias in regression analysis. Here I show how item level data offers the opportunity to improve a broad range of economic, social and psychometric studies; an opportunity now enhanced significantly by the new release of item response level data for the Armed Forces Qualifying Test in the 1979 National Longitudinal Survey of Youth.

Keywords

Measurement error; NLSY79; AFQT

1. Introduction

Throughout the social sciences, researchers use test scores as proxies of latent variables. Most researchers acknowledge these proxies have measurement error but hope it is small enough that it need not be modeled. This paper demonstrates that under the assumptions of the psychometric models used to design, construct and score such tests, the measurement error is heteroskedastic and correlated with the latent variable.

I use the 1979 National Longitudinal Survey of Youth (NLSY79), arguably one of the most analyzed longitudinal data sets in the social sciences (Pierret, 2005) to make my case, though the conclusion applies more widely. A key feature of the NLSY79 is its breadth, particularly the Armed Services Vocational Aptitude Battery (ASVAB; DoD, 1984) and the Armed Forces Qualifying Test (AFQT; DoD, 1984) scores for 94% of the sample respondents.

*The work was supported in part by Award Number R21HD069778 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The content is solely the responsibility of the author and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health and Human Development or the National Institutes of Health.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Many studies use the AFQT to proxy for human capital and/or cognitive proficiency (e.g., Neal and Johnson, 1996; Lang and Manove, 2011). For example, Mears and Cochran (2013) examine the relationship between AFQT scores and criminal behavior. Lang and Manove (2011) use the AFQT as a control to examine if blacks and whites obtain the same amount of education conditional on cognitive ability. Most of these studies assume the AFQT contains no measurement error.

Griliches and Mason (1972) note the AFQT is an estimate of the true cognitive ability score, θ for each individual containing measurement error. If the measurement error is not modeled, there will be bias (Fuller, 1987) in the estimates of the effect of the both the AFQT and any covariate(s) associated with the AFQT on the response variable.

New AFQT data in the NLSY79 has recently been released to the public which will include item response level data for each of the respondents who took the AFQT; this data has been previously unavailable. Heretofore, the data contained an estimate of a respondent's proficiency on each AFQT subtest. The item response level data provides several (i.e., the number of items on the subtest) measures of an individual's subtest ability. While each item provides only a crude measure, taken together, the item responses can estimate both the AFQT subtest score and its measurement error. Because of the unique character of the measurement error in cognitive test scores, I demonstrate that instrumental variables (IV) methods will not solve the bias problem. I suggest a solution lies in estimating simultaneously the AFQT score and the regression coefficients in a structural equations model as in Junker, Schofield, and Taylor (2012).

2. The AFQT Score

In 1980, the NLSY79 respondents were administered the ASVAB for the purpose of constructing new national norms for the aptitudes of the nation's youth (Bock and Mislevy, 1981a). "The Profiles of American Youth," produced scores on each of the ten subtests of the ASVAB for all respondents who took the tests. For each subtest, the NLSY79 currently reports raw scores (or total correct), item response theory (IRT) scale scores, standard errors, and a sampling weight. The NLSY79 also contains unofficial derived AFQT scores (CHRR, 2001).

In order to ensure the reliability and validity of the test scores, the scale scores and standard errors were estimated using an item response theory (IRT) model (Bock and Mislevy, 1981). Below I review IRT models—the models used to design, construct, and score the AFQT in order to demonstrate that the assumptions underlying the IRT models are inconsistent with methods that are commonly used by researchers (e.g., OLS and IV) to estimate the effect of the AFQT on future outcomes.

3. Item Response Theory

Item response theory (IRT; van der Linden and Hambleton, 1997) models posit that the latent trait θ underlying performance on a test can be described by the item response function (IRF), a monotonically increasing function (see Figure 1). Most often, θ is assumed to be a continuous, unbounded hypothetical construct.

A standard unidimensional (the test only measures one latent trait) model is the 3-PL model (Birnbaum, 1968) which postulates that the probability that individual *i* responds correctly to item *j* is conditional on the latent trait θ, and three item parameters, a_j, b_j and c_j,

$$P_j\left(\theta_i\right) \equiv P\left[X_{ij}=1\right]=c_j+\frac{1-c_j}{1+exp\left[-\alpha_j\left(\theta_i-b_j\right)\right]}, \quad (1)$$

where x_{ij} equals 1 when individual *i* answers item *j* correctly and is 0 otherwise.

The “discrimination” parameter a_j affects the slope of the IRF and measures how influential changes in θ are on changes in $P(X = 1)$ (van der Linden and Hambleton, 1997). The “difficulty” parameter b_j affects the location of the IRF along the ability scale. As b_j increases, the probability that most examinees will answer the item correctly decreases. The “guessing” parameter c_j affects the location of the *y*–intercept of the IRF and measures the probability that an item can be answered correctly through guessing.

3.1. Estimation of in IRT models

A common assumption in IRT models is local independence which assumes that conditional on θ the item responses are statistically independent of one another (Lord and Novick, 1968). Under local independence,the joint likelihood of a vector of item responses is

$$L\left(x_{1,1},\ldots,x_{N,J}|\theta_1,\ldots,\theta_N\right)=\prod_i^N\prod_j^J P(x_{ij}|\theta_i)^{x_{ij}}\left(1-P\left(x_{ij}|\theta_i\right)\right)^{1-x_{ij}}. \quad (2)$$

IRT models are commonly estimated using marginal maximum likelihood methods (MML, Bock and Aitkin, 1981). In standard MML practice, θ is first treated as missing for all examinees and assigned an underlying probability distribution, e.g., $N(0, 1)$. Item parameter estimates are obtained by integrating θ out of the joint likelihood function and then taken to be fixed. Estimation of θ proceeds, commonly using maximum likelihood methods.

Standard errors of $\hat{\theta}$ are determined using the Fisher Information, $I_j(\theta_i)$,

$$SE\left(\theta_i\right)=\frac{1}{\sqrt{\sum_{j=1}^{J} I_j\left(\theta_i\right)}}. \quad (3)$$

In the 3PL model the Fisher information is

$$I_j\left(\theta_i\right)=\alpha_j^2\frac{\left(P_j\left(\theta_i\right)-c_j\right)^2}{\left(1-c_j\right)^2}\frac{\left(1-P_j\left(\theta_i\right)\right)}{P_j\left(\theta_i\right)} \quad (4)$$

Estimates of θ in IRT models increase in precision by increasing *J*, the number of test items. Thus, $SE\left(\hat{\theta}\right)$ tends toward 0 as $J \rightarrow \infty$ and $\hat{\theta}$ will be on average more precise for tests with more items. Additionally, $SE\left(\hat{\theta}\right)$ varies for different θ. Measurement error is generally greatest for examinees with very low or very high θ, and least for examinees near the middle of the distribution.

4. Attenuation Bias

The AFQT is often used as a predictor in a regression-based model,

$$y_i=\beta_0+\beta_1 AFQT_i+\beta_2 Z_i+\varepsilon_i, \quad \varepsilon_i \sim N(0,1) \qquad (5)$$

where y_i is the dependent variable of interest, and Z_i are a set of covariates (e.g., race, gender, educational attainment). Standard regression models like (5) assume that the predictor variables have been measured precisely and account only for error in y_i.

The measurement error in the AFQT scores, however, will result in bias whenever it is not explicitly modeled. This bias will underestimate the effect of the AFQT on *y*. The size of the bias in $\hat{\beta}_1$ is determined by the amount of measurement error in the AFQT. Bias will also occur in the estimates of the effect of any covariates *Z* correlated with θ including race, gender, and educational attainment.The size and direction of the bias in $\hat{\beta}_2$ depends on the size and direction of the correlation between the AFQT, *Z* and *y*.

It is often incorrectly assumed that the measurement error is small, normally distributed and homoskedastic resulting in small biases.

4.1. Measurement Error in the NLSY79 AFQT

In the NLSY79, Bock and Mislevy (1981) used BILOG software (Mislevy and Bock, 1983) to estimate the item and person parameters using MML methods (Bock and Mislevy, 1981, Appendix A) Their report provides estimates of the item parameters for the *J* = 30, 35, 15, and 25 items on the arithmetic reasoning, word knowledge, paragraph comprehension, and mathematical knowledge subtests respectively (Bock and Mislevy, 1981, Appendix D).

Using Bock and Mislevy's (1981) item parameters, I calculated $SE(\theta)$ for the range of values of θ in the NLSY79 using (3). In Figure 2, I show the heteroskedastic relationship between θ and $SE(\theta)$ for each of the four subtests. In all subtests of the AFQT, respondents with very high and very low levels of the latent trait have larger standard errors. The AFQT is a less precise proxy of human capital for those individuals with very high or very low levels of ability.

In addition, I simulated one million θs. Using Bock and Mislevy's item parameters, I simulated a set of item responses for each θ on each subtest and estimated the MLE using (2). I calculated the measurement error *u* for each $\hat{\theta}$. In addition to the heteroskedasticity, the measurement error is correlated with the estimates of θ. In Figure 3, I show the relationship of the mean measurement error $E(u_i)$ for each MLE, for each of the four subtests. For individuals with low estimates of θ, the error is on average positive and for individuals with high estimates of θ the error is on average negative.

Figure 2 and Figure 3 show the fallacy in the assumption that the measurement error is small, homoskedastic, or uncorrelated with the MLE. While the arguments above are based on population values of θ, $\hat{\theta}$ and the measurement error, we expect to see similar relationships in the sample of individuals in the NLSY79.

4.2. Item Level Data for the AFQT

The vast majority of researchers who use the AFQT as a predictor ignore the error entirely. A small set of examples exist where the error is modeled. Bollinger (1996) follows Klepper and Leamer (1984) to estimate (rather large) bounds of the regression coefficients when θ is a predictor. Cunha, Heckman, and Schennach (2010) extend a classical measurement error model using a MIMIC approach (Joreskog and Goldberger 1975) in applications in which multiple latent variables are used as predictors.

One might consider instrumenting one subtest of the AFQT with another subtest as a solution to the measurement error. But the measurement error is related to estimates of $\hat{\theta}$, thus violating the assumption in IV that the instrument is uncorrelated with the error. Furthermore, as Williams (2012) notes, there is a fundamental problem of identification when attempting to identify a continuous latent variable from a finite collection of binary proxy variables (such as item responses). Black, Berger, and Scott (2000) provide a solution to this problem in the limited case where the variable of interest is binary. Hu and Schennach (2008) discuss nonparametric identification of models in which the continuous variable of interest is measured with error, but their identification requires that at least one of the instruments be continuous.

Junker, Schofield, and Taylor (2012) suggest a structural equations modeling solution which simultaneously estimates θ and the regression coefficients. Their method requires data at the item response level. Fortunately, the NLSY79 has released data of the AFQT item responses which will allow researchers to model the measurement error of the AFQT in a way that has never been possible before.

5. Discussion

This paper points to a common, but often ignored source of bias in economic, social and psychometric studies that use latent variables measured by a test score as a predictor of future outcomes. It points to the unique opportunity now available with the release of the item response level data for the AFQT in the NLSY79 data. When test score proxies are used as predictors in regression based analyses, bias will occur in estimates of the effect of the latent variable measured with error and in any covariate associated with the latent variable. This paper has focused on the AFQT test in the NLSY79 data, because it has been used to proxy human capital and cognitive skills in thousands of studies in the social sciences. However, measurement error problems exist for any latent variable. Researchers should be aware of the measurement error and employ appropriate methods for handling the error in their analyses to avoid the bias that results. When possible, survey institutions should release item response level data so that researchers can estimate the measurement error directly.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinees ability. In: Lord, FM.; Novick, MR., editors. Statistical theories of mental test scores. Addison-Wesley; Reading MA: 1968.

Black D, Berger M, Scott F. Bounding parameter estimates with non-classical measurement error. Journal of the American Statistical Association. 2000; 95:739–748.

Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika. 1981; 4(6):443–459.

Bock, RD.; Mislevy, RJ. The profile of American youth: Data quality analyses of the Armed Services Vocational Aptitude Battery. National Opinion Research Center; Chicago: 1981.

Bollinger C. Measurement error in human capital and the black-white wage gap. Review of Economics and Statistics. 2003; 85:578–85.

Center for Human Resource Research. [Retrieved November 20, 2013] NLSY79 users guide. 2001. from http://www.bls.gov/nls/79guide/2001/nls79g0.pdf

Cunha F, Heckman JJ, Schennach SM. Estimating the technology of cognitive and noncognitive skill formation. Econometrica. 2010; 78(3):883–931. [PubMed: 20563300]

Department of Defense. Armed Services Vocational Aptitude Battery (ASVAB) test manual. Military Entrance Processing Command; Chicago, IL: 1984. (DoD 1304.12AA)

Fuller, WA. Measurement error models. Wiley; New York, NY: 1987.

Griliches Z, Mason WM. Education, income, and ability. Journal of Political Economy. 1972; 80(3):S74–S103.

Hu Y, Schennach S. Instrumental variable treatment of nonclassical measurement error models. Econometrica. 2008; 76(1):195216.

Joreskog KG, Goldberger AS. Estimation of a model with multiple indicators and multiple causes of a single latent variable. Journal of the American Statistical Association. 1975; 70:631–639.

Junker BW, Schofield LS, Taylor L. The use of cognitive ability measures as explanatory variables in regression analysis. IZA Journal of Labor Economics. 2012; 1:4.

Klepper S, Leamer E. Consistent sets of estimates for regressions with errors in all variables. Econometrica. 1984; 52:163–183.

Lang K, Manove M. Education and labor market discrimination. American Economic Review. 2011; 101:1467–1496.

Lord, FM.; Novick, MR. Statistical theories of mental test scores. Addison-Welsley; Reading, MA: 1968.

Mears DP, Cochran JC. What is the effect of IQ on offending? Criminal Justice and Behavior. 2013; 40(11):1280–1300.

Mislevy, RJ.; Bock, RD. BILOG: Item analysis and test scoring with binary logistic models [Computer program]. Scientific Software, Inc.; Mooresville, IN: 1983.

Neal D, Johnson W. The role of pre-market factors in black-white wage differences. Journal of Political Economy. 1996; 104:869–895.

Pierret C. The National Longitudinal Survey of Youth: 1979 cohort at 25. Monthly Labor Review. 2005; 128(2):3–7.

Williams, B. A measurement model with discrete measurements and continuous latent variables. 2013. Under Review

van der Linden, WJ.; Hambleton, RK., editors. Handbook of Modern Item Response Theory. Springer-Verlag; New York, NY: 1997.

- Test scores, which serve as a proxy for human capital are measured with error.
- Failure to model the error in the proxies creates bias in regression estimates.
- I show the error is often large, non-normal, and heteroskedastic.
- The NLSY79 has released new item level data for each subtest of the AFQT.
- The data offers opportunities for better estimates of the proxies and the error.

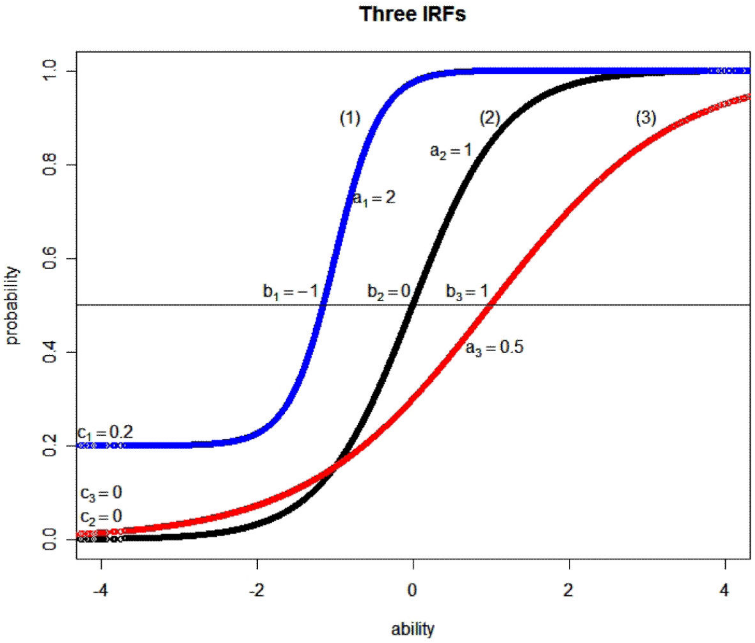

Figure 1.

Three typical IRFs of a 3-PL model

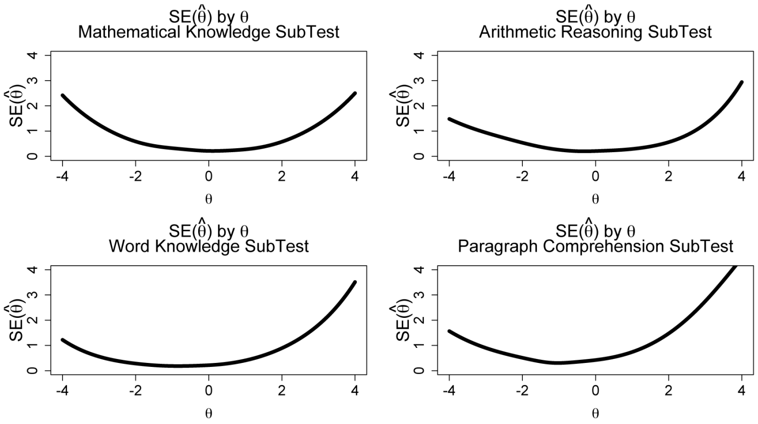

Figure 2.

$SE\left(\hat{\theta}\right)$ by θ for the four subtests of the AFQT

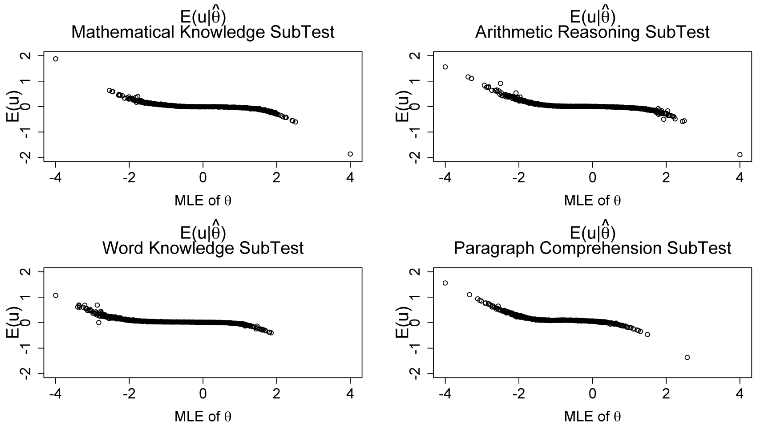

Figure 3.

Mean of the measurement error of $\hat{\theta}$ by $\hat{\theta}$ for the four subtests of the AFQT