Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Econometrics 121 (2004) 39–98
JOURNAL OF Econometrics
www.elsevier.com/locate/econbase

The effect of schooling and ability on achievement test scores☆

Karsten T. Hansen[a], James J. Heckman[b,c,*], Kathleen J. Mullen[b]

[a] *Kellogg School of Management, Northwestern University, Evanston, IL 60657, USA*
[b] *Department of Economics, The University of Chicago, 1126 East 59th Street, Chicago, IL 60637, USA*
[c] *The American Bar Foundation, 750 North Lake Shore Drive, Chicago, IL 60611, USA*

Abstract

This paper develops two methods for estimating the effect of schooling on achievement test scores that control for the endogeneity of schooling by postulating that both schooling and test scores are generated by a common unobserved latent ability. These methods are applied to data on schooling and test scores. Estimates from the two methods are in close agreement. We find that the effects of schooling on test scores are roughly linear across schooling levels. The effects of schooling on measured test scores are slightly larger for lower latent ability levels. We find that schooling increases the AFQT score on average between 2 and 4 percentage points, roughly twice as large as the effect claimed by Herrnstein and Murray (1994) but in agreement with estimates produced by Neal and Johnson (1996) and Winship and Korenman (1997). We extend the previous literature by estimating the impact of schooling on measured test scores at various quantiles of the latent ability distribution.

JEL classification: C35; C15; I21

Keywords: Education; Ability; Latent variables; Selection; MCMC

1. Introduction

There are two widely held and mutually inconsistent conceptions of ability and scholastic achievement tests. The first view claims that cognitive ability is essentially

☆ Supplementary data associated with this article can be found at http://athens.src.uchicago.edu/jenni/JOE/

* Corresponding author. Department of Economics, The University of Chicago, 1126 East 59th Street, Chicago, IL 60637, USA. Tel.: +1-773-702-0634; fax: +1-773-702-8490.
E-mail addresses: karsten-hansen@northwestern.edu (K.T. Hansen), jjh@uchicago.edu (J.J. Heckman), kjmullen@midway.uchicago.edu (K.J. Mullen).

0304-4076/$ - see front matter
doi:10.1016/j.jeconom.2003.10.011

fixed at a relatively early age (around age eight) and is virtually unchanged afterward. According to this view, achievement tests and IQ tests measure the same fundamental cognitive skill. The correlation between IQ and achievement tests is high and proponents of this view use these two types of tests interchangeably. According to scholars who advocate this point of view, schooling and other influences barely budge measured IQ. (See the evidence summarized in Herrnstein and Murray, 1994, Appendix B.) A consensus estimate in this literature is that a year of schooling raises measured IQ by about one point (Jencks, 1972; Herrnstein and Murray, 1994). [1]

A second widely held view claims that schooling raises achievement measured by tests and more successful types of schooling raise measured achievement more. This is the premise of large-scale testing programs designed to monitor the performance of schools. Debates about the effectiveness of vouchers and interventions hinge on their effects on measured achievement (see, e.g., Hanushek, 2002). This literature implicitly separates out latent ability (IQ) from measured ability and views schooling as a mechanism for either enhancing or revealing ability. Proponents of this view argue that schooling can increase measured ability by as many as 2–4 IQ points (Winship and Korenman, 1997; Neal and Johnson, 1996), or 2.9–5.7 AFQT points (2.7–5.4 percentage points).

This paper presents evidence that the measure of IQ used by Herrnstein and Murray is strongly affected by schooling. Postulating that latent ability cannot be affected by schooling, we test whether manifest ability is affected by schooling when both schooling and manifest ability are affected by latent ability. Manifest ability is widely regarded as a determinant of socioeconomic success. Gaps in test scores across socioeconomic groups are widely viewed as a major source of social problems (Jencks and Phillips, 1998; Herrnstein and Murray, 1994). We examine whether measured ability gaps can be eliminated by schooling.

Our measures of ability are the ASVAB achievement (competency) tests used to screen persons entering the military. ASVAB stands for Armed Services Vocational Aptitude Battery and is described in more detail below. We find that schooling, especially in the high school years, is an important determinant of measured achievement. It operates differently at different latent ability levels.

In order to establish these conclusions, we need to address the problem of reverse causality. There is a well-established empirical regularity that measured test scores predict schooling. Individuals choose to attend school in part based on their own intelligence which is measured by test scores. In addition, admission into colleges and fellowship support is based, in part, on scores on tests like the SAT. The central econometric question addressed in this paper is how to characterize and solve the problem of joint causality: schooling causing test scores and test scores causing schooling. Our

[1] IQ is assumed to have mean 100 and standard deviation 15. Many of the papers in the literature obtain estimates using the Armed Forces Qualification Test (AFQT), the test used in this paper, which has a scale of 0–105. Typically estimates are converted into "IQ points" by computing the effect of education in terms of standardized AFQT score and then, assuming that a standard deviation increase in AFQT is equivalent to a standard deviation increase in IQ score, by further multiplying by 15. Herrnstein and Murray (1994) estimate an increase of 1.1 IQ points per year of education, or 1.6 AFQT points (1.5 percentage points), using our estimate of the standard deviation of AFQT, 21.6.

solution is based on a model of test scores as manifestations of latent ability (and other determinants) with schooling determined by latent ability (and other determinants). Our framework accounts for ceiling effects (on some easy tests, students with very different ability levels get perfect scores) and endogeneity of schooling (which includes choice of date of entry into schooling as well as choice of final schooling level).

We find that the effects of schooling on test scores for a given level of ability are approximately linear across schooling levels. Effects are slightly larger for those with lower ability. One year of schooling increases the AFQT score between 2 and 4 percentage points on average. This is roughly twice as large as the effect claimed by Herrnstein and Murray (1994).

The plan of this paper is as follows. Section 2 presents our framework and discusses some important conceptual issues. Section 3 applies the method of control functions, developed by Heckman (1976, 1980) and Heckman and Robb (1985, 1986), to identify schooling effects on tests using a special feature of the National Longitudinal Survey of Youth (NLSY) data. A nonparametric version of the method is developed, but it suffers from certain practical limitations in more general cases. Section 4 presents a parametric econometric model motivated by choice theory for the joint determination of schooling and test scores. This method allows us to supplement the nonparametric control function method to impose additional identifying information to develop a method for determining the effects of schooling on test scores in more general data sets than the NLSY and to account for ceiling effects on tests. Section 5 presents empirical results. Section 6 concludes. Appendix A describes the data. The estimation algorithm for the control function approach is presented in Appendix B. The likelihood and the Bayesian computational methods used to estimate it are presented in Appendix C.

2. The relationship between ability and schooling

Let $T(s)$ be the test score of a person with s years of schooling at the time the test is taken. For notational simplicity, we keep implicit the conditioning on all other variables that determine $T(s)$ except latent ability f. The other variables might include age, socioeconomic status of the parents, and other environmental and genetic factors. We account for some of these additional variables in our empirical work.

Our model of test scores is based on an extension of the factor analysis model used in psychometrics (see e.g., Lord and Novick, 1968). Test score $T(s)$ is a manifestation of latent ability f mediated by schooling:

$$T(s) = \mu(s) + \lambda(s)f + \varepsilon(s), \tag{1}$$

where it is assumed that $\varepsilon(s)$ is independent of f. Both f and $\varepsilon(s)$ are assumed to have zero means. This amounts to a normalization and a definition of the mean, $\mu(s)$. We extend the standard model of factor analysis by allowing the level of s selected to depend on f. For externally-manipulated levels of schooling, $\mu(s)$ in Eq. (1) is the effect of schooling that is uniform across latent ability levels and $\lambda(s)$ is the effect of schooling on revealing or transforming latent ability f. The marginal causal effects of

changing schooling from s' to s on levels and slopes are $\mu(s)-\mu(s')$ and $\lambda(s)-\lambda(s')$ respectively using the usual *ceteris paribus* logic familiar to all economists.[2]

The psychometric and educational testing literatures are fundamentally ambiguous about what constitutes cognitive ability. Is it f, $T(s)$, $\mu(s)$ or $\mu(s)+\lambda(s)f$? Neal and Johnson (1996), Winship and Korenman (1997), Winship (2001), and Herrnstein and Murray (1994) take measured test scores ($T(s)$) to be cognitive ability. Yet the logic of IQ testing interprets f as cognitive ability. A reinterpretation of Eq. (1) writes $\lambda(s)f$ as ability determined at schooling level s. Knowing only $T(s)$ and $S=s$, we cannot decide which of these interpretations is correct. Without further information, the model is fundamentally underidentified because we do not observe f. We can identify the combination of parameters in

$$E(T(s)\,|\,S=s)=\mu(s)+\lambda(s)E(f\,|\,S=s)+E(\varepsilon(s)\,|\,S=s) \tag{2}$$

but the causal status of an estimated effect of S is unclear because both $E(f\,|\,S=s)$ and $E(\varepsilon(s)\,|\,S=s)$ may depend on S. Thus latent cognitive ability (f) may determine S and so may measured ability $T(s)$ (and hence $\varepsilon(s)$ given f). If the test studied does not directly affect schooling decisions, e.g., through its use in admission criteria, as is the case for the test analyzed in this paper, then $E(\varepsilon(s)\,|\,S=s)=0$.[3]

The empirical literature recognizes the problem of reverse causality, and adopts different strategies for identifying different parameters. Herrnstein and Murray (1994) and Winship and Korenman (1997) implicitly adopt $\mu(s)$ as their parameter of interest, assume that $\mu(s)=s\beta$ (linearity) and use an "early" test score (obtained at an earlier age) to proxy f.[4] Let proxy P be

$$P=\gamma_0+\gamma_1 f+\varepsilon(P), \tag{3}$$

where f and $\varepsilon(P)$ are independent, and γ_0, γ_1 are assumed not to be functions of the S in (1). Solving for f and substituting into (1) we obtain

$$T(s)=\mu(s)-\lambda(s)\frac{\gamma_0}{\gamma_1}+\frac{\lambda(s)}{\gamma_1}P+\left[\varepsilon(s)-\frac{\lambda(s)\varepsilon(P)}{\gamma_1}\right]. \tag{4}$$

Observe that the composite error is correlated with P unless P is a perfect proxy for f (so $\varepsilon(P)=0$), the implicit assumption used by Herrnstein and Murray (1994).[5] Herrnstein and Murray also implicitly assume that $\lambda(s)$ does not depend on s. Then, using ordinary least squares applied to (4), they estimate the marginal effect of schooling which in their setup is β ($\mu(s)=s\beta$). If $\lambda(s)=\lambda$, but $\varepsilon(P)\neq 0$, and if $\lambda\gamma_1>0$ (so that f affects P and $T(s)$ in the same way), then least squares based on (3) is upward biased. More generally, if λ depends on s, the bias is ambiguous and depends on specific parameter configurations. The combination of parameters $\mu(s)-\lambda(s)\gamma_0/\gamma_1$ becomes the implicit parameter estimated and it does not answer the questions posed in the literature.

[2] Throughout this paper we maintain the traditional separable-in-the-errors model of Eq. (1). A more general nonseparable model would be desirable but is beyond the scope of this paper.

[3] Even though the tested know their scores they are not directly used by schools or firms to screen persons and we assume that they do not affect subsequent actions.

[4] These authors also include additional control variables which we do not discuss.

[5] Olley and Pakes (1996) make the same assumption in a different context.

Winship and Korenman (1997) consider the problem of measurement error in their proxy P. They draw on work by Ashenfelter and Krueger (1994) who claim that the reliability (the proportion of variance of P that is true, $\gamma_1^2\sigma_f^2$, relative to the total variance $\gamma_1^2\sigma_f^2+\sigma_\varepsilon^2$) of IQ measures is typically above 0.9.

Winship and Korenman carry out a variety of sensitivity analyses, estimating the model under different assumptions about the reliability of the early IQ score, which they let take on values between 0.8 and 1. They obtain a wide range of estimates of the effect of schooling on AFQT, from 1.5 to 5 points. Correcting for measurement error under what Winship and Korenman believe to be "reasonable assumptions about the extent of measurement error," they estimate the effect of education to be 2.7 IQ points per year of school and they state that "a year of education most likely increases IQ by somewhere between 2 and 4 points."

Neal and Johnson take a different approach. They choose β in the specification $\mu(s)=s\beta$ as their parameter of interest and use month of birth (which determines years of schooling attained by a given birth cohort) as an instrument to avoid dependence of S on f.[6] This forced variation in schooling attained among children of the same nominal birth cohort is a source of identifying variation. We estimate a larger set of parameters and consider how schooling affects test scores at different levels of the latent ability distribution. However, our estimates of the same parameter are in agreement with theirs.

A variety of other studies, surveyed by Ceci (1991), rely on various "natural experiments" of uncertain quality. Winship and Korenman (1997) survey and criticize this literature.

In this paper, we estimate $\mu(s)$ and $\lambda(s)$ for different levels of schooling without imposing the parametric restrictions used in the previous literature. We explicitly account for the endogeneity of completed schooling. In addition we estimate the distribution of latent ability (f) and compare it with measured ability. We can identify the effect of schooling on measured test scores at different latent ability (f) levels. This allows us to identify where in the overall distribution of ability schooling interventions are the most effective. We first develop estimators based on the principle of control functions.

3. Simple identification strategies based on control functions

Our first approach to this problem exploits an unusual feature of the NLSY data. The test we study is given to a nationally representative sample of people. Some people who take the test are in school while others have finished school. We observe completed schooling for all individuals. Let S_T denote schooling that a person has at the date of the test. We observe the test score $T(S_T)$ which can be expressed as

$$T(S_T)=\mu(S_T)+\lambda(S_T)f+\varepsilon(S_T). \tag{5}$$

[6] In most school districts, in a given year any 5-year-old child whose birthday falls after October 1 must wait to start school in the following year.

Letting S denote the final level of schooling that is actually attained, $S \geqslant S_T$. Let A be the age at which a person is tested. If we redefine age so that schooling starts at age 0 and if we assume that dropouts do not return to school, then we observe $S_T = A < S$ if the test date comes before a person has completed his schooling.[7] If he has completed schooling by the time of the test, then we observe $S_T = S$.

Using the control function approach introduced in Heckman (1976, 1980) and Heckman and Robb (1985, 1986), and assuming no maturation effects (no independent effect of age on performance on the test) and that everyone starts school at the same age, we may write observed tests conditional on final schooling and schooling at the test date as

$$E(T(S_T) \mid S_T = s_T, S = s) = \mu(s_T) + \lambda(s_T)E(f \mid S_T = s_T, S = s) + E(\varepsilon(S_T) \mid S_T = s_T, S = s). \tag{6}$$

To simplify the notation we keep other conditioning variables implicit.

Because sampling is random across ages, if individuals consider only their final schooling level when making schooling decisions, irrespective of their path to schooling and there is no dropping out and re-entry, conditional on $S = s$ the observed S_T is random with respect to f. Thus $E(f \mid S_T = s_T, S = s) = E(f \mid S = s)$. Further, if the test is not used to make decisions about schooling, $E(\varepsilon(S_T) \mid S_T = s_T, S = s) = 0$.

Under these assumptions we obtain

$$E(T(S_T) \mid S_T = s_T, S = s) = \mu(s_T) + \lambda(s_T)E(f \mid S = s). \tag{7}$$

From this equation it is clear that we cannot identify the scale of f without some normalization. Setting $\lambda(1) = 1$ is one such normalization. We can identify $\lambda(s_T)$ up to the normalization because for two different schooling levels s, $s' \geqslant s_T$, $s \neq s'$,

$$E(T(S_T) \mid S_T = s_T, S = s) - E(T(S_T) \mid S_T = s_T, S = s') = \lambda(s_T)[E(f \mid S = s) - E(f \mid S = s')].$$

Assuming $\lambda(s_T) \neq 0$, we may form the ratio

$$\frac{E(T(S_T) \mid S_T = s'_T, S = s) - E(T(S_T) \mid S_T = s'_T, S = s')}{E(T(S_T) \mid S_T = s_T, S = s) - E(T(S_T) \mid S_T = s_T, S = s')} = \frac{\lambda(s'_T)}{\lambda(s_T)}$$

for two values $s_T \neq s'_T$, both less than or equal to s, s'. Therefore with one normalization we can identify all of the $\lambda(s_T)$, $s_T = 1, \ldots, \bar{S} - 1$ (we cannot identify $\lambda(\bar{S})$ because there is only one possible value of s for $S_T = \bar{S}$).

Taking expectations with respect to S_T alone we obtain

$$E(T(S_T) \mid S_T = s_T) = \mu(s_T) + \lambda(s_T)E[f \mid S_T = s_T].^{8} \tag{8}$$

[7] Cameron and Heckman (2001) present evidence that few high school dropouts return to school.

[8] Note that $E(\varepsilon(S_T) \mid S_T = s_T) = E[E[\varepsilon(S_T) \mid S_T = s_T, S = s] \mid S_T = s_T] = 0$.

Recall that we know $\lambda(s_T)$, $s_T = 1,\ldots,\bar{S}-1$, from the preceding argument. Subtracting (8) from (7) we obtain

$$E(T(S_T)\,|\,S_T = s_T, S = s) - E(T(S_T)\,|\,S_T = s_T)$$
$$= \lambda(s_T)[E(f\,|\,S = s) - E(f\,|\,S_T = s_T)]$$

so we can identify for $s \geqslant s_T$, $s_T = 1,\ldots,\bar{S}-1$,

$$E(f\,|\,S = s) - E(f\,|\,S_T = s_T)$$
$$= \frac{E(T(S_T)\,|\,S_T = s_T, S = s) - E(T(S_T)\,|\,S_T = s_T)}{\lambda(s_T)}. \tag{9}$$

Let $E(f\,|\,S = s) = a_s$ and $E(f\,|\,S_T = s_T) = b_{s_T}$. From the data we can form a matrix of the following identifiable combination of parameters:

$$\begin{pmatrix} a_{\bar{S}} - b_1 & a_{\bar{S}-1} - b_1 & \ldots\ \ldots\ \ldots & a_1 - b_1 \\ a_{\bar{S}} - b_2 & a_{\bar{S}-1} - b_2 & \ldots\ \ldots\ \ldots & \sim \\ \ldots & \ldots & \ldots\ \ldots\ \sim & \sim \\ a_{\bar{S}} - b_{\bar{S}-1} & a_{\bar{S}-1} - b_{\bar{S}-1} & \ldots\ \sim\ \sim & \sim \\ \sim & \sim & \ldots\ \sim\ \sim & \sim \end{pmatrix}$$

where $\sim$ in a cell denotes the absence of data on the entry. We also know as a consequence of $E(f) = 0$ that, if we define $P_j = Pr(S = j)$,

$$\sum_{j=1}^{\bar{S}} a_j P_j = 0. \tag{10}$$

Letting $\tilde{P}_j$ be $Pr(S_T = j)$, we also obtain

$$\sum_{j=1}^{\bar{S}} b_j \tilde{P}_j = 0. \tag{11}$$

Taking a weighted sum across row 1 of the matrix, we identify b_1 since

$$\sum_{j=1}^{\bar{S}} P_j(a_j - b_1) = \sum_{j=1}^{\bar{S}} a_j P_j - b_1 \sum_{j=1}^{\bar{S}} P_j = -b_1$$

by (10) and the fact that $\sum_{j=1}^{\bar{S}} P_j = 1$. Going across the first row element by element, we obtain a_j, $j = 1,\ldots,\bar{S}$. Going down the first column, we obtain the remaining b_j, $j = 1,\ldots,\bar{S}-1$. Using (11) we identify $b_{\bar{S}}$. Thus the model is fully identified except for $\lambda(\bar{S})$ and $\mu(\bar{S})$.

Attractive as these results are, there are three reasons to be cautious about estimates derived from this identification strategy: (a) Age effects (maturation effects) may affect test scores independently of any effect of schooling because persons may acquire life experiences that raise their test scores independently of their schooling at the date of the test. Our procedure has to be modified to distinguish age effects from schooling effects. (b) Persons start school at different ages. Less able people (those with lower

f) may start school at later ages, making an assumption of an identical school starting age for all persons problematic. Simply conditioning on the starting age N to solve this problem is not satisfactory given its likely dependence on f. (c) In principle there might be a separate N effect on test scores apart from the dependence of N on f if there are discouragement effects (students older than their classmates may feel inferior and be less motivated). The confluence of an endogenous N and independent age effects is problematic.

Modeling the starting age N along with the schooling level S does not pose any conceptually new problem as long as there are no age-at-test effects. We can use different (S_T, N), (S, N) pairs and replace (S_T, S) in the preceding analysis. Data cells may thin out but the previous identification strategy works.

Allowing for age in addition to N produces a fundamental identification problem if we maintain the "no return to school for dropouts" assumption. Observe that by definition $S_T = \min\{A - N, S\}$, so that S, S_T and $A - N$ cannot be freely varied. A more general model that incorporates age and entry writes the test score T as $T(A, S_T, S, N)$ where A is the age at the test, S_T is the level of schooling at the date of the test, S is the final level of schooling attained and N is the age at which the person enters school. For simplicity normalize $N = 0$ to be the "normal age" of starting school. If A and S_T both affect the measured test score directly, while S and N do not directly affect the test score but potentially are stochastically dependent on latent ability f, we may write

$$T(A, S_T, S, N) = \mu(A, S_T) + \lambda(A, S_T) f + \varepsilon(A, S_T).^{9} \tag{12}$$

Then conditioning on observable (A, S_T, S, N) we obtain

$$\begin{aligned} &E(T(A, S_T, S, N) \mid A = a,\ S_T = s_T,\ S = s,\ N = n) \\ &\quad = \mu(a, s_T) + \lambda(a, s_T) E(f \mid S = s,\ N = n), \end{aligned} \tag{13}$$

where we assume $\varepsilon(A, S_T)$ is independent of all other variables. Observe that when $S_T < S$, fixing N and S_T determines A:

$$A = S_T + N. \tag{14}$$

This exact linear dependence does not apply to persons with completed schooling ($S_T = S$). In that subpopulation, S and S_T cannot be independently varied so the control function identification strategy previously developed breaks down but the exact linear dependence (14) does not hold so that we can independently vary A and $S_T = S$ for each N. If we parameterize $\mu(A, S_T)$ and $\lambda(A, S_T)$, we can identify separate effects of age and schooling at the test date.[10] With sufficient structure, we can extrapolate $\mu(A, S_T)$ and $\lambda(A, S_T)$ back to ages and schooling levels at schooling levels $S_T < S$. This method is pursued in Section 4.

[9] If N causally affects the test, then (12) is modified to read $T(A, S_T, S, N) = \mu(A, S_T, N) + \lambda(A, S_T, N) f + \varepsilon(A, S_T, N)$.

[10] Thus with $\mu(A, S_T) = \varphi_1(A) + \varphi_2(S_T)$ and $\lambda(A, S_T) = \eta_1(A) + \eta_2(S_T)$ we can break these linear dependencies by extrapolating back to the subpopulation where the linear dependencies hold. Multiplicative versions can work as well. This is the strategy used in Section 4 to achieve identification of these effects. See the closely related identification analysis of Heckman and Vytlacil (2001).

In our data, there are effectively two starting ages $N \in \mathcal{N} = \{0, 1\}$. Given our "no return to school for dropouts" assumption, in the sample $S > S_T$, people who start school one year later are also one year older at schooling level S_T than are people who start school at a normal age. We cannot identify a separate N effect from an A effect.

If we condition on each value of $N = n$ and repeat the preceding identification argument for each N, we identify $\mu(s_T, n)$ and $\lambda(s_T, n)$, s, $s' \geqslant s_T$ from the sample $S \geqslant S_T$ by conditioning on $S_T = s_T$ and $N = n$ in (8). When $N = 0$, $S_T = A$ if schooling is incomplete at the test date ($S > S_T$). We identify a *joint* schooling and age effect for each N. When $N = 1$, we can identify the effect of being one year older on $\mu(s_T)$ and $\lambda(s_T)$ for samples in which $S > S_T$. This effect is indistinguishable from the effect of starting one year later. We can test for an age (at test or entry) effect by testing $\mu(s_T, 0) = \mu(s_T, 1)$ and $\lambda(s_T, 0) = \lambda(s_T, 1)$.[11] This argument can be modified in a straightforward way to account for the case of more than two elements in $\mathcal{N}$.

While intuitively appealing, the method based on control functions does not exploit all of the information in the $S = S_T$ sample. Data where $S = S_T$ is the more commonly occurring case. It is not straightforward to use the control function method to account for ceiling effects. When $S = S_T$, it is possible in principle to isolate separate A, N and S effects. We now present a different method designed to analyze the entire sample more fully.

4. A discrete-continuous econometric model of schooling and test scores

This section develops a more explicitly structured semiparametric model that does not rely on special features of the NLSY data and that enables us to condition more finely. The model also enables us to link our work to more conventional models of schooling and wages, and identify separate S, A and N effects. Initially we assume $S = S_T$ and then we extend the analysis to allow for the case $S > S_T$.

Unlike the control function method developed in Section 3, the method discussed in this section requires more than one test. Suppose that we have data on $K(\geqslant 2)$ tests associated with different levels of schooling $S = s$. Array the tests into a vector

$$T(s,x) = \mu(s,x) + Q(s,x),$$

where the kth component of Q, $Q_k(s,x)$, has a factor structure $Q_k(s,x) = \lambda_k(s) f + \varepsilon_k(s)$, $k = 1, \ldots, K$, $s = 1, \ldots, \bar{S}$ like the one used in Sections 2 and 3. Exact stochastic specifications are given in Section 4.1.

We use the following notation:

$$T(s,x) = \begin{pmatrix} T_1(s,x) \\ \vdots \\ T_K(s,x) \end{pmatrix},$$

[11] Observe that for persons for whom $S = S_T$, age at test is not restricted by (14). Thus we can in principle identify age effects when we use $S = S_T$ observations, but we cannot use the control function method developed in this section to solve the selection problem.

$$\mu(s,x) = \begin{pmatrix} \mu_1(s,x) \\ \vdots \\ \mu_K(s,x) \end{pmatrix},$$

and

$$Q(s,x) = \begin{pmatrix} Q_1(s,x) \\ \vdots \\ Q_K(s,x) \end{pmatrix}.$$

We initially work with Q and produce a semiparametric identification theorem for the distribution of Q and other variables. Then we identify the distributions of the components of the Q. The X are determinants of tests. We assume $Q(s,x) \perp\!\!\!\perp X$ throughout. We observe $T(s,x)$ only if $S = s$. The schooling states in this section can be defined in a sufficiently general way to include different schooling-entry ages as different states. Other definitions for the states are possible (e.g. the Cartesian product of schooling, entry age, schooling quality, etc.), so S can be interpreted in a general way.

In order to account for the endogeneity of schooling, we construct the following model of schooling choice, which we adjoin to the system of test scores:

$$V(S) = \varphi_s(Z) + \eta(s), \quad s = 1,\ldots,\bar{S}, \tag{15}$$

where $V(s)$ is the utility associated with schooling level s, and Z is a vector of determinants of utility. We assume that $\eta = (\eta(1),\ldots,\eta(\bar{S}))$ is absolutely continuous with support $\mathscr{R}^{\bar{S}}$. This joint system of test scores and choice equations is a mixed discrete-continuous choice model as in Heckman (1974a,b). Optimal schooling is $\hat{s} = \arg\max_s \{V(s)\}_{s=1}^{\bar{S}}$. The Z variables may be state-specific or general. Sufficient conditions for nonparametric identifiability of versions of this model are available in the literature.[12] We present a new analysis.

We observe $T(s,x)$ for each schooling level conditional on $\hat{s} = s$. We assume:

$(Q(s,x),\eta) \perp\!\!\!\perp (Z,X)$, *for all* $s = 1,\ldots,\bar{S}$.
The $(Q(s,x),\eta)$ *have zero means and finite variances.* (A-1)

$(Q(s,x),\eta)$ *for all* $s = 1,\ldots,\bar{S}$ *are absolutely continuous with support* $\mathscr{R}^{K+\bar{S}}$. (A-2)

[12] Matzkin (1993) and Thompson (1989) consider the special case where utility functions are identical across choices. In the linear-in-parameters case, they assume $\gamma(s) = \gamma$. See Cameron and Heckman (1998) for a more general analysis.

Under these assumptions, we can write

$$\begin{aligned}&\Pr(T(s,x) < t \mid \hat{s} = s,\ X = x,\ Z = z)\\ &\quad = \Pr(Q(s,x) < (t - \mu(s,x)) \mid V(s) > V(s'),\ s' \neq s, s' = 1,\ldots,\bar{S}),\\ &\quad\quad s = 1,\ldots,\bar{S},\end{aligned} \tag{16}$$

where both t and $T(s,x)$ are vectors.

Adapting an argument from Heckman and Honoré (1990), for each choice $\hat{s} = s$ we can trace out each of the components of $\mu(s,x)$ over their supports for each corresponding component of t up to intercepts which we can obtain by a limit argument presented below. [13]

In this paper we assume the following functional form for utility. For Z a $1 \times J$ vector of variables affecting choices we assume a linear-in-parameters model:

$$\varphi_s(Z) = Z\gamma(s).$$

We define

$$\varphi_{s,s'}(Z) = Z(\gamma(s) - \gamma(s')),$$

and

$$\eta(s,s') = \eta(s) - \eta(s').$$

If the jth coordinate of $\gamma(s)$ is zero, the variable does not affect the sth level of utility. We adopt the notational convention that the first coordinate of Z is the intercept. Array the contrasts of the unobservables into a vector of length $\bar{S} - 1$ where the entry $\eta(s,s)$ $(=0)$ is deleted:

$$\eta_{(s)} = (\eta(s,1),\ldots,\eta(s,\bar{S})).$$

As a consequence of these assumptions, we may write

$$\begin{aligned}&Pr(T(s,x) < t \mid \hat{s} = s,\ X = x,\ Z = z)Pr(\hat{s} = s \mid X = x, Z = z)\\ &\quad = Pr(Q(s,x) < (t - \mu(s,x)), \eta(1,s) < \varphi_{s,1}(z),\ldots,\eta(\bar{S},s) < \varphi_{s,\bar{S}}(z)),\\ &\quad\quad s = 1,\ldots,\bar{S}.\end{aligned} \tag{17}$$

We know the left-hand side of these expressions and seek to determine all of the parameters generating the right-hand side including the joint distribution of the unobservables. We have already established how to identify the components of $\mu(s,x)$ up to intercepts. These can be obtained without assuming any structure for $\gamma(s)$, $s = 1,\ldots,\bar{S}$.

First consider identification of the test system by a limit argument. We assume that the coordinates of the contrast-in-choices vector are "variation free" or more precisely

[13] The easiest way to see how this argument works is to integrate out all components of T except the kth. For different (t_k,x) values, we can trace out pairs that keep the left side of (16) constant. (Recall that we know this CDF.) Applied sequentially, this produces the components of $\mu(s,x)$ up to constants.

that they are measurably separated, so they can be independently varied over their supports:

Support$([\varphi_{s,1}(Z), \varphi_{s,2}(Z), \ldots, \varphi_{s,s-1}(Z), \varphi_{s,s+1}(Z), \ldots, \varphi_{s,\bar{S}}(Z)]) = \mathscr{R}^{\bar{S}-1}$ *for all* $s = 1, \ldots, \bar{S}$, *where the components are measurably separated with respect to each other* ("*variation free*").[14] (A-3)

This assumption says that the support of the difference in the deterministic portions of the contrasts in utility functions matches the support of the corresponding error terms and that we can independently manipulate each argument holding the other arguments fixed.[15] As a consequence of (A-3) and our choice of functional forms for $\varphi_s(Z)$, there exist limit sets $\mathscr{Z}_s$ for each $s = 1, \ldots, \bar{S}$ such that as $Z \to \mathscr{Z}_s$, $Pr(\hat{s} = s \mid Z = z) \to 1$ for $s = 1, \ldots, \bar{S}$. These limit sets can be constructed by making coordinates of Z arbitrarily large or small. In these limit sets, we can identify

$$Pr(Q(s,x) < t - \mu(s,x)) \tag{18}$$

for each $s = 1, \ldots, \bar{S}$. Coordinate by coordinate, we can identify the intercepts of $\mu(s,x)$ since the mean of each coordinate of $Q(s,x)$ exists and is known.[16] For each coordinate, we may form $t_k - \mu_k(s,x)$, $k = 1, \ldots, K$, for each fixed s,x. From (18), in each limit set we may identify the joint distribution of $Q(s,x)$ from the variation in the t_k which traces out the cumulative distribution function of $Q(s,x)$, $s = 1, \ldots, \bar{S}$.[17]

Turning to identification of the choice system, consider choice system s with $\bar{S} - 1$ contrasts

$$V(s) - V(l), \quad l = 1, \ldots, \bar{S}; \ \ l \neq s.$$

Define the set of variables that appear with nonzero coefficients in the s and l utility systems by index sets on the Z and the associated γ coefficients:

$$\mathscr{L}_{c,s,l} = \{j \mid \gamma_j(s) \neq 0 \text{ and } \gamma_j(l) \neq 0\},$$

where $\gamma_j(k)$ is the jth coordinate of the kth system of utility coefficients associated with the jth component of Z. The variables that are common to all (s,l) pairs, $l = 1, \ldots, \bar{S}$, $l \neq s$, are associated with the subscripts in

$$\mathscr{L}_{c,s} = \bigcap_{\substack{l=1 \\ l \neq s}}^{\bar{S}} \mathscr{L}_{c,s,l}.$$

[14] See Florens, Mouchart and Rolin (1990) for a precise definition of measurable separability.

[15] The supports of both can be bounded by straightforward modifications of the initial assumptions. Then we require that the supports of the deterministic functions contain the supports of the error terms.

[16] We can alternatively use a median zero assumption.

[17] Use of these limit sets raises the possibility that identification is achieved only on null sets. Using a version of the argument presented in Aakvik, Heckman and Vytlacil (1999) adapted to this context shows that this possibility is not relevant.

Define the set of unique variables (relative to s, l) as those with nonzero coefficients in s or l but not both:

$$\mathscr{L}_{u,s,l} = \{j \mid \gamma_j(s) = 0 \text{ or } \gamma_j(l) = 0 \text{ but not both}\}.$$

These coefficients are unique within the (s, l) pair (in s or l, but not both). Many intermediate cases may arise where variables are common between s and l but not between s and l', for various l and l' values ($l, l' \neq s$).

Consider the binary choice between s and l. Suppose that (A-3) is satisfied. In particular suppose that for all choices l' ($s, l \neq l'$) apart from s and l there are variables with zero coefficients in $\gamma(s)$ and $\gamma(l)$ with nonzero coefficients in $\gamma(l')$ that have full support in $\mathscr{R}$. This produces (A-3) given our assumed functional form for utility. The following explicit exclusion condition produces identification:

There are nonempty sets of indices
$B_{s,l,l',l''} = \{j \mid j > 1,\ j \notin \mathscr{L}_{c,s,l}, j \notin \mathscr{L}_{u,s,l}, j \in (\mathscr{L}_{c,l',l''} \cup \mathscr{L}_{u,l',l''})\}$ *for all* $l', l'' \neq s, l$.
Thus for some $\gamma_j(l'), \gamma_j(l'')$, *and* $j > 1$, *with zero coefficients in* s *and* l, *the support of the associated* Z_j *is* $\mathscr{R}$, *for all* $l', l'' = 1, \ldots, \bar{S}, l', l'' \neq s, l$. (A-3)′

Setting these variables to limit values, we obtain a limit binary choice model

$$\Pr(V(s) > V(l) \mid Z) = F_{\tilde{\eta}(s,l)}\left(\frac{(\gamma(s) - \gamma(l))Z}{(\sigma(s,l))^{1/2}}\right),$$

where $\sigma(s,l) = \mathrm{Var}(\eta(s) - \eta(l))$ and $\tilde{\eta}(s,l) = \eta(s,l)/(\sigma(s,l))^{1/2}$. By an argument due to Manski (1988), if we assume that

$$Z \in \mathscr{R}^J \text{ is of full rank},^{18} \tag{A-4}$$

we can identify

$$\frac{\gamma_j(s) - \gamma_j(l)}{(\sigma(s,l))^{1/2}}, \quad j \in \mathscr{L}_{c,s,l},$$

and either

$$\frac{\gamma_j(s)}{(\sigma(s,l))^{1/2}} \quad \text{or} \quad \frac{\gamma_j(l)}{(\sigma(s,l))^{1/2}}, \quad j \in \mathscr{L}_{u,s,l},$$

for variables excluded from s or l (but not both). By virtue of (A-3), we can identify the marginal distribution of $\eta(s,l) = \eta(s) - \eta(l)$ up to scale. The mean of this distribution is assumed to be zero. This allows us to identify the intercept of the s, l contrast. In addition, we can identify the marginal distribution of $\eta(s) - \eta(l)$ up to scale, $F_{\tilde{\eta}(s,l)}$, $s = 1, \ldots, \bar{S}$, $l = 1, \ldots, \bar{S}$, $l \neq s$.

[18] Clearly this is a sufficient condition. We only need to have the components of Z with nonzero coefficients possessing full rank, i.e., the components of $\{j \mid j \in (\mathscr{L}_{c,s,\ell} \cup \mathscr{L}_{u,s,\ell})\}$.

We can repeat this argument for each utility contrast (s with $l' \neq l$), and identify either contrasts in parameters (for those common across all utility contrasts) or unique parameters. Using parameters that are unique across the $\mathscr{L}_{u,s,l}$ sets for various s values we can identify ratios of variances from ratios of utility contrasts. For example, suppose that $\gamma_j(s) = 0$ while at the same time for various l values, $\gamma_j(l) \neq 0$. At the same time suppose $\gamma_j(s') = 0$ but $\gamma_j(l) \neq 0$ then we can identify

$$\frac{\gamma_j(l)/\sigma(s,l)^{1/2}}{\gamma_j(l)/\sigma(s',l)^{1/2}} = \left[\frac{\sigma(s',l)}{\sigma(s,l)}\right]^{1/2}.$$

We can repeat this argument for all $s=1,\ldots,\bar{S}$; $l=1,\ldots,\bar{S}$ to identify different combinations of parameters. Depending on the various configurations of $\mathscr{L}_{u,s,l}$, $\mathscr{L}_{c,s',l}$, $s \neq s'$, $l = 1,\ldots,\bar{S}$, $l \neq s$ or s', respectively, we can identify different ratios of variances.

Exclusions of the type just utilized are not strictly required to identify the model. As noted by Cameron and Heckman (1998) and extended by Aakvik et al. (1999), the choice model can be identified with no exclusions if the contrast vectors are linearly independent:

$$[\gamma(s) - \gamma(l)]_{s,l=1,l\neq s}^{\bar{S}} \textit{ is of full rank, and the number of continuous } Z \textit{ variables with support in } \mathscr{R} \textit{ is } \bar{S} - 1 \textit{ or greater.} \tag{A-5}$$

Assumption (A-5) constitutes an alternative to identification by exclusion. The essential idea in this argument is that we can fix each contrast and vary the others (off to limit values), achieving a limit binary choice model. In this case (under (A-4)), we can obtain identification of the marginals $F_{\tilde{\eta}(s,l)}$, $s=1,\ldots,\bar{S}$; $l=1,\ldots,\bar{S}$, $l \neq s$ and the normalized contrasts

$$\frac{\gamma(s) - \gamma(l)}{\sigma(s,l)^{1/2}}, \quad l = 1,\ldots,\bar{S},\ l \neq s,\ s = 1,\ldots,\bar{S}.$$

However in this case, without exclusions, we cannot identify the ratios of variances obtained with exclusions. For details of this argument see Cameron and Heckman (1998) and Aakvik et al. (1999).

From exclusion restrictions or rank conditions on the coefficients of contrast vectors in utilities, we can obtain identification of the choice system and the utility contrasts up to scale. We state a more general result for the joint choice-test score system. We can identify the full joint distribution of $(Q(s,x), \eta_{(s)})$ under the following assumption:

$$\textit{Support}([\varphi_{s,1}(Z), \varphi_{s,2}(Z), \ldots, \varphi_{s,s-1}(Z), \varphi_{s,s+1}(Z), \ldots, \varphi_{s,\bar{S}}(Z), \mu(s,x)]) = \mathscr{R}^{\bar{S}-1+K},\ s = 1,\ldots,\bar{S},\ \textit{an assumption that the components are measurably separated (“variation free”) with respect to each other.} \tag{A-6}$$

This is an assumption which guarantees that we can vary the coordinates of the $(\varphi_{s,s'}(z), \mu)$ freely. We can obtain the $\varphi_{s,s'}(z)$ (up to scale) using either exclusions or rank conditions. Exploiting this assumption, we obtain the following theorem.

Theorem 1. *Under assumptions* (A-1) – (A-4) *and* (A-6), $\mu(s,x)$, $\gamma(s)-\gamma(l)$ (*up to scale* $[\sigma(s,l)]^{1/2}$), $s=1,\ldots,\bar{S}$, $l=1,\ldots,\bar{S}$ *and the joint distributions of* $(Q(s,x),\eta_{(s)})$ (*the second coordinate up to scale*) $s=1,\ldots,\bar{S}$ *are identified.*

Proof. We have already established identification of $\mu_k(s,x)$, $k=1,\ldots,K$, $(\gamma(s)-\gamma(l))/[\sigma(s,l)]^{1/2}$, and the marginal distribution of $\eta(s,s')$ up to scale and joint distributions of $Q(s,x)$. Under (A-6), we can vary each component of $\varphi_{s,s'}(z)$ and $\mu(s,x)$ for each $s'=1,\ldots,\bar{S}$, $s\neq s'$, holding the other components fixed. For all possible values of upper limits, we can trace out the joint distribution of $(Q(s,x),\eta_{(s)})$ nonparametrically. We can do this for all s. ■

With exclusion restrictions we can improve on Theorem 1 by identifying ratios of the scale $[\sigma(s,l)/\sigma(s',l)]^{1/2}$ for some l and s, s' as previously discussed. Note that either (A-3)′ or (A-5) can be used to implement (A-3) but (A-3) is the key condition.

This proof can be adapted to the case where T are indicator functions of latent variables using the argument in Carneiro et al. (2003). Thus we can nonparametrically identify the distribution of the unobservables generating choices and test scores. In addition we can nonparametrically identify the $\mu(s,x)$ and the contrasts in utilities up to scale. We next turn to a factor analysis of the distributions of unobservables.

4.1. Factor models

In this paper we assume that the error term in the utilities has a one-factor specification, [19]

$$\eta(s)=\alpha(s)f+u(s),\quad s=1,\ldots,\bar{S}. \tag{19}$$

Define the $1\times\bar{S}$ vector u as

$$u=(u(1),\ldots,u(\bar{S})),$$

where the $u(s)$ are mutually independent. We now assume $K\geqslant 2$ test scores at each schooling level with a factor structure $Q_k(s)=\lambda_k(s)f+\varepsilon_k(s)$, $k=1,\ldots,K$, so test scores can be written as

$$T_k(s)=\mu_k(s)+\lambda_k(s)f+\varepsilon_k(s),\quad k=1,\ldots,K. \tag{20}$$

The $\mu_k(s)$ may be functions of X as may the $\lambda_k(s)$. For the rest of this section, we keep dependence on X implicit for the sake of notational simplicity. Array these K tests into a vector equation system for each schooling level s:

$$T(s)=\mu(s)+\lambda(s)f+\varepsilon(s),\quad s=1,\ldots,\bar{S}, \tag{21}$$

where $T(s)=(T_1(s),\ldots,T_K(s))$, $\mu(s)=(\mu_1(s),\ldots,\mu_K(s))$, $\lambda(s)=(\lambda_1(s),\ldots,\lambda_K(s))$, and $\varepsilon(s)=(\varepsilon_1(s),\ldots,\varepsilon_K(s))$. We assume that the components of $\varepsilon(s)$ are mutually independent within and across each s and are independent of f.

We assume for the factor structure model:

$$\begin{aligned}&\textit{Independence for the full model}: (X,Z)\perp\!\!\!\perp(f,u,\varepsilon(s));\\ &f\perp\!\!\!\perp u\perp\!\!\!\perp\varepsilon(s),\quad s=1,\ldots,\bar{S}.\end{aligned} \tag{A-7}$$

[19] Heckman (1981) and McFadden (1984) use factor structure error terms for discrete choice models. We extend their models to accommodate both discrete and continuous random variables.

Error terms $u(s)$ for the choice model are mutually independent, with
$$\mathrm{Var}(u(s)) = \sigma^2(s), \quad s = 1, \ldots, \bar{S}. \tag{A-8}$$

Some normalizations are needed for identification of the choice model. One possible normalization is $\sigma^2(s) = 1$. Other normalizations are possible and are developed below.

The input for the factor analysis is the joint distribution of the unobservables produced from Theorem 1. Since we can only identify contrasts in latent utility levels, there are $\bar{S}$ systems with K tests each and $\bar{S} - 1$ utility-normalized contrasts.

The utility contrasts and the test scores form $\bar{S}$ systems of $K + \bar{S} - 1$ random variables to which standard factor analysis (e.g. Anderson and Rubin, 1956) can be applied. Initially we assume no exclusion restrictions so that ratios of variance of $\eta(s) - \eta(l)$ and of $\eta(s') - \eta(l)$ are not known ($s \neq s'$). We develop the case of exclusion restrictions at the end of this section. Under these definitions and normalizations, we obtain from (19) and (20) the following system of covariances for each system $s = 1, \ldots, \bar{S}$:

$$\sigma(s, s') = \mathrm{Var}(\eta(s, s')) = (\alpha(s) - \alpha(s'))^2 \sigma_f^2 + \sigma^2(s) + \sigma^2(s') \tag{22}$$

$$\frac{\mathrm{Cov}(\eta(s, s'), \eta(s, s''))}{\sigma(s, s')^{1/2} \sigma(s, s'')^{1/2}} = \frac{(\alpha(s) - \alpha(s'))(\alpha(s) - \alpha(s''))\sigma_f^2 + \sigma^2(s)}{\sigma(s, s')^{1/2} \sigma(s, s'')^{1/2}},$$

$$s = 1, \ldots, \bar{S}, \quad s \neq s', \quad s''. \tag{23}$$

Recalling that $Q_k(s) = \lambda_k(s) f + \varepsilon_k(s)$, we obtain

$$\frac{\mathrm{Cov}(Q_k(s), \eta(s, s'))}{\sigma(s, s')^{1/2}} = \frac{\lambda_k(s)(\alpha(s) - \alpha(s'))\sigma_f^2}{\sigma(s, s')^{1/2}},$$

$$s' = 1, \ldots, \bar{S}, \quad s' \neq s, \quad k = 1, \ldots, K. \tag{24}$$

$$\mathrm{Cov}(Q_k(s), Q_{k'}(s)) = \lambda_k(s) \lambda_{k'}(s) \sigma_f^2, \quad k \neq k'. \tag{25}$$

The left-hand sides of (23), (24) and (25) are known as a consequence of Theorem 1. If we make one normalization, e.g. $\lambda_1(1) = 1$, and if the conditions of Theorem 1 apply we can identify all of the contrasts $[\alpha(s) - \alpha(s')]/\sigma(s, s')^{1/2}$, $s' = 1, \ldots, \bar{S}$, $s' \neq s$, $s = 1, \ldots, \bar{S}$, and the factor loadings $\lambda_k(s)$, $s = 1, \ldots, \bar{S}$, $k = 1, \ldots, K$, and σ_f^2, provided that $K \geqslant 2$ and $K + \bar{S} - 1 \geqslant 3$.

To see this, suppose $s = 1$. From system (24) with $s = 1$, we may form the ratios

$$\frac{\mathrm{Cov}(Q_k(1), \eta(1, s'))}{\mathrm{Cov}(Q_1(1), \eta(1, s'))} = \lambda_k(1), \quad k = 1, \ldots, K.$$

From (25), for $s = 1$ we can obtain σ_f^2 since we know $\lambda_k(1)$ and $\lambda_{k'}(1)$, for all k, $k' = 1, \ldots, K$, assuming one normalization. From (24), given $\lambda_k(1)$ and σ_f^2 we can obtain $[\alpha(s) - \alpha(s')]/\sigma(s, s')^{1/2}$, $s' = 1, \ldots, \bar{S}$. In this analysis we assume that $\lambda_k(s) \neq 0$, $k = 1, \ldots, K$, $s = 1, \ldots, \bar{S}$. [20]

Turning to the system $s = 2$, armed with σ_f^2, we can identify all factor loadings $\lambda_k(2)$, $k = 1, \ldots, K$, from (24) for $s = 2$, using our knowledge of $[\alpha(1) - \alpha(2)]/\sigma(1, 2)^{1/2}$, and

[20] If this is not so then the effective dimension of the test system is reduced to the number of tests with nonzero factor loadings. A comparable analysis applies to the utility system.

σ_f^2. From (23), for $s=2$, $s' \neq 1$, we can identify $[\alpha(s)-\alpha(s')]/\sigma(s,s')^{1/2}$, $s'=3,\ldots,\bar{S}$. By the same line of reasoning, we can identify all of the $\lambda_k(s)$, $k=1,\ldots,K$, $s=1,\ldots,\bar{S}$.

Using (23), we can identify

$$\frac{\sigma^2(s)}{\sigma(s,s')^{1/2}\sigma(s,s'')^{1/2}}$$

$$=\frac{\mathrm{Cov}(\eta(s,s'),\eta(s,s''))}{\sigma(s,s')^{1/2}\sigma(s,s'')^{1/2}} - \frac{(\alpha(s)-\alpha(s'))(\alpha(s)-\alpha(s''))\sigma_f^2}{\sigma(s,s')^{1/2}\sigma(s,s'')^{1/2}} \tag{26}$$

since we know all of the right-hand side terms either from data or the preceding argument. If we normalize $\sigma(s,s')=1$ and $\sigma(s,s'')=1$ for all s,s', we identify $\sigma^2(s)$, $s=1,\ldots,\bar{S}$.[21] If we normalize $\sigma^2(s)=\frac{1}{2}$, then

$$\sigma(s,s')=(\alpha(s)-\alpha(s'))^2\sigma_f^2+1.$$

We have identified (by the previous argument)

$$\frac{(\alpha(s)-\alpha(s'))\sigma_f}{[\sigma(s,s')]^{1/2}}=\tau(s,s'),$$

where

$$|\tau(s,s')|<1.$$

Thus this normalization is equivalent to the normalization

$$\sigma(s,s')=\frac{1}{1-[\tau(s,s')]^2}>1.$$

When $\bar{S}+K-1<3$, the argument breaks down. Since $\bar{S}=2$ is the minimum number of choices for the system to be interesting, the breakdown comes with one test and two choices.[22] In this case, the only information is in (24) which is

$$\frac{\lambda_1(1)(\alpha(1)-\alpha(2))\sigma_f^2}{[\sigma(1,2)]^{1/2}},$$

$$\frac{\lambda_1(2)(\alpha(2)-\alpha(1))\sigma_f^2}{[\sigma(1,2)]^{1/2}}.$$

Even normalizing $\lambda_1(1)=1$, we can only identify $\lambda_1(2)$ and the combination of parameters $(\alpha(1)-\alpha(2))\sigma_f^2$ up to an unknown scale. Additional normalizations must be made to identify these components separately.

From the joint distribution of (17) we can identify the distribution of f and the distributions of the uniqueness $(\varepsilon_1(s),\ldots,\varepsilon_K(s)$, and $u(s)$, $s=1,\ldots,\bar{S})$. To see why, recall that from Kotlarski's Theorem (1967) that if

$$X_1=Y+Z_1,$$

$$X_2=Y+Z_2,$$

[21] Obviously the choice of these particular normalizations is arbitrary.

[22] In that case we lose the information in (23) and (25).

where $Y \perp\!\!\!\perp Z_1 \perp\!\!\!\perp Z_2$, from the joint distribution of (X_1, X_2) we can identify the distributions of Y, Z_1, Z_2 under a mean zero assumption for Z_1 and Z_2 ($E(Z_1) = 0$; $E(Z_2) = 0$) or for Y ($E(Y) = 0$). From the analysis of Theorem 1 we know the joint distribution of $T(s)$, s=1,…,$\bar{S}$, k=1,…,K. Using (20) and invoking the normalizations previously discussed in the text following Eq. (26), we can write for $\lambda_k(s) \neq 0$,

$$\frac{T_k(s) - \mu_k(s)}{\lambda_k(s)} = f + \frac{\varepsilon_k(s)}{\lambda_k(s)}, \quad k = 1, \ldots, K, \ s = 1, \ldots, \bar{S}.$$

The expression on the left is known since $\lambda_k(s)$, $\mu_k(s)$, $s = 1, \ldots, \bar{S}$, $k = 1, \ldots, K$, are identified by the previous argument. Applying Kotlarski's Theorem we can identify the distribution of f nonparametrically and the distributions of $\varepsilon_k(s)/\lambda_k(s)$, $k = 1, \ldots, K$, $s = 1, \ldots, \bar{S}$, and hence the distributions of $\varepsilon_k(s)$, $k = 1, \ldots, K$, $s = 1, \ldots, \bar{S}$.

From the joint distributions of

$$\frac{\eta(s, s')}{(\sigma(s, s'))^{1/2}} = \left(\frac{\alpha(s) - \alpha(s')}{(\sigma(s, s'))^{1/2}} \right) f + \frac{u(s) - u(s')}{(\sigma(s, s'))^{1/2}}, \quad s' = 1, \ldots, \bar{S}, \ s' \neq s,$$

we obtain a two-factor model with the distribution of the first factor (f) known from the preceding analysis (as well as its factor loading). $u(s)$ is a second factor that is common across all outcomes based on s-contrasts and its factor loading is known by the normalizations previously presented. $u(s')$ is independent of $u(s)$ and $u(s'')$, $s'' \neq s$, s' and f by assumption. Using deconvolution we can remove f from the marginal distributions of $\eta(s, s')/(\sigma(s, s'))^{1/2}$ and apply Kotlarski's theorem to identify the joint distribution of $u(s)$ and $u(s')$, $s' = 1, \ldots, \bar{S}$, $s' \neq s$.[23] The model is strongly overidentified when going across s systems.

Thus far we have not exploited the information available through exclusion restrictions. Suppose that there is at least one variable in $V(s)$ that does not appear in $V(s')$, s, $s' = 1, \ldots, \bar{S}$, $s' \neq s$, with full support ($\mathscr{R}$). Then we can identify $\sigma(s, l)$, $s \neq l$, $s = 1, \ldots, \bar{S}$, $l = 1, \ldots, \bar{S}$ up to a common scale. Thus we can identify the $\alpha(s) - \alpha(s')$ up to a common scale for all s, s'. With this information in hand, fewer normalizations have to be imposed. Thus we can relax one of the normalizations given under Eq. (20). If the exclusions are only partial, we identify various $\sigma(s, l)$ up to different common scales depending on the particular exclusions employed. We do not develop this topic further in this paper.

Recall that we have defined "S" in a general way. It can consist of different combinations of years of schooling completed (S) and age at entry date (N) and other states. Thus we can work with an indicator variable $D(S, N, \ldots)$ that defines schooling states for all $S, N, \ldots$ combinations as discussed in Section 3. This is the model that we estimate.

[23] Since we know the distribution of f from the analysis of the test score data, we can write the density of $\eta(s, s')/(\sigma(s, s'))^{1/2}$ which is known by virtue of Theorem 1 since

$$g_\eta \left(\frac{\eta(s, s')}{\sigma(s, s')^{1/2}} \right) = g_f \left(\frac{\alpha(s) - \alpha(s')}{\sigma(s, s')^{1/2}} f \right) * g_u \left(\frac{u(s) - u(s')}{\sigma(s, s')^{1/2}} \right)$$

where $*$ denotes convolution. We know the first term on the right-hand side. Thus we can form the characteristic functions of $\eta(s, s')/(\sigma(s, s'))^{1/2}$, $([\alpha(s) - \alpha(s')]/(\sigma(s, s'))^{1/2}) f$ and using the inversion theorem identify the density of $[u(s) - u(s')]/(\sigma(s, s'))^{1/2}$, $s' = 1, \ldots, \bar{S}$, $s' \neq s$. For each s system, $u(s)$ constitutes a separate factor apart from f.

4.2. Allowing for tests taken during schooling and age effects

The preceding framework is for the analysis of data on completed schooling ($S=S_T$), where S_T is schooling at the test date. From the assumption that persons who drop out do so only once,[24] and recalling that the age at the test date is A, we obtain $S_T=A-N$ (from (14)) if the individual is still in school at the date of the test. Conditioning on N and A, S_T is a number, not a random variable.

Assuming that sampling is random with respect to A, we can write the density of S_T as the convolution of N (which we model) and a random variable A independent of N (and all other variables) whose distribution we know from the sampling rule. We abstract from any issues of selective survival from mortality since the sample is young.

The density of S_T conditional on $X=x$ and $Z=z$ is

$$g(s_T \mid X=x, Z=z)=\sum_{a=\underline{A}}^{\overline{A}} g_N(a-s_T \mid X=x, Z=z)P(A=a)$$

where $[\underline{A},\overline{A}]$ is the range of survey ages (14–21 in the NLSY data we analyze) and

$$P(A=a)=\frac{1}{\overline{A}-\underline{A}+1}.$$

The density of test scores in the preceding section is conditional on $S_T=S$ (an event which was assumed to hold with probability one). Now we postulate that the event $S_T<S$ (further schooling after the test) may occur. Conditional on $S_T<S$, S_T is a degenerate random variable given A,N. Thus S_T is exogenous given A,N and $S_T<S$ (i.e. $S_T \perp\!\!\!\perp f \mid N,A,S_T<S$).

We may pool the data on S_T for $S_T<S$ with the data on S_T for $S_T \geqslant S$ using this insight. Details about the likelihood for the pooled data are given in Appendix C.

4.3. Accounting for ceiling effects

In the NLSY data a substantial number of test score observations "hit the ceiling," i.e., they achieve the maximum score on a particular test component. This is documented in Table 15 (see Appendix A). To account for these ceiling effects use a latent test score T_k^* so that

$$T_k(s)=\begin{cases} T_k(s) & \text{if } T_k^*(s)<c_k, \\ c_k & \text{if } T_k^*(s)\geqslant c_k, \end{cases}$$

where c_k is the maximum attainable score on test component k. Let the latent test score for an individual with schooling level s at the test date be

$$T_k^*(s)=X\beta_k(s)+\lambda_k(s)f+\varepsilon_k(s),$$

[24] As previously noted this assumption is supported for schooling through high school in Cameron and Heckman (2001).

where X is a set of observed covariates and f is the unobserved factor. Identification with censored random variables can be established by a straightforward modification of Theorem 1 given sufficient support on X.[25]

5. Empirical results

We now present findings from estimating the joint schooling and test score model on the NLSY data discussed in Appendix A. We consider four completed schooling groups: high school dropouts, high school graduates, individuals with some college, and 4-year college graduates. We group GEDs with high school dropouts.[26] We group associate's degrees (junior college graduates) with some college. In addition we group respondents into two categories by age at entry into schooling. Let $N=0$ if an individual began schooling at age 6 or earlier; let $N=1$ otherwise.[27] We estimate a choice model with $4 \times 2 = 8$ potential outcomes (combinations of completed schooling and age at entry).

Over two fifths of the sample (870 individuals, or 42.11% of the sample) had yet to complete high school as of July–October 1980 when the ASVAB was administered. As a consequence we are able to break up this group into three subgroups of schooling level at the test date—those with 9 years of schooling or less (205), those with 10 years of schooling (322), and those with 11 years of schooling including some dropouts with more than 11 reported years of schooling (342). We are thus able to trace out schooling and ability effects for six levels of schooling, including high school graduation and college attendance. Appendix A describes the features of our sample and the variables used to estimate the models.

5.1. Control function estimates

We first present nonparametric estimates from the control function estimators outlined in Section 3. Appendix B describes the econometric procedure used to produce the estimates. It is written for the specific case analyzed in this paper, with six values of S_T and four values of S.

Tables 1 and 2 present estimates for the simple case analyzed at the beginning of Section 3, where we do not control for age effects or endogeneity of entry into schooling. Table 1 reports estimates of the factor loadings $\lambda(s_T)$. Since the model is

[25] A prototype for this proof is in Carneiro et al. (2003), who show how to identify a related model under the case that analysts only observe $1(T_k^*(s) < c_k)$ or $1(T_k^*(s) \geqslant c_k)$. Extension to the censored case is straightforward and for the sake of brevity is omitted here.

[26] The GED is an exam certification for high school equivalency for those who do not earn the degree the traditional route by finishing high school. Our grouping is based on work by Cameron and Heckman (1993).

[27] Of the 1404 individuals in the "normal/ahead" category ($N=0$), 1087 (77.42%) entered school at age 6 and 317 (22.58%) entered school at an earlier age. Since we model choice of schooling and age at entry jointly, further stratifying into 3 age-at-entry categories would produce a model with 12 possible choices and some cells would be very small. Specification checks suggest that combining the "normal" and "ahead" groups is innocuous.

Table 1
Nonparametric estimates of factor loadings

Comparison groups (s,s')								
	(1,2)	(1,3)	(1,4)	(2,3)	(2,4)	(3,4)	MD[a]	χ^2
$\hat{\lambda}(2)$	2.26	1.38	1.00	1.03	0.81	0.68	0.83	6.31
	(1.52)	(0.45)	(0.15)	(0.54)	(0.16)	(0.27)	(0.13)	($p = 0.28$)
$\hat{\lambda}(3)$	0.71	0.73	0.87	0.73	0.89	0.98	0.87	0.83
	(0.72)	(0.27)	(0.13)	(0.40)	(0.16)	(0.34)	(0.12)	($p = 0.98$)
$\hat{\lambda}(4)$				0.89	0.66	0.52	0.61	3.01
				(0.41)	(0.12)	(0.19)	(0.12)	($p = 0.22$)
$\hat{\lambda}(5)$						0.56	0.56	N/A
						(0.20)	(0.20)	N/A

We normalize $\lambda(1) = 1$.
[a]MD = minimum distance.

Table 2
Nonparametric estimates of intercepts and control functions

Intercepts				
$\hat{\mu}(1)$	$\hat{\mu}(2)$	$\hat{\mu}(3)$	$\hat{\mu}(4)$	$\hat{\mu}(5)$
56.91	65.93	71.63	75.24	82.58
(1.27)	(0.95)	(0.78)	(0.61)	(0.63)

Control functions	
$E[f \mid S = \text{dropout}]$	−15.89
	(1.15)
$E[f \mid S = \text{high school}]$	−10.70
	(0.74)
$E[f \mid S = \text{some college}]$	1.95
	(0.99)
$E[f \mid S = \text{college}]$	19.70
	(0.76)
$E[f \mid S_T = \text{9th grade or less}]$	−10.99
	(0.95)
$E[f \mid S_T = \text{10th grade}]$	−2.23
	(0.87)
$E[f \mid S_T = \text{11th Grade}]$	−2.23
	(0.75)
$E[f \mid S_T = \text{high school}]$	−3.63
	(0.56)
$E[f \mid S_T = \text{some college}]$	13.33
	(0.71)
$E[f \mid S_T = \text{college}]$	19.70
	(7.33)

$\chi^2 = 10.61$ ($p = 0.30$)

overidentified we can compute estimates of $\lambda(s_T)$ using information for different completed schooling groups.[28] In Table 1 both the unrestricted estimates and estimates obtained by imposing the overidentifying restrictions using a minimum distance approach are shown.[29] The χ^2-statistics do not reject the overidentifying restrictions. Recalling that $\lambda(1)$ has been normalized to one, the estimates of the remaining $\lambda(s_T)$'s indicate a decreasing effect of latent ability on test scores as schooling at the date of the test increases (the estimates of $\lambda(s_T)$ are decreasing with s_T). Table 2 reports the minimum distance estimates of the intercepts and control functions. Again the χ^2-test fails to reject the overidentifying restrictions implied by the model. Estimated schooling effects (for the average person, with $f=0$) range from 3.61 to 9.02 AFQT points per year of schooling. The estimates imply an expected test score function which is roughly linear in schooling. As expected, the estimates of the control functions (which are conditional expectations of the factor f) are increasing in schooling. The control functions for the different completed schooling categories are clearly statistically different from one another. We identify the scale of f by normalizing $\lambda(1)=1$. Thus, any comparisons are conditional on this normalization, a feature shared with the structural estimates reported below.

We can interpret $\lambda(s_T)[E[f \mid S=s]-E[f \mid S=s']]$ as the expected difference in test scores for two individuals with the same schooling at the test date, s_T, but with different levels of completed schooling. The fact that λ is declining in s_T implies that the test score difference between individuals with different completed schooling levels declines with schooling at the test date. In other words, the test magnifies differences in latent ability at low schooling levels and dampens differences at higher schooling levels.[30]

Tables 3 and 4 present nonparametric control function estimates for the case where f depends on the age of schooling entry, $E(f \mid N=n, S_T=s_T)$, but N does not otherwise enter the model. In this case we can identify $\lambda(\bar{S})$ by differencing test scores conditioning on fixed S_T, $S=\bar{S}$ and varying $N=0,1$. Appendix B.1 presents the estimation procedure used to construct these estimates. Knowing $\lambda(\bar{S})$ we can identify $\mu(\bar{S})$. Table 3 reports the loadings estimated using the minimum distance approach. Again we fail to reject the overidentifying restrictions. In addition, the pattern of declining estimates with schooling is still present. The loading for college $\lambda(6)$ is estimated to be zero. This is most likely caused by the presence of ceiling effects as this pattern is not found to the same extent using the structural model (see next section). Table 4 presents the estimated intercepts and control functions. There is now less evidence for the restrictions implied by the model (the p-value for the χ^2-test is 0.01). However, the estimated test score function (assuming $f=0$) is quite similar to the one estimated without entry effects—especially during the high school years before diverging slightly at the "Some College" level. Estimated schooling effects therefore remain high,

[28] We obtain six different estimates of $\lambda(2)$ by comparing different completed schooling groups following the discussion in Section 3 and in Appendix B.

[29] Since we can only identify ratios of the $\lambda(s_T)$ we have normalized $\lambda(1)$ to one.

[30] This could be due to ceiling effects. The structural model estimates reported in the next section takes ceiling effects into account.

Table 3
Nonparametric estimates of factor loadings controlling for endogenous start date

	Factor loadings	χ^2 and P values
$\hat{\lambda}(2)$	1.23	1.13
	(0.22)	($p = 0.98$)
$\hat{\lambda}(3)$	0.96	1.45
	(0.18)	($p = 0.96$)
$\hat{\lambda}(4)$	0.94	0.01
	(0.23)	($p = 0.92$)
$\hat{\lambda}(5)$	0.41	N/A
	(0.25)	N/A
$\hat{\lambda}(6)$	0.00	N/A
	(0.90)	N/A

Table 4
Nonparametric estimates of intercepts and control functions controlling for endogenous start date

Intercepts

$\hat{\mu}(1)$	$\hat{\mu}(2)$	$\hat{\mu}(3)$	$\hat{\mu}(4)$	$\hat{\mu}(5)$	$\hat{\mu}(6)$
57.13	66.06	72.25	74.62	87.10	95.10
(1.35)	(0.96)	(0.77)	(0.65)	(0.54)	(1.04)

Control functions

$E[f \mid S = \text{dropout},\ N = \text{normal}]$	−11.14	$E[f \mid S = \text{dropout},\ N = \text{behind}]$	−17.03
	(1.34)		(1.38)
$E[f \mid S = \text{high school},\ N = \text{normal}]$	−5.24	$E[f \mid S = \text{high school},\ N = \text{behind}]$	−12.02
	(0.74)		(1.15)
$E[f \mid S = \text{some college},\ N = \text{normal}]$	2.26	$E[f \mid S = \text{some college},\ N = \text{behind}]$	−1.26
	(0.95)		(1.73)
$E[f \mid S = \text{college},\ N = \text{normal}]$	15.80	$E[f \mid S = \text{college},\ N = \text{behind}]$	14.08
	(0.71)		(1.38)
$E[f \mid S_T = \text{9th grade or less},\ N = \text{normal}]$	−6.73	$E[f \mid S_T = \text{9th grade or less},\ N = \text{behind}]$	−12.45
	(2.26)		(1.14)
$E[f \mid S_T = \text{10th grade},\ N = \text{normal}]$	0.84	$E[f \mid S_T = \text{10th grade},\ N = \text{behind}]$	−7.52
	(0.71)		(1.33)
$E[f \mid S_T = \text{11th grade},\ N = \text{normal}]$	1.21	$E[f \mid S_T = \text{11th grade},\ N = \text{behind}]$	−8.23
	(0.86)		(1.22)
$E[f \mid S_T = \text{high school},\ N = \text{normal}]$	−0.52	$E[f \mid S_T = \text{high school},\ N = \text{behind}]$	−6.30
	(0.56)		(1.07)
$E[f \mid S_T = \text{some college},\ N = \text{normal}]$	11.19	$E[f \mid S_T = \text{some college},\ N = \text{behind}]$	7.85
	(0.98)		(1.22)
$E[f \mid S_T = \text{college},\ N = \text{normal}]$	N/A	$E[f \mid S_T = \text{college},\ N = \text{behind}]$	N/A
	N/A		N/A

$\chi^2 = 37.40$ ($p = 0.01$)

between 2.37 and 8.93 AFQT points per year of schooling. The control functions now depend on both schooling and entry age. As expected the estimates are increasing in completed schooling and entry state (individuals who start at an older age have on average lower cognitive ability). Note, however, that entry state has a much smaller effect for the "Some College" and "College" category than for the lower schooling categories.

Table 5 presents estimated intercepts and control functions for a model allowing direct N-effects on the test score in addition to controlling for potential dependence of f on N. As discussed in Section 3, estimating a model controlling for both entry age N and age at the test date A requires more structure in order to break the fundamental identification problem resulting from the confluence of N and A. The estimated intercepts $\mu(N, S_T)$ are uniformly larger for late-starters (who are older when they take the test) than they are for those who begin their schooling at the normal time. [31] Recall, however, that if there are independent age effects, then the difference $\mu(1, S_T) - \mu(0, S_T)$ captures those effects as well as any discouragement effects. As noted in Section 3 we cannot identify an independent age effect. [32] However, we can reject the joint hypothesis of no A and N effects. The evidence points to a much stronger role for age

[31] Note that in order to estimate the model we must restrict $\mu(S_T, 0) = \mu(S_T, 1)$ for some S_T. We report estimates for the model imposing the restriction for $S_T = 5$. To see why we must impose equality between at least one pair of intercepts, note that the moment conditions for this case are (letting $\bar{T}$ denote the conditional mean of T and $c(s, n) = E[f \mid S = s, N = n]$):

$$\bar{T}(s, s_T, n) = \mu(s_T, n) + \lambda(s_T)c(s, n), \qquad \forall s, s_T, n, \tag{27}$$

where $c(s, n) \equiv E[f \mid S = s, N = n]$. In the sample, $n = 0, 1$ and $s = 1, \ldots, 4$. This gives us $(8 - 1) = 7$ control functions to estimate since the weighted sum of the control functions is zero. Note that given s_T and n we have a maximum of four conditions for determining $\mu(s_T, n)$:

$$\bar{T}(s, s_T, n) = \mu(s_T, n) + \lambda(s_T)c(s, n), \qquad s = 1, \ldots, 4. \tag{28}$$

How much data is needed to identify the control functions? Suppose we consider only one s_T value, say $s_T = 1$, and $n = 0, 1$. This yields eight moment conditions:

$$\bar{T}(s, 1, 0) = \mu(1, 0) + \lambda(1)c(s, 0), \qquad s = 1, \ldots, 4,$$

$$\bar{T}(s, 1, 1) = \mu(1, 1) + \lambda(1)c(s, 1), \qquad s = 1, \ldots, 4.$$

Note that under our previous assumptions we had $\mu(1, 0) = \mu(1, 1) \equiv \mu(1)$. If this restriction holds the model is identified, since by taking contrasts we can identify the 7 differences $c(s, 0) - c(4, 1)$ and using the sum restriction on the control functions we get that all of the $c(s, n)$ are identified and then the single intercept $\mu(1)$ is identified. Recall the argument in Section 3.

If we allow for separate intercepts, $\mu(1, 0) \neq \mu(1, 1)$, the model is no longer identified, since we can now only identify the differences $c(s, 0) - c(4, 0)$ and $c(s, 1) - c(4, 1)$. Thus, we can only identify six differences and so we cannot identify the control function elements. Note that this problem persists no matter how many s_T values we use. We can only identify the six differences mentioned above. To obtain the required normalization we can restrict $\mu(s_T, n = 0) = \mu(s_T, n = 1)$ for one s_T value.

[32] Given our "no return to school for dropouts" assumption, people who start school one year later are also one year older at schooling level S_T than are people who start school at a normal age *if they have not completed their schooling at the test date* ($S = S_T$). However, in order to estimate the model we must include individuals with completed schooling at the test date ($S = S_T$) in order to observe the boundary group $S = 1$. Conditioning on the entire sample means that varying N is not equivalent to varying A and, even in the absence of N effects, we cannot identify an independent age effect using this procedure.

Table 5
Nonparametric estimates of intercepts and control functions allowing for N-effects in intercepts and controlling for endogenous start date

Intercepts					
$\hat{\mu}(S_T=1,N=0)$	$\hat{\mu}(S_T=2,N=0)$	$\hat{\mu}(S_T=3,N=0)$	$\hat{\mu}(S_T=4,N=0)$	$\hat{\mu}(S_T=5,N=0)$	$\hat{\mu}(S_T=6,N=0)$
53.68	62.15	69.98	72.30	86.65	95.10
(3.40)	(1.76)	(1.35)	(1.25)	(0.57)	(1.14)
$\hat{\mu}(S_T=1,N=1)$	$\hat{\mu}(S_T=2,N=1)$	$\hat{\mu}(S_T=3,N=1)$	$\hat{\mu}(S_T=4,N=1)$	$\hat{\mu}(S_T=5,N=1)$	$\hat{\mu}(S_T=6,N=1)$
62.81	74.27	77.03	79.40	86.65	95.11
(2.75)	(3.25)	(2.65)	(2.60)	(0.57)	(2.58)

Control functions			
$E[f \mid S=\text{dropout},\ N=\text{normal}]$	−8.14	$E[f \mid S=\text{dropout},\ N=\text{behind}]$	−22.81
	(1.88)		(2.75)
$E[f \mid S=\text{high school},\ N=\text{normal}]$	−2.68	$E[f \mid S=\text{high school},\ N=\text{behind}]$	−17.61
	(1.34)		(2.64)
$E[f \mid S=\text{some college},\ N=\text{normal}]$	4.74	$E[f \mid S=\text{some college},\ N=\text{behind}]$	−6.06
	(1.42)		(2.72)
$E[f \mid S=\text{college},\ N=\text{normal}]$	18.05	$E[f \mid S=\text{college},\ N=\text{behind}]$	9.51
	(1.17)		(2.33)

$\chi^2 = 30.32$

Note: $N=0$ is normal, $N=1$ is behind.

(maturation) in influencing test scores than any discouragement effects from being held back as people who are older at any schooling level have higher test scores.

In the next section we present estimates from the structural, semi-parametric model discussed in Section 4. Taking a structural approach to the problem we can estimate a more general model of schooling and test scores allowing for both age effects, endogenous entry into schooling and testing ceiling effects. We can also condition on covariates such as family background and local labor market variables which may influence the choice of schooling. However, the estimates from the control function approach are in broad agreement with estimates from the structural model.

5.2. Estimates from the structural model

We now present empirical results from the structural model of schooling and test scores presented in Section 4. We use Bayesian MCMC methods to estimate the sample likelihood for the model of Section 4. Details of the algorithm are presented in Appendix C. Our use of Bayesian methods is only a computational convenience. Under our identifying assumptions, the priors we use are asymptotically irrelevant. Our identification analysis is strictly classical.

Table 6 reports exclusion and inclusion restrictions for each equation of the structural model. The common variables in the choice system (included in all but the "college/behind" index) are family background—urban status, broken home status, number of siblings, southern dummy, mother's and father's education, family income—and birth cohort dummies. Choice-specific variables are: local wage and unemployment rate for high school dropouts, high school graduates, and those with some college for

Table 6
Covariates included in structural model

	Utility associated with schooling, entry normal/ahead of peers				Utility associated with schooling, entry behind peers				Test score
	HS dropout	HS grad.	Some coll.	Coll. grad.	HS dropout	HS grad.	Some coll.	Coll. grad.	equations
Intercept	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Yes
Urban dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Yes
Broken home dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Yes
Number of siblings	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Yes
South dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Yes
Mother's education	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Yes
Father's education	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Yes
Family income	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Yes
Cohort dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	—
Local wage for dropouts	Yes	—	—	—	Yes	—	—	—	—
Local dropout unempl. rate	Yes	—	—	—	Yes	—	—	—	—
Local wage for HS grads	—	Yes	—	—	—	Yes	—	—	—
Local HS grad. unempl. rate	—	Yes	—	—	—	Yes	—	—	—
Local wage for some college	—	—	Yes	—	—	—	Yes	—	—
Local some coll. unempl. rate	—	—	Yes	—	—	—	Yes	—	—
Tuition to nearest 4-yr. coll.	—	—	—	Yes	—	—	—	Yes	—
Distance to nearest 4-yr. coll.	—	—	—	Yes	—	—	—	Yes	—
Quarter of birth dummies	—	—	—	—	Yes	Yes	Yes	Yes	—
Age as of December 31, 1980	—	—	—	—	—	—	—	—	Yes
In school at test date	—	—	—	—	—	—	—	—	Yes
Factor	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Yes
Conditional on:	—	—	—	—	—	—	—	—	Schooling at test date

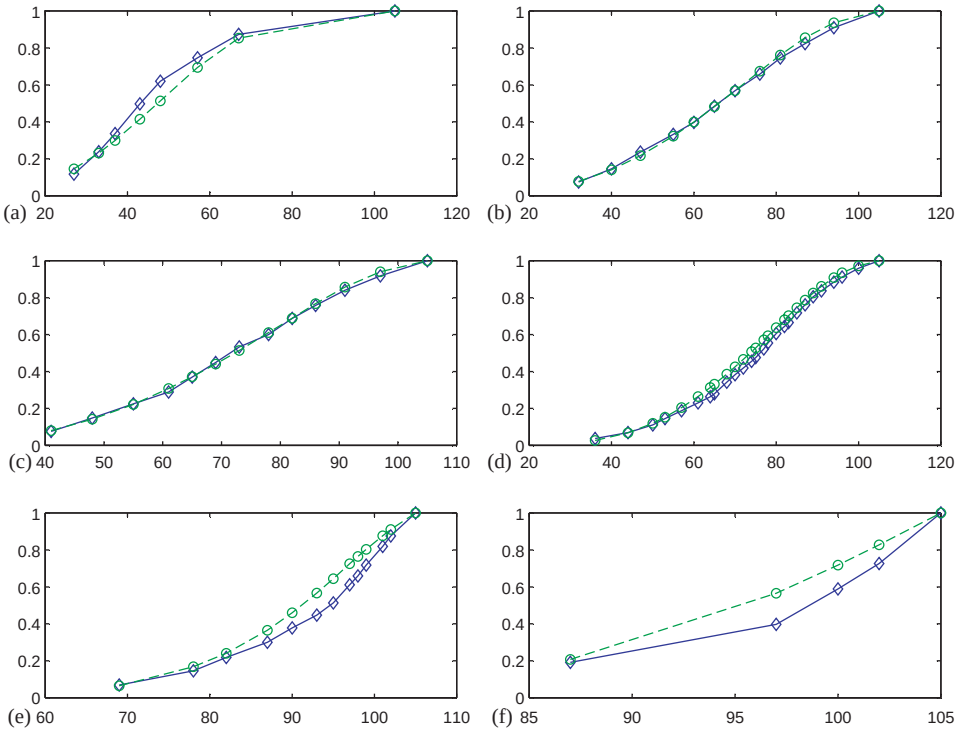

Fig. 1. Actual (diamond) vs. predicted (circle) AFQT cumulative distribution functions conditional on schooling at test date: (a) ninth grade or less; (b) tenth grade; (c) eleventh grade; (d) high school graduate; (e) some college; (f) college graduate.

equations with the corresponding schooling groups, and tuition and distance to 4-year college in the college equations. Quarter-of-birth dummies are included in the "behind" equations. We invoke identification assumption (A-5) because we lack exclusions. We adopt linear-in-parameters utility functions.

We parameterize the latent test score equations as follows:

$$T_k^*(s) = X\beta_k(s) + \lambda_k(s)f + \varepsilon_k(s), \quad k = 1,\ldots,K;\ s = 1,\ldots,\bar{S}_T,$$

where X is a set of observed covariates, including age, which we restrict to have a linear effect. Covariates in the test score equations include family background variables, age (as of December 31, 1980), and a dummy variable for in-school status at the test date. We estimate 24 test equations: four equations for each AFQT component (Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning and Mathematics Knowledge) for each of six levels of schooling at the test date.[33]

[33] In addition to the covariates above we included a dummy variable in the test score equations for having completed strictly less than 9 years of school to allow for possible heterogeneity in the grade school and ninth grade composite group; the coefficient on this dummy was statistically insignificant for all tests.

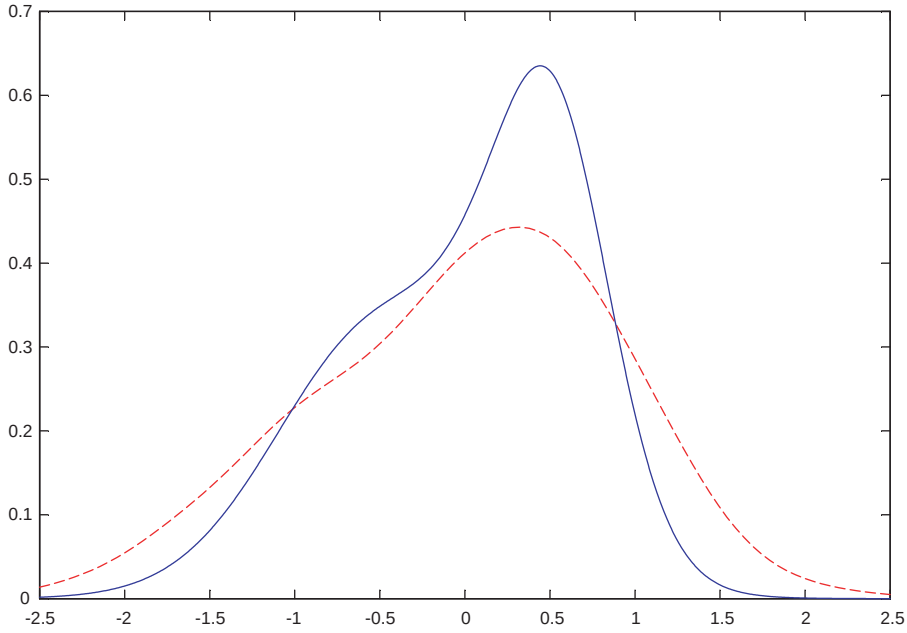

Fig. 2. Estimated factor (solid) vs. residualized AFQT (dashed) distributions.

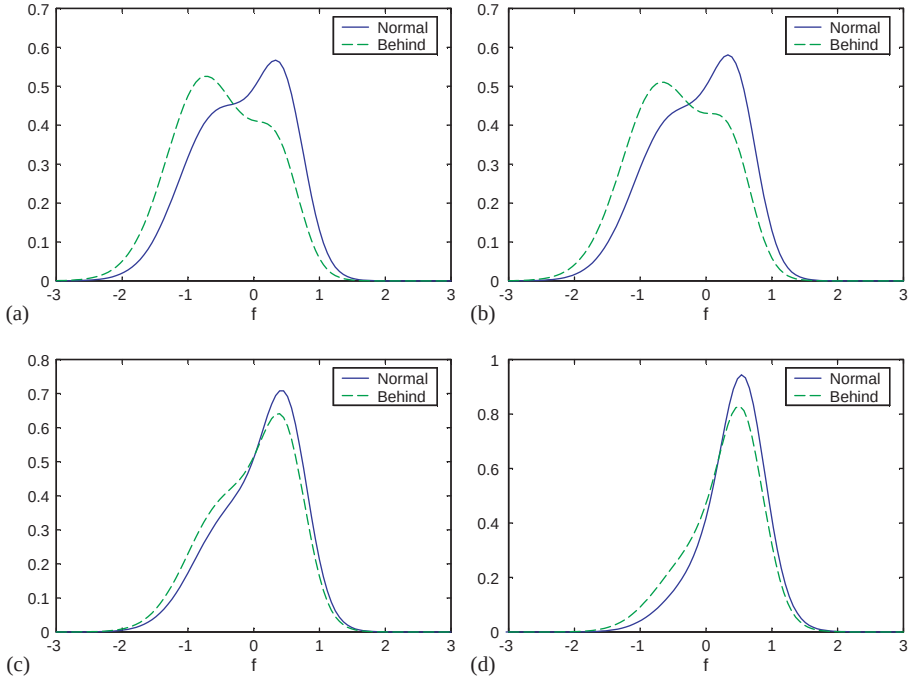

Fig. 3. Factor densities conditional on age at entry by final schooling group: (a) high school dropouts; (b) high school graduates; (c) some college; (d) college graduates.

The computational algorithm used to estimate the model parameters is discussed in detail in Appendix C. Due to space constraints detailed parameter estimates of the models are posted at http://home.uchicago.edu/~kjmullen/Schooling_JOE.htm.

5.2.1. Model fit

We first discuss the fit of the estimated model to the data. Tables 7 and 8 describe the fit of the model to the data for the schooling choice and test systems, respectively. The fit reported in Table 7 is quite good both overall and in partitions of the data on selected covariates. Figs. 1(a)–(f) plot the fitted AFQT test score distribution against the actual empirical CDF for each schooling group. We pass χ^2 goodness of fit tests at conventional levels of significance for most groups though the figures reveal a slight tendency to underpredict scores for the lowest schooling category and to overpredict test scores for the highest two schooling groups. The goodness of fit statistics are reported in Table 8. The fit is worst for the most heterogeneous groups, "ninth grade or less" and "some college." In fact, the poor fit in the "some college" category causes us to fail the overall test of fit.[34] Excluding that group we would pass the overall test.

5.2.2. Estimated cognitive ability distribution

Fig. 2 displays the estimated latent ability or factor distribution plotted against "residualized AFQT" (constructed by running an ordinary least squares regression of standardized AFQT score on family background and cohort dummies). Recall that the location and scale of the latent ability distribution must be set since they are not identified in the model. This is a standard result in factor analysis. Recall that we set the location by constraining the unconditional mean of the factor to be zero (note the residualized AFQT distribution also has mean zero by construction). The scale is set by a normalization in one of the test score equations. Specifically, we set to 1 the coefficient on the factor in the equation for the Word Knowledge test component (standardized to have within-sample mean 0 and variance 1) estimated for individuals who had completed 11 years of schooling at the test date.[35]

The estimated factor density is not normal and closely tracks but does not completely resemble the conventional residualized AFQT density. Residualized AFQT computed by OLS (not accounting for schooling or selection effects) is an imperfect measure of cognitive ability. While the mean of the factor is fixed to 0, the estimated median, 0.1158, is positive, so that more than half of the population has above average ability. However the estimated range of the factor distribution is skewed negative: a person who is at the 2.5th percentile in ability is more than half a standard deviation further away from the average (at -1.4846) than a person at the 97.5th percentile (with ability 1.1131).

[34] The p-value for the overall fit of the model excluding some college is 0.1050.
[35] The estimated standard deviation of the factor is 0.7027.

Table 7
χ^2-statistics for choice model: average choice probabilities in selected groups

		Dropout	HS grad.	Some coll.	Coll. grad.	χ^2-statistic	p-value
Overall ($N = 2066$)							
Normal/ahead	Actual	0.0726	0.2435	0.1331	0.2304		
	Predicted	0.0716	0.2452	0.1359	0.2273		
Behind	Actual	0.0871	0.1060	0.0581	0.0692		
	Predicted	0.0843	0.1052	0.0605	0.0697	0.6775	0.9985
Individuals from urban area ($N = 1553$)							
Normal/ahead	Actual	0.0734	0.2292	0.1301	0.2473		
	Predicted	0.0722	0.2330	0.1335	0.2420		
Behind	Actual	0.0837	0.1011	0.0592	0.0760		
	Predicted	0.0802	0.1007	0.0622	0.0760	0.9032	0.9962
Individuals from rural area ($N = 513$)							
Normal/ahead	Actual	0.0702	0.2865	0.1423	0.1793		
	Predicted	0.0698	0.2822	0.1433	0.1830		
Behind	Actual	0.0975	0.1209	0.0546	0.0487		
	Predicted	0.0968	0.1188	0.0555	0.0504	0.1345	1.0000
Individuals with less than 3 *siblings* ($N = 968$)							
Normal/ahead	Actual	0.0610	0.2231	0.1312	0.3171		
	Predicted	0.0562	0.2383	0.1405	0.2926		
Behind	Actual	0.0589	0.0857	0.0496	0.0733		
	Predicted	0.0607	0.0886	0.0501	0.0721	4.0925	0.7691
Individuals with 3 *or more siblings* ($N = 1098$)							
Normal/ahead	Actual	0.0829	0.2614	0.1348	0.1539		
	Predicted	0.0851	0.2513	0.1319	0.1698		
Behind	Actual	0.1120	0.1239	0.0656	0.0656		
	Predicted	0.1052	0.1198	0.0697	0.0675	3.1838	0.8675
Avg. parents' education < 12 *years* ($N = 803$)							
Normal/ahead	Actual	0.1270	0.2827	0.1009	0.0934		
	Predicted	0.1234	0.2730	0.1173	0.1204		
Behind	Actual	0.1694	0.1469	0.0523	0.0274		
	Predicted	0.1407	0.1365	0.0556	0.0331	13.3411	0.0642
Avg. parents' education $\geqslant 12$ *years* ($N = 1263$)							
Normal/ahead	Actual	0.0380	0.2185	0.1536	0.3175		
	Predicted	0.0386	0.2276	0.1478	0.2953		
Behind	Actual	0.0348	0.0800	0.0618	0.0958		
	Predicted	0.0485	0.0852	0.0636	0.0929	8.3124	0.3059
Four-year college tuition $\leqslant \$2000$ ($N = 1008$)							
Normal/ahead	Actual	0.0863	0.2004	0.1290	0.2500		
	Predicted	0.0791	0.2214	0.1364	0.2370		
Behind	Actual	0.1032	0.0982	0.0605	0.0724		
	Predicted	0.0969	0.0995	0.0568	0.0722	4.4722	0.7241

Table 7 (*continued*)

		Dropout	HS grad.	Some coll.	Coll. grad.	χ^2-statistic	p-value
Four-year college tuition > \$2000 ($N = 1058$)							
Normal/ahead	Actual	0.0595	0.2845	0.1371	0.2117		
	Predicted	0.0644	0.2679	0.1354	0.2181		
Behind	Actual	0.0718	0.1134	0.0558	0.0662		
	Predicted	0.0723	0.1106	0.0641	0.0672	2.9245	0.8919
Zero distance to 4-*year college* ($N = 1552$)							
Normal/ahead	Actual	0.0689	0.2397	0.1308	0.2577		
	Predicted	0.0686	0.2398	0.1357	0.2493		
Behind	Actual	0.0754	0.1018	0.0541	0.0715		
	Predicted	0.0772	0.0983	0.0577	0.0730	1.3616	0.9867
Nonzero distance to 4-*year college* ($N = 514$)							
Normal/ahead	Actual	0.0837	0.2549	0.1401	0.1479		
	Predicted	0.0805	0.2617	0.1365	0.1608		
Behind	Actual	0.1226	0.1187	0.0700	0.0623		
	Predicted	0.1057	0.1259	0.0692	0.0597	2.4023	0.9343

Table 8
χ^2-statistics for predicted AFQT distributions: conditional on schooling level at test date

Schooling level	N	No. bins	χ^2-statistic	p-Value
Ninth grade or less	205	8	16.8349	0.0185
Tenth grade	322	12	7.2653	0.7772
Eleventh grade	343	13	12.3787	0.4158
High school graduate	747	26	30.5079	0.2058
Some college	376	13	35.3758	0.0004
College graduate	73	5	10.3990	0.0342
Overall	2066	77	112.7616	0.0040

Note: Bins were chosen to include approx. equal numbers of observations in each cell. No. bins was chosen to average roughly 25–30 people per bin, except the last group due to small size.

5.2.3. *Allowing for age and endogenous entry dates*

By using the sample $S = S_T$, we break the dependence between A and N given by (14). We parameterize age effects on test scores by assuming

$$\lambda(A,S) = \lambda(S),^{36}$$

$$\mu(A,S) = \beta_1(S)A + \beta_2(S),$$

where $\beta_1(S)$ and $\beta_2(S)$ are unrestricted functions of S. In this paper we explicitly model the relationship between entry date N and latent ability f. The model specifies a joint $S \times N$ space. How important is it to account for endogeneity of N? Fig. 3 plots

[36] Attempts to estimate age-dependent λ led to very imprecise estimates.

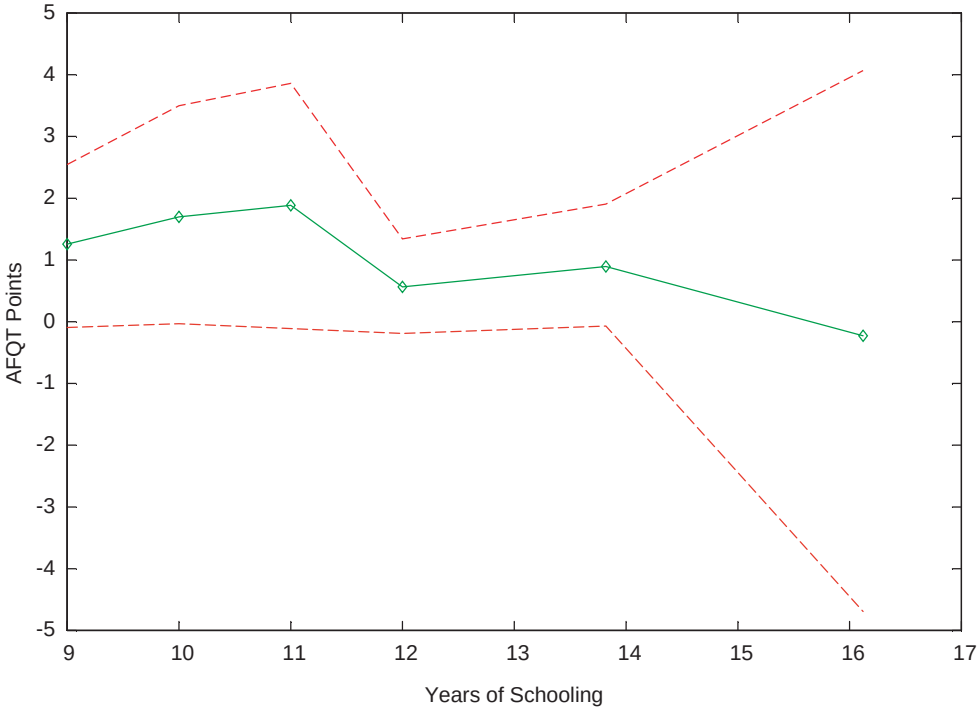

Fig. 4. Estimated effect of one additional year of age on AFQT score (with 95% confidence bands).

the distributions of latent ability f conditional on entry status N. Note that individuals who are behind their peers on average have lower latent ability than their counterparts who are age-grade normal or ahead, especially those who do not attend any college. Failing to correct for endogenous entry effects would lead us to underestimate the effect of cognitive ability on dropping out versus graduating high school, especially at lower levels. [37]

Fig. 4 plots the estimated age at test effects for each $N \in \mathcal{N} = \{0, 1\}$ group. The estimated maturation effects are roughly constant across ages. As in the control function estimates, the net effect of age at the test on measured test scores is positive.

5.2.4. Schooling behavior

The structural approach models the schooling decision explicitly. Thus we can estimate the relationship between cognitive ability f and schooling choice. Correcting for endogenous schooling effects on AFQT turns out to have some interesting implications for inference about the effects of ability on schooling choice. Figs. 5(a) and (b) plot

[37] Note that in the structural model we do not allow for direct N-effects on test scores, which would increase the dimension of the test score system to $2 \times 4 \times 6 = 48$ equations. We do, however, allow for linear age effects in the means (these are graphed in Fig. 4). The evidence from the control function approach outlined in Section 5.1 supports the idea that age effects are more important than "late starter" effects.

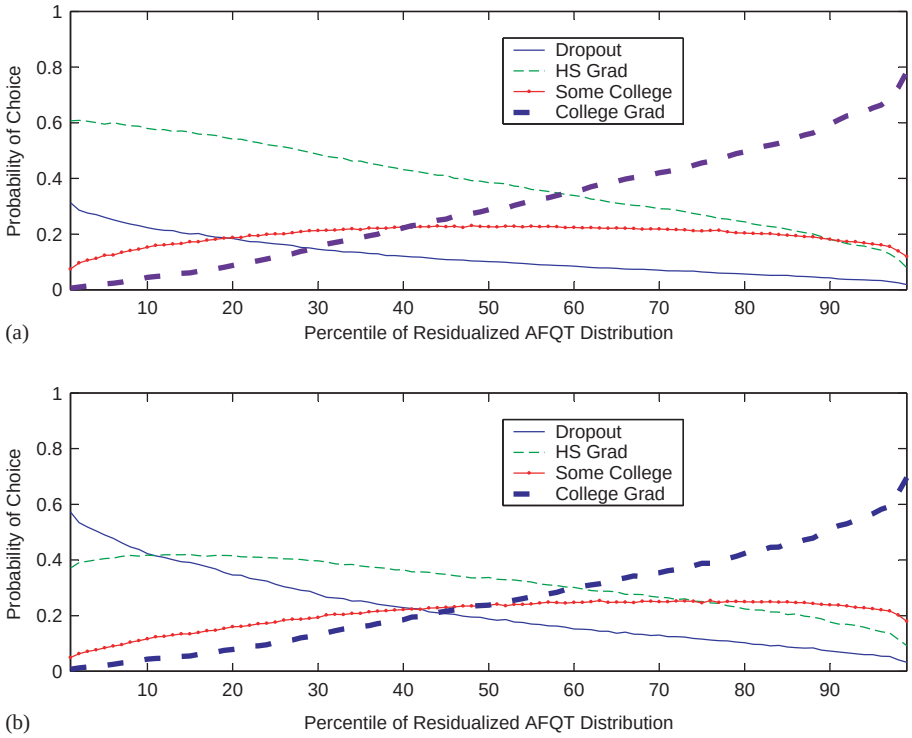

Fig. 5. Schooling choice probabilities as a function of residualized AFQT, no factor: (a) normal/ahead of peers; (b) behind peers.

schooling choice probabilities as a function of observed ability for a simple multinomial choice model that conditions on residualized AFQT (i.e., observed minus predicted AFQT, where predicted AFQT is formed by regressing standardized raw AFQT score on family background characteristics and cohort dummies, not correcting for schooling effects), stratified by entry age. This model assumes that residualized AFQT is a perfect proxy for latent ability so $\sum_{k=1}^{K} T_k(s) - X\beta = \alpha_1 + \alpha_2 f$ where α_1, α_2 are constants. In this conventional specification, measured ability is a strong predictor of schooling decisions, especially high school dropout and college-going decisions. For those who are age-grade normal or ahead of their age-peers in their schooling, the probability of dropping out conditional on a low residualized AFQT score (e.g., a score of -1.8 at the 2.5th percentile) is about 31.9% compared to the population rate of 10.7%. For individuals who are behind their peers the difference is even more pronounced: the predicted probability of dropping out conditional on a score at the 2.5th percentile is 57.7%, compared with a population rate of 27.2%. At the upper end of the AFQT distribution, the estimated probabilities of graduating college with an AFQT score of 1.46 at the 97.5th percentile are 71.2% and 67.5% for the "normal/ahead" and "behind" groups, respectively. The corresponding population rates are 33.9% and 21.6%, respectively.

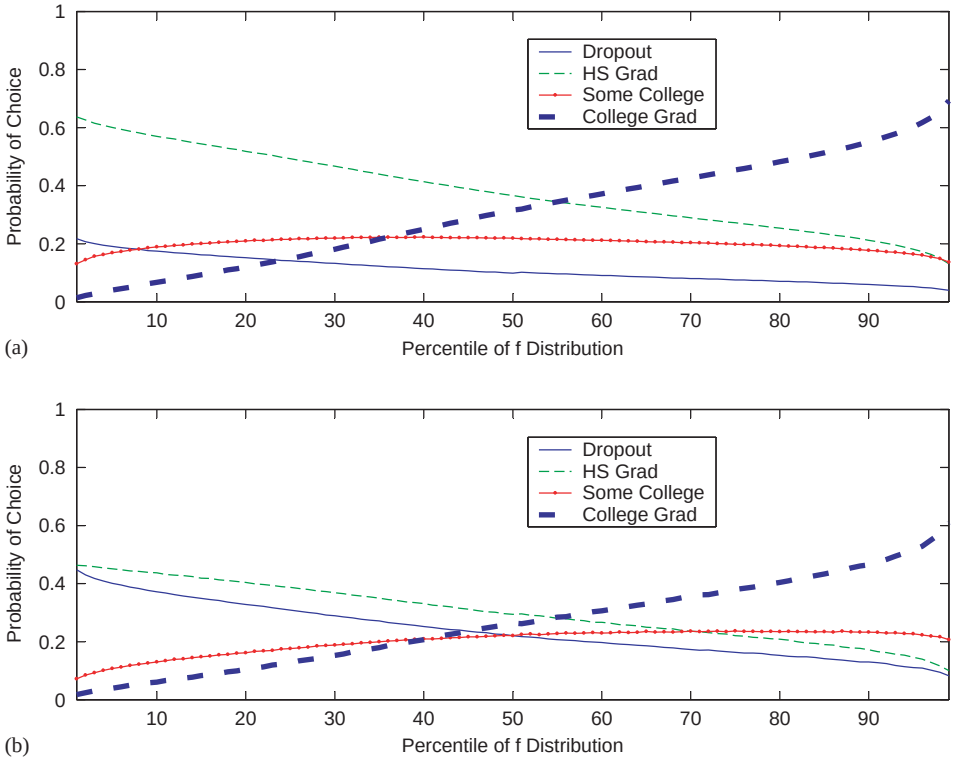

Fig. 6. Schooling choice probabilities as a function of factor: (a) normal/ahead of peers; (b) behind peers.

Figs. 6(a) and (b) plot estimates from the model of Section 4. For small values of the factor f (i.e., at the 2.5th percentile corresponding to the estimated factor distribution) the probabilities of dropping out of high school (21.2% and 42.8%) are almost 11 and 15 percentage points smaller than the comparable probabilities estimated by the simple model above. Larger factor values imply college probabilities of 65.8% and 56.9% which are just over 10 percentage points lower than those estimates produced by the model using residualized AFQT.

Aside from measurement error bias, there is a fundamental econometric problem associated with estimating a schooling choice model which conditions on a measure of ability which has been constructed without accounting for reverse causation (i.e., that schooling affects measured ability). Ignoring the simultaneity problem leads to substantial overstatement of the role of cognitive ability in explaining schooling decisions.

Figs. 7 and 8 demonstrate the importance of this point. Figs. 7(a) and (b) plot the estimated residualized AFQT densities conditional on completed schooling, stratified by entry status. The estimated densities are standardized so that the unconditional density has variance 1 to facilitate comparison with the structural model estimates. A key feature to note is the degree of separation in the conditional "ability" distributions.

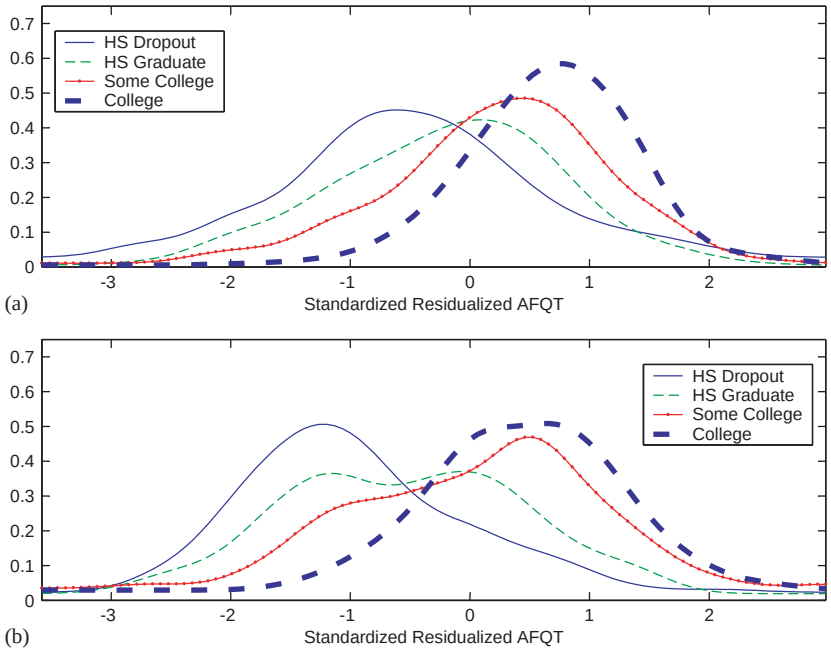

Fig. 7. Kernel estimated densities of standardized residualized AFQT conditional on schooling: (a) normal/ahead of peers; (b) behind peers.

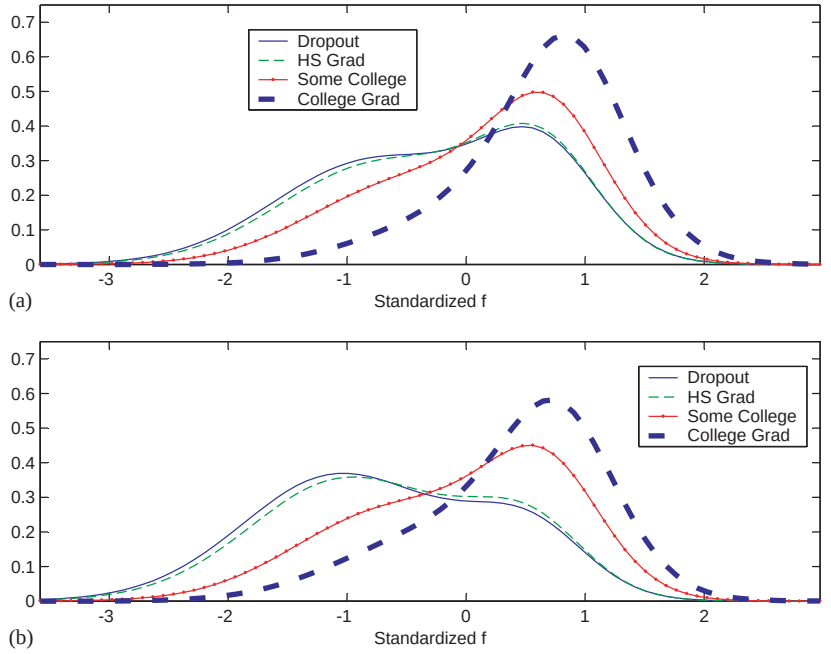

Fig. 8. Standardized factor densities conditional on schooling: (a) normal/ahead of peers; (b) behind peers.

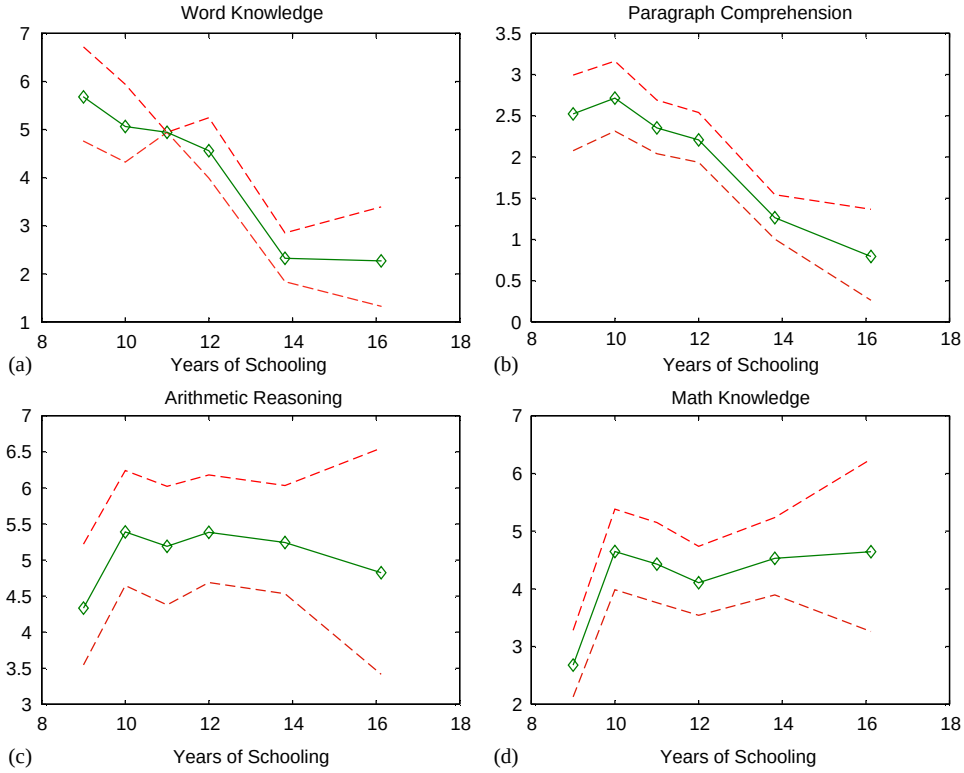

Fig. 9. Marginal effect of standard deviation increase in factor on AFQT components conditional on schooling (with 95% confidence bands): (a) word knowledge; (b) paragraph comprehension; (c) arithmetic reasoning; (d) math knowledge.

Failing to correct for schooling effects on measured ability leads one to predict a strong causal relationship between schooling choice and cognitive ability. Figs. 8(a) and (b) show the estimated factor distributions (again, standardized) estimated from our corrected model conditional on completed schooling and entry date. In the corrected model, the cognitive ability distributions are much less stratified.

Taken together, these findings suggest that the previous literature has overstated the role of latent cognitive ability on explaining schooling. This leaves more room for non-cognitive factors. (See the evidence on the importance of noncognitive factors in Heckman and Rubinstein, 2001.)

5.2.5. *Effect of ability on AFQT*

An important feature of the structural approach developed in this paper which distinguishes it from conventional models of determinants of achievement test scores is that we can estimate the effects of latent ability on manifest test scores in addition to causal schooling effects on test scores. The usual approach treats unobserved ability as a nuisance variable that biases the parameter of interest, the effect of schooling on

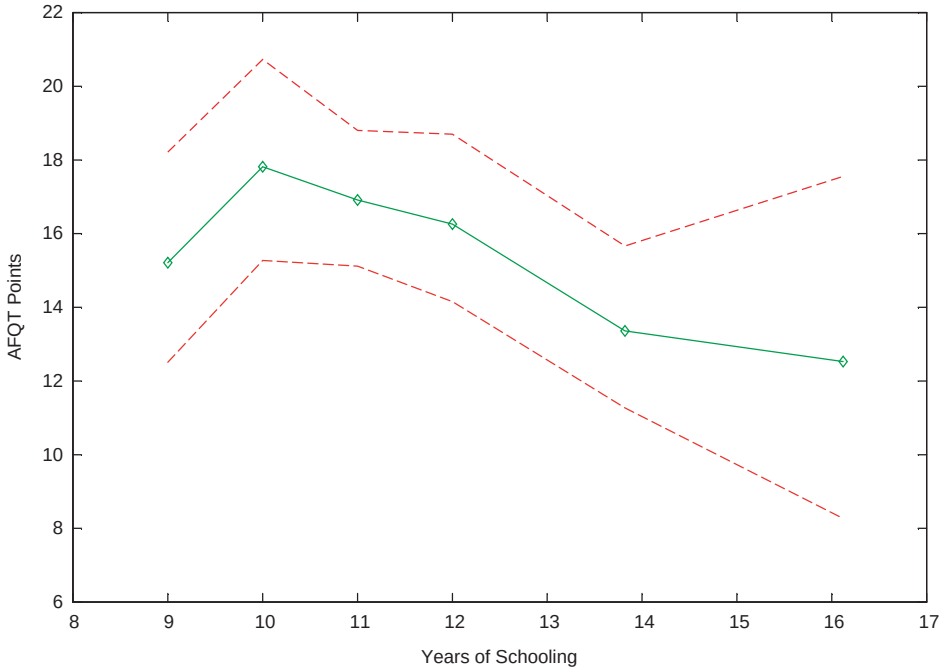

Fig. 10. Marginal effect of a standard deviation increase in latent ability factor on overall AFQT score conditional on schooling level (with 95% confidence bands).

measured ability, and focuses on ways to eliminate its influence. We model latent ability and its influence on schooling explicitly, allowing us to investigate the relationship between latent ability and measured ability.

Figs. 9(a)–(d) show the marginal effect of a standard deviation increase in latent ability on each of the four AFQT test components for different schooling levels at the test date.[38] In several cases the gap in the expected test score between two persons one standard deviation apart in intelligence (unless one of them is at or near the maximum score) is quite large, up to 20% of the total number of points possible on the test. Schooling affects verbal and mathematical skills differently. Moreover, we can see that while the marginal effects of ability decrease with additional schooling for the two verbal test components, the marginal effects of ability on the mathematics components are roughly constant or slightly increasing over schooling levels.

[38] In Fig. 9(a) the collapse of the confidence bands at 11 years of schooling is due to a normalization, i.e., setting $\lambda(S_T = 3) = 1$ in an equation where standardized Word Knowledge (WK) is the dependent variable. Multiplying λ by the standard deviation of the original WK test (7.0327) converts the effect into test score points. Multiplying further by the standard deviation of the latent factor (0.7027) gives the effect in points on the WK test of increasing the latent ability by one standard deviation.

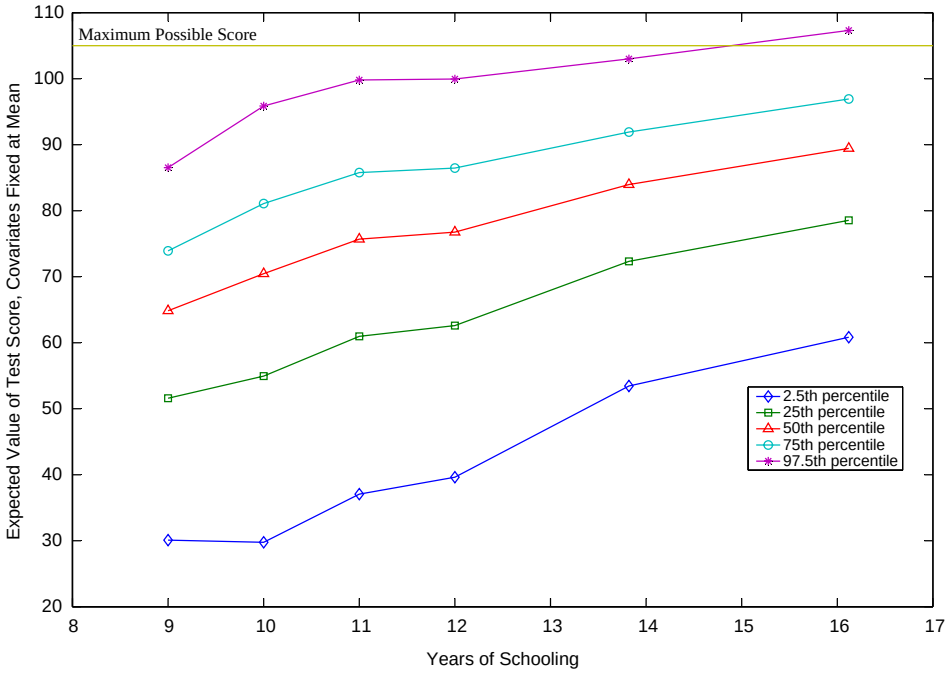

Fig. 11. Effect of schooling on AFQT score conditional on factor.

Fig. 10 shows that the marginal effect of a standard deviation increase in cognitive ability, aggregated over the four test components, ranges from 12.5 to 17.8 AFQT points, or about 12–17% of the maximum possible score of 105. The effect increases initially from ninth to tenth grade where it reaches its peak and appears to fall thereafter. In a formal test we cannot reject the hypothesis that the marginal effect of latent ability on AFQT score is the same across all schooling levels at the 5% significance level. Recall that this is the implicit assumption used by Winship and Korenman (1997), Herrnstein and Murray (1994) and Neal and Johnson (1996).

5.2.6. Effect of schooling on AFQT

In Fig. 11 we switch perspectives and show the effect of schooling on test scores for fixed levels of latent ability. This figure demonstrates strong effects of schooling on test scores. From grade school to college a given individual can expect to improve his performance on the AFQT by about 18–31 points (16–29.5 percentage points), depending on his initial ability level. Fig. 11 shows that the largest schooling effects are found for individuals with very low ability levels. However even with more than 15 years of schooling the test scores of the individuals at the 2.5th percentile do not quite reach the average test score that persons at the median achieve with just a ninth grade education.

Individuals at the very top of the ability distribution (the 97.5th percentile) are within roughly five points of the test score ceiling by 11 years of schooling. The estimated

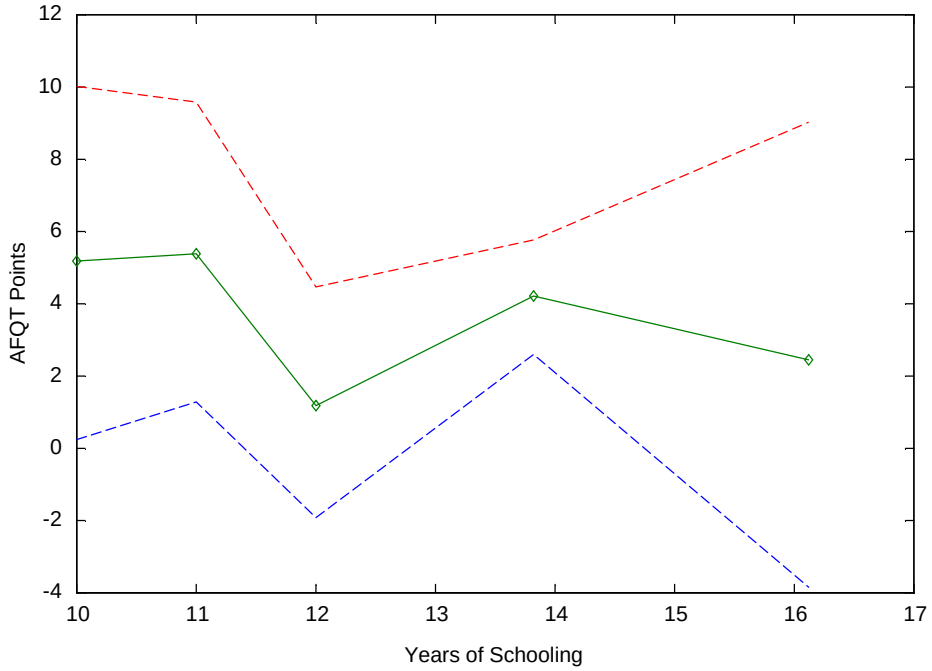

Fig. 12. Estimated annualized school effects for person with average ability (with 95% confidence bands).

AFQT test score functions are roughly parallel across ability levels. In fact the gap in AFQT scores between the 25th and 75th percentiles closes only four points (after widening slightly at first) between ninth grade and the college years. The functions are roughly linear.

Fig. 11 shows that the effects of schooling on AFQT are highest during the early high school years for all ability levels, between grades 9 and 11, between 3.5 and 6.6 points on average per year, varying by ability level. After ninth grade schooling effects decrease with latent ability. The average estimated annual schooling effect, varying across ability levels, is between 3.4 and 4.1 AFQT points (or 3.2–3.9 percentage points). Fig. 12 plots estimated schooling effects with 95% confidence bands for the average person (with $f = 0$), which vary from 1.2 points (transiting from 11th to 12th grade) to 5.4 points (from 9th to 10th grade), with an average schooling effect of 3.7 AFQT points (3.5 percentage points). In other words, a one year increase in schooling is associated on average (across schooling levels) with a 0.16–0.19 standard deviation increase in the AFQT score across ability levels; the estimated increase in AFQT score per year of education for the average person ($f = 0$) is 0.17 standard deviation.

5.3. *Comparison of control function and structural model results*

Figs. 13 and 14 summarize our estimates obtained from using both nonparametric and structural approaches. Fig. 13 plots the estimated ratios of the factor loadings

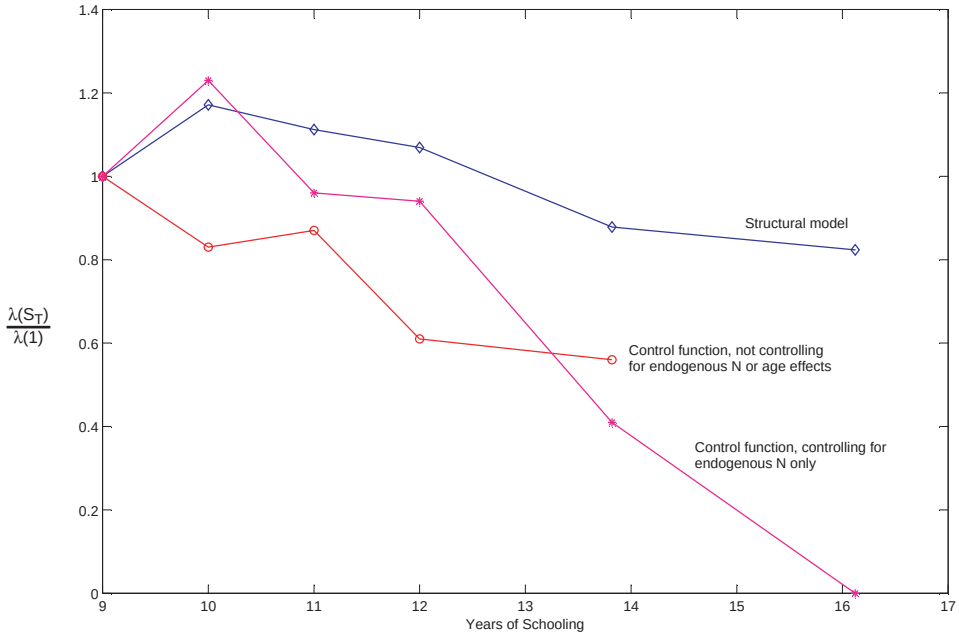

Fig. 13. Comparison of control function and structural estimates of ratios of factor loadings.

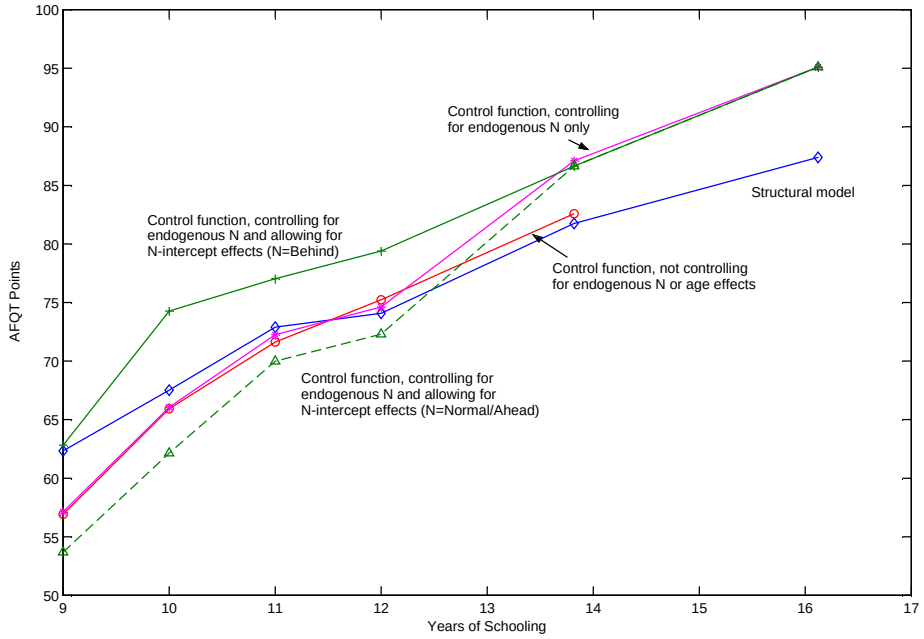

Fig. 14. Comparison of control function vs. structural estimates of expected test score ($f = 0$).

$\lambda(s_T)/\lambda(1)$.[39] Note that the general pattern of declining ability effects is consistent across all models although it is more pronounced for the nonparametric estimates.

Fig. 14 plots estimated intercepts $\mu(s_T)$ for all models.[40] Again the models are in agreement especially for the high school years. The control function estimates (which do not control for other determinants of test scores) tend to be steeper than the structural model estimates which control for family background variables. The control function estimates range from 2.37 to 9.02 AFQT points while the structural estimates vary less and tend to be smaller, ranging from 2.79 to 4.2 points on average. Overall, the agreement is close.

6. Summary, conclusions and applications to the returns to schooling

This paper develops two methods for estimating the effect of schooling on measured test scores when both schooling and measured test scores depend on latent ability. The methods are applied to NLSY data, and produce estimates that are in general agreement with each other.

We find that schooling increases the AFQT score on average between 2.79 and 4.2 points per additional year of education. The effect of schooling on test scores is constant across schooling levels. Our estimates are roughly twice as large as the estimates reported in Herrnstein and Murray (1994). They are in line with the estimates reported in the literature reviewed by Winship and Korenman (1997) who report schooling effects on the order of 2–4 IQ points, or 2.9–5.7 AFQT points. Our analysis shows that schooling has small equalizing effects on measured test scores especially for those with low ability and low levels of schooling. Our analysis also demonstrates that the estimated effect of latent cognitive ability on attending school has been overstated in the previous literature that does not correct for reverse causality between schooling and test scores.

Our analysis also has important implications for the empirical literature on the effects of ability and/or tests on wages. Suppose that

$$\ln W = \alpha_0 + \alpha_1 s + \alpha_2 f + \xi_1(W).^{41} \tag{29}$$

The causal effect of a unit increase in schooling is α_1. A common strategy in the empirical literature on wage equations is to proxy f by T to avoid ability bias arising

[39] The structural model estimates of the factor loadings of the four AFQT components are converted into estimated marginal effects for overall AFQT score, as in Fig. 9.

[40] Recall that the structural model controls for covariate effects. The appropriate comparison to the control function parameter $\mu(s_T)$ is therefore the expected AFQT score evaluated at $f = 0$ and fixing covariates at the mean.

[41] Recall that the scale of f is set by our normalization of a factor loading in the test score measurement system.

Table 9
Estimates from OLS regression of log wage in 1998 on years of schooling and residualized AFQT

Variable	Coefficient	Std. error	t	$p > \|t\|$	95% confidence interval	
Years of schooling in 1998	0.1022	0.0101	10.08	0.000	0.0823	0.1221
Experience	−0.5657	0.0457	−1.24	0.216	−0.1463	0.0332
Experience^2	0.0028	0.0013	2.23	0.026	0.0003	0.0053
OLS-residualized AFQT	0.0988	0.0221	4.48	0.000	0.0555	0.1421
Constant	1.4812	0.4749	3.12	0.002	0.5493	2.4132

Source	SS	Degrees of freedom	MS
Model	62.4505	4	15.6126
Residual	249.9216	977	0.2558
Number of obs. = 982			
$F(4, 977) = 61.03$			
Prob. $> F = 0.0000$			
$R^2 = 0.1999$			
Adjusted $R^2 = 0.1966$			
Root MSE = 0.5058			

Note: Regressions estimated on observations with nonmissing wages.

from the dependence of f on s. Using (1) in Eq. (29) to solve out for f, we obtain

$$\ln W = \alpha_0 + \alpha_1 s + \alpha_2 \frac{(T(s) - \mu(s) - \varepsilon(s))}{\lambda(s)} + \xi_1(W)$$

$$= \alpha_0 + \alpha_1 s - \frac{\alpha_2 \mu(s)}{\lambda(s)} + \frac{\alpha_2}{\lambda(s)} T(s) + \xi_1(W) - \frac{\alpha_2}{\lambda(s)} \varepsilon(s).$$

This is a bad proxy for two reasons: (a) The usual problem that $\alpha_2 \varepsilon(s)/\lambda(s)$ is correlated with $T(s)$; and (b) a novel problem that even if $\varepsilon(s) = 0$, so $T(s)$ is a perfect proxy for f, we acquire additional s-dependent terms arising from the fact that schooling determines test scores, and the estimated marginal effect of schooling on earnings is biased for α_1 unless $\mu(s) = \mu$ and $\lambda(s) = \lambda$ (that is, unless schooling has no effect on test scores).[42] Thus, we get biased estimates of the causal effect of schooling from the proxy even if $\varepsilon(s) \equiv 0$.

Using estimates of the structural model we construct a measure of ability f, $\hat{f}$, correcting for endogenous schooling at the test date (as well as correcting for family background and age effects).[43] Estimates are reported in both Tables 9 and 10. We find

[42] Take, for example, the simple case where $\mu(s) = s\beta$, $\lambda(s) = \lambda$, and $\varepsilon(s) \equiv 0$. Then the estimated coefficient on s is $\alpha_1 - \beta/\lambda$. Including the proxy will lead to downward-biased estimates of α_1 if $\beta > 0$ (assuming the test is positively related to latent ability f, i.e. $\lambda > 0$).

[43] Let the symbol $\hat{}$ denote a consistent estimate of a model parameter (see Appendix C for the estimation algorithm). Then for any individual with characteristics x and schooling at the test date s_T let $\hat{f} = \sum_{k=1}^{K} (\widehat{\lambda_k(s_T)}/\sigma_k(s_T)^2)(T_k - x(s_T)\widehat{\beta_k(s_T)})$ where K is the number of test scores observed.

Table 10
Estimates from OLS regression of log wage in 1998 on years of schooling and schooling corrected ability measure

Variable	Coefficient	Std. error	t	$p > \|t\|$	95% confidence interval	
Years of schooling in 1998	0.1176	0.0094	12.46	0.000	0.0991	0.1362
Experience	−0.0589	0.0461	−1.28	0.202	−0.1494	0.0316
Experience^2	0.0029	0.0013	2.29	0.022	0.0004	0.0054
$\hat{f}$	0.1483	0.0734	2.02	0.044	0.0042	0.2923
Constant	1.2708	0.4758	2.67	0.008	0.3371	2.2045

Source	SS	Degrees of freedom	MS
Model	58.3832	4	14.5958
Residual	253.9890	977	0.2600
Number of obs. = 982			
$F(4, 977) = 56.14$			
Prob. $> F = 0.0000$			
$R^2 = 0.1869$			
Adjusted $R^2 = 0.1836$			
Root MSE = 0.5099			

Note: Regressions estimated on observations with nonmissing wages.

that substituting our schooling-corrected measure of ability for OLS-residualized AFQT in a regression of log wages on years of schooling, ability, experience and experience squared increases the estimated coefficient on schooling by over 1.5 percentage points, from an estimated return of 10.22% to a higher 11.76% in our sample of white males. Previously used measures of ability include the effect of schooling on ability. Purging the measure of ability for this effect results in a substantially larger estimated effect of schooling on wages.

If the true model for wages is instead

$$\ln W = \alpha_0 + \alpha_1 S + \alpha_2 T + \xi_2(W), \tag{30}$$

so that the test score directly affects earnings as a signal of productivity (see, e.g., Altonji and Pierret, 2001), the true marginal return to schooling is

$$\frac{\partial \ln W}{\partial s} = \alpha_1 + \alpha_2 \frac{\partial T}{\partial s}, \tag{31}$$

where

$$\frac{\partial T(s)}{\partial s} = \frac{\partial \mu(s)}{\partial s} + \frac{\partial \lambda(s)}{\partial s} f. \tag{32}$$

Assuming $\alpha_2 > 0$, an approach that ignores the effect of S on T understates the total effect of schooling on wages because it ignores its indirect effect through measured ability. Tests of the relative importance of schooling and signals (T) on wages ignore the effect of S on T. An estimated increase in AFQT score of 3–4 points per year of

Table 11
Estimates from OLS regression of AFQT score on years of schooling at test date

Variable	Coefficient	Std. error	t	$p > \|t\|$	95% confidence interval	
Years of schooling	5.5800	0.2867	19.4600	0.0000	5.0177	6.1423
Urban status	0.2610	0.8569	0.3000	0.7610	−1.4195	1.9415
Broken home	0.9417	0.9729	0.9700	0.3330	−0.9662	2.8496
Number of siblings	−0.5773	0.1898	−3.0400	0.0020	−0.9496	−0.2051
Southern	−2.6561	0.8535	−3.1100	0.0020	−4.3298	−0.9823
Mother's education	1.9069	1.3426	1.4200	0.1560	−0.7261	4.5400
Father's education	8.0419	1.2635	6.3600	0.0000	5.5640	10.5197
Family income	0.1250	0.0299	4.1900	0.0000	0.0665	0.1836
Age	0.1239	0.2562	0.4800	0.6290	−0.3785	0.6264
In school	8.6112	0.9174	9.3900	0.0000	6.8122	10.4103
Constant	−11.9674	4.0363	−2.9600	0.0030	−19.8830	−4.0517

Source	SS	Degrees of freedom	MS
Model	407339.225	10	40733.923
Residual	556311.768	2055	270.711

Number of obs. = 2066
$F(10, 2055) = 150.47$
Prob. $> F = 0.0000$
$R^2 = 0.4227$
Adjusted $R^2 = 0.4199$
Root MSE = 16.453

Note: Instruments for years of schooling: quarter of birth dummies, urban status, broken home, number of siblings, southern residence, mother's education, father's education, family income and cohort dummies.

additional schooling biases downward estimates of the return to schooling on wages by 1.28 to 1.71 percentage points. Accordingly, in their analysis Altonji and Pierret overstate the contribution of signalling to the growth of labor market earnings because they neglect the role of schooling in producing the signal.

These results, and the results reported in Section 5, show that it is important to address carefully the problem of endogenous schooling effects when using measures of cognitive ability. Simply proxying latent ability with an available test score in a wage equation does not solve the problem of ability bias when estimating a return to schooling even if measurement error is zero unless the test score is unrelated to schooling. Similarly, even if the measured test score, as opposed to underlying latent ability, has a causal effect on wages, ignoring schooling effects will lead one to underestimate the effect of schooling on wages. To identify the effects of schooling on test scores it is necessary to control for the endogeneity of schooling decisions. Otherwise, schooling effects on ability are overstated. When we regress the test score on schooling, we get an average effect of 5.58 AFQT points per year of schooling (see Table 11). Using quarter of birth as an instrument, as reported in Table 12, we get a lower effect of 4.52, which is still larger than the estimate from the structural model, although not

Table 12
Estimates from instrumental variables regression of AFQT score on years of schooling at test date

Variable	Coefficient	Std. error	t	$p > \|t\|$	95% confidence interval	
Years of schooling	4.5164	0.9203	4.9100	0.0000	2.7117	6.3212
Urban status	0.2174	0.8605	0.2500	0.8010	−1.4702	1.9050
Broken home	0.8077	0.9823	0.8200	0.4110	−1.1187	2.7341
Number of siblings	−0.6708	0.2054	−3.2700	0.0010	−1.0735	−0.2680
Southern	−2.8466	0.8705	−3.2700	0.0010	−4.5538	−1.1393
Mother's education	2.2952	1.3844	1.6600	0.0970	−0.4197	5.0102
Father's education	8.5196	1.3271	6.4200	0.0000	5.9169	11.1223
Family income	0.1347	0.0310	4.3400	0.0000	0.0739	0.1955
Age	0.7564	0.5799	1.3000	0.1920	−0.3809	1.8936
In school	9.6624	1.2624	7.6500	0.0000	7.1868	12.1381
Constant	−13.1093	4.1571	−3.1500	0.0020	−21.2619	−4.9567

Source	SS	Degrees of freedom	MS
Model	403614.616	10	40361.462
Residual	560036.378	2055	272.524

Number of obs. = 2066
$F(10, 2055) = 114.26$
Prob. $> F = 0.0000$
$R^2 = 0.4188$
Adjusted $R^2 = 0.4160$
Root MSE = 16.508

Note: Instruments for years of schooling: quarter of birth dummies, urban status, broken home, number of siblings, southern residence, mother's education, father's education, family income and cohort dummies.

far from it. Our approach goes beyond the standard IV method to explore the impact of schooling interventions on persons at different places of the latent ability distribution.

Acknowledgements

This research is supported by grants from NICHD-40-4043-000-85-261 and NSF SES-0099195. We thank Chris Winship for a stimulating discussion which influenced this paper. (See his related research reported in Winship, 2001.) We have benefitted from numerous comments by Derek Neal and Chris Winship on various aspects of this paper. We also thank Joseph Kaboski, Salvador Navarro, Sergio Urzua, participants of the Empirical Economics workshop and the Theory and Econometrics workshop at the University of Chicago, and an anonymous referee for helpful comments.

Appendix A. Data

This paper uses data from the National Longitudinal Survey of Youth (NLSY) to estimate the joint model of schooling and test scores presented in Section 3. The NLSY is a representative sample of American young men and women between the ages of 14 and 21 at the time of the first interview in 1979. The NLSY is comprised of three subsamples: (1) a random sample of 6111 noninstitutionalized civilian youths; (2) a supplemental sample of 5295 youths designed to oversample civilian Hispanics, blacks, and economically disadvantaged whites; (3) a sample of 1280 youths who were ages 17–21 as of January 1, 1979, and who were enlisted in the military as of September 30, 1978. The NLSY collects information on parental background, schooling decisions, labor market experiences, cognitive and noncognitive test scores and other behavioral measures on these individuals on an annual basis.

Our analysis is restricted to a sample of 2066 white males from the main subsample for whom there is information on schooling, parental background, and other variables affecting schooling decisions. Parental background may include mother's and father's education, family income, number of siblings, geographic information such as urban status and region of the country in which the family resides, and whether or not the individual comes from a broken home (i.e. non-traditional family). Where information on mother's education, father's education, and/or family income is missing, we impute values for the missing variables. (Exact imputation rules are to be found at our Website.)[44] In addition, direct and implicit (opportunity) costs of schooling are needed. These variables are introduced in the relevant schooling choice equations. These include tuition, distance to school, and local labor market variables such as local wages and unemployment rates (stratified by completed schooling level). Distance to nearest 4-year college is constructed as follows: if there exists a college in the county where a person resides then distance to nearest college is zero; otherwise we compute distance in miles to the nearest county with a college, measuring distance between two counties as the distance between their two centers. This distance is constructed using county of residence at age 17; for those individuals older than 17 in 1979 we use the county of residence in 1979. Tuition at age 17 is the average tuition in colleges in county of residence at age 17. If there is no college in the county, then average tuition in the state is used instead. Local labor market variables for the county of residence are gathered from the 5% sample in the 1980 census. We compute local unemployment rates and average local wages for high school dropouts, high school graduates, and individuals with some college. We assume that the 1980 variables are a close proxy for local labor market conditions in the years in which NLSY respondents are assumed to be making the schooling decisions analyzed in this paper. Tables 13 and 14 present means and standard deviations for the variables stratified by final schooling and by schooling at the test date, respectively, and overall.

[44] http://athens.src.uchicago.edu/jenni/JOE/

Table 13
Sample means by final attainment status NLSY, white males

Variables	Overall	By final years of education			
		HS dropout	HS graduate	Some college	College graduate
No. observations	2066	330	722	395	619
Urban dummy	0.7517	0.7394	0.7105	0.7443	0.8110
	(0.4321)	(0.4396)	(0.4538)	(0.4368)	(0.3918)
From broken home	0.1999	0.4000	0.1814	0.1924	0.1196
	(0.4000)	(0.4906)	(0.3856)	(0.3947)	(0.3247)
Number of siblings	2.9942	3.6900	3.0997	3.0152	2.4862
	(1.9805)	(2.3302)	(1.8728)	(2.1147)	(1.6559)
Southern dummy	0.2493	0.3818	0.2119	0.2506	0.2213
	(0.4327)	(0.4866)	(0.4089)	(0.4339)	(0.4155)
Mother's education	12.1343	10.5019	11.5817	12.2393	13.4105
($N = 1884$)	(2.3326)	(2.3080)	(1.9751)	(1.9270)	(2.2662)
Father's education	12.4202	10.1341	11.4407	12.6711	14.4179
($N = 1942$)	(3.3135)	(3.0890)	(2.7352)	(2.9093)	(3.1170)
Family income (thousands)	22.3244	14.7412	20.7361	22.5783	28.1310
($N = 1695$)	(14.3756)	(9.7335)	(11.6325)	(13.2189)	(17.5746)
Born in first quarter	0.2464	0.2364	0.2493	0.2658	0.2359
	(0.4310)	(0.4255)	(0.4329)	(0.4423)	(0.4249)
Born in second quarter	0.2483	0.2606	0.2396	0.2329	0.2617
	(0.4321)	(0.4396)	(0.4271)	(0.4232)	(0.4399)
Born in third quarter	0.2672	0.2727	0.2659	0.2709	0.2633
	(0.4426)	(0.4460)	(0.4421)	(0.4450)	(0.4408)
Behind peers	0.3204	0.5455	0.3033	0.3038	0.2310
	(0.4668)	(0.4987)	(0.4600)	(0.4605)	(0.4218)
Born in 1957	0.1026	0.0970	0.0983	0.1266	0.0953
	(0.3035)	(0.2964)	(0.2980)	(0.3329)	(0.2939)
Born in 1958	0.0978	0.0606	0.0886	0.1139	0.1179
	(0.2971)	(0.2390)	(0.2844)	(0.3181)	(0.3228)
Born in 1959	0.1094	0.1000	0.1260	0.1038	0.0985
	(0.3122)	(0.3005)	(0.3321)	(0.3054)	(0.2983)
Born in 1960	0.1336	0.1394	0.1482	0.1114	0.1276
	(0.3403)	(0.3469)	(0.3555)	(0.3150)	(0.3339)
Born in 1961	0.1317	0.1272	0.1343	0.1392	0.1260
	(0.3382)	(0.3338)	(0.3413)	(0.3466)	(0.3321)
Born in 1962	0.1641	0.1667	0.1634	0.1671	0.1616
	(0.3704)	(0.3732)	(0.3700)	(0.3735)	(0.3683)
Born in 1963	0.1389	0.1636	0.1316	0.1215	0.1454
	(0.3459)	(0.3705)	(0.3383)	(0.3271)	(0.3528)
Local dropout wage	6.5651	6.5853	6.5993	6.5913	6.4976
	(1.2256)	(1.3347)	(1.2512)	(1.2165)	(1.1377)
Local dropout	0.0697	0.0684	0.0718	0.0710	0.0672
Unemployment rate	(0.0231)	(0.0237)	(0.0231)	(0.0240)	(0.0219)
Local HS graduate wage	7.5509	7.5600	7.5186	7.5742	7.5689
	(1.4599)	(1.5218)	(1.3438)	(1.4226)	(1.5780)
Local HS unempl. rate	0.0573	0.0552	0.0609	0.0588	0.0531
	(0.0254)	(0.0266)	(0.0254)	(0.0264)	(0.0235)
Local wage for some	7.6666	7.6679	7.6692	7.7600	7.6033
college	(1.4020)	(1.4245)	(1.4148)	(1.4042)	(1.3733)

Table 13 (*continued*)

Variables	Overall	By final years of education			
		HS dropout	HS graduate	Some college	College graduate
Local unempl. rate for some college	0.0371 (0.0156)	0.0355 (0.0160)	0.0395 (0.0155)	0.0378 (0.0162)	0.0347 (0.0148)
4-year college tuition (tens)	19.8694 (7.8463)	18.0718 (7.2133)	21.3384 (7.7863)	19.5033 (8.3439)	19.3481 (7.6350)
Distance to 4-year college	8.1149 (16.4639)	10.4564 (19.9424)	8.3852 (15.9885)	9.2578 (16.8611)	5.8219 (14.3320)
Word knowledge	26.8930 (7.03266)	20.1000 (8.1892)	25.1911 (6.5495)	28.6886 (5.0356)	31.3538 (3.6514)
Paragraph comprehension	10.9719 (3.3182)	7.7848 (3.5822)	10.1801 (3.2502)	11.8152 (2.5227)	13.0565 (1.6166)
Arithmetic reasoning	19.7333 (7.2253)	13.0242 (5.8957)	17.6191 (6.5290)	20.9747 (5.9915)	24.9838 (5.0460)
Math knowledge	14.6438 (6.5385)	8.3272 (4.0018)	11.9834 (5.0292)	15.5873 (5.5759)	20.5121 (4.5122)
Overall AFQT	72.2420 (21.6023)	49.2364 (18.7146)	64.9737 (18.1901)	77.0658 (16.0706)	89.9063 (12.2652)

Table 14
Sample means by education at test date NLSY, white males

Variables	Overall	By years of education at test date (July–October 1980)					
		$\leqslant 9$	10	11	HS graduate	Some college	College graduate
No. Observations	2066	205	322	343	747	376	73
Urban dummy	0.7517 (0.4321)	0.7171 (0.4515)	0.7360 (0.4415)	0.7405 (0.4390)	0.7349 (0.4417)	0.8138 (0.3898)	0.8219 (0.3852)
From broken home	0.1999 (0.4000)	0.3805 (0.4867)	0.2671 (0.4431)	0.2187 (0.4139)	0.1620 (0.3687)	0.1197 (0.3250)	0.1096 (0.3145)
Number of siblings	2.9942 (1.9805)	3.7366 (2.4968)	2.9534 (1.8897)	2.9767 (1.9330)	2.9411 (1.8523)	2.8830 (2.0375)	2.2877 (1.3487)
Southern dummy	0.2493 (0.4327)	0.4341 (0.4969)	0.2484 (0.4328)	0.2449 (0.4307)	0.2129 (0.4096)	0.2340 (0.4240)	0.2055 (0.4068)
Mother's education ($N = 1884$)	12.1343 (2.3326)	10.4892 (2.4653)	11.7412 (2.0647)	12.0709 (2.3630)	11.9636 (2.1499)	13.0822 (2.1671)	13.7455 (2.0925)
Father's education ($N = 1942$)	12.4202 (3.3135)	9.9712 (3.3274)	12.0570 (3.2026)	12.0709 (3.1889)	12.0971 (3.0071)	13.8783 (3.1258)	14.2364 (3.3331)
Family income (thous.) ($N = 1695$)	22.3244 (14.3756)	15.0775 (9.0086)	20.1549 (12.0373)	22.7529 (13.1427)	22.7557 (13.0989)	27.9788 (18.1897)	25.9860 (17.8028)
In school at test date	0.5034 (0.5001)	0.3902 (0.4890)	0.7702 (0.4214)	0.7055 (0.4565)	0.2249 (0.4178)	0.7261 (0.4466)	0.3973 (0.4927)
Born in 1957	0.1026 (0.3035)	0.0634 (0.2443)	0.0248 (0.1559)	0.0321 (0.1764)	0.1098 (0.3128)	0.1622 (0.3692)	0.5068 (0.5034)
Born in 1958	0.0978 (0.2971)	0.0293 (0.1690)	0.0155 (0.1238)	0.0350 (0.1840)	0.1017 (0.3025)	0.1835 (0.3876)	0.4658 (0.5023)
Born in 1959	0.1094 (0.3122)	0.0634 (0.2443)	0.0248 (0.1559)	0.0437 (0.2048)	0.1446 (0.3519)	0.2128 (0.4098)	0.0274 (0.1644)

Table 14 (*continued*)

Variables	Overall	By years of education at test date (July–October 1980)					
		⩽ 9	10	11	HS graduate	Some college	College graduate
Born in 1960	0.1336	0.0585	0.0404	0.0816	0.1754	0.2447	0.0000
	(0.3403)	(0.2353)	(0.1971)	(0.2742)	(0.3805)	(0.4305)	(0.0000)
Born in 1961	0.1317	0.0780	0.0373	0.0816	0.1954	0.1862	0.0000
	(0.3382)	(0.2689)	(0.1897)	(0.2742)	(0.3968)	(0.3898)	(0.0000)
Born in 1962	0.1641	0.1561	0.0776	0.2391	0.2637	0.0080	0.0000
	(0.3704)	(0.3638)	(0.2680)	(0.4271)	(0.4409)	(0.0891)	(0.0000)
Born in 1963	0.1389	0.1415	0.2609	0.4840	0.0094	0.0027	0.0000
	(0.3459)	(0.3494)	(0.4398)	(0.5005)	(0.0964)	(0.0516)	(0.0000)
Word knowledge	26.8930	18.4488	24.2888	26.2828	27.5636	31.7766	32.9452
	(7.03266)	(7.2581)	(7.3532)	(6.3737)	(5.9219)	(3.5437)	(2.2292)
Paragraph comprehension	10.9719	7.0878	9.8758	10.9679	11.2503	13.0000	13.4384
	(3.3182)	(3.3302)	(3.6382)	(3.1589)	(2.8090)	(1.8504)	(1.2582)
Arithmetic reasoning	19.7333	12.0683	17.0870	18.9650	19.9545	25.0532	26.8767
	(7.2253)	(5.6755)	(6.9550)	(6.7120)	(6.4590)	(4.9727)	(3.7228)
Math knowledge	14.6438	0.3805	12.7702	13.7843	14.1245	20.1197	21.8493
	(6.5385)	(0.4867)	(6.3278)	(6.1909)	(5.7016)	(4.7017)	(3.7256)
Overall AFQT	72.2420	3.7366	64.0217	70.0000	72.8929	89.9495	95.1096
	(21.6023)	(2.4968)	(21.5587)	(19.3215)	(17.8026)	(12.4515)	(8.9047)

Table 15
ASVAB test information: AFQT subcomponents

	Word knowledge	Paragraph comprehension	Arithmetic reasoning	Math knowledge	Overall AFQT
Number of questions	35	15	11	25	86
Time (in min)	11	13	36	24	84
Max. possible raw score	35	15	30	25	105
Fraction of observations censored in each subsystem					
Schooling at test date	Word knowledge	Paragraph comprehension	Arithmetic reasoning	Math knowledge	At least 1 AFQT component
9 years or less	0.0000	0.0098	0.0049	0.0000	0.0146
10 years	0.0155	0.0311	0.0248	0.0217	0.0745
11 years	0.0496	0.0787	0.0437	0.0408	0.1370
12 years/HS graduate	0.0535	0.0656	0.0348	0.0388	0.1392
Some college	0.2128	0.1649	0.1622	0.1303	0.4521
College graduate	0.3151	0.1781	0.2740	0.2192	0.5890
Overall	0.0799	0.0789	0.0634	0.0557	0.1893

In 1980, NLSY participants were administered a series of achievement tests known as the Armed Forces Vocational Aptitude Battery (ASVAB). The math and verbal scores of the ASVAB can be aggregated into a measure called the Armed Forces Qualifica-

tion Test (AFQT). These include tests of Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning and Mathematics Knowledge. In our application AFQT is constructed as the sum of these four tests. Table 15 presents the number of questions, allotted completion time, and maximum possible score for each test. In addition, the fraction of individuals who attain the maximum possible score on each test are presented, overall and by schooling group at the test date. Accounting for top censoring by modeling the AFQT distribution as the sum of censored-normal subcomponents is empirically relevant: while only seven individuals out of 2066 (less than half of 1%) achieve the maximum possible AFQT score (by achieving the maximum score on all four test components), 19% of the sample (391 individuals) attain the maximum score on at least one of the four tests. The table reveals that accounting for top censoring is more important for people with higher levels of schooling at the time of the test; in some cases 20–30% of the individuals in those groups are censored on one or more components.

The NLSY contains longitudinal data on highest grade completed, year last enrolled in school (if not currently enrolled), high school degree or equivalency status, type of degree (diploma or GED) for the years 1979–2000 as well as highest degree attained and year highest degree attained after 1988. Final schooling categories were constructed primarily using degree information from last year observed provided that the respondent was 21 or older if the final state was coded as high school dropout or high school graduate or that the respondent was 25 or older if the individual attended some college. GED recipients were classified as high school dropouts. For those individuals without specific degree information, the highest grade completed variable was used. For the remaining 2% of the sample, the age restriction was relaxed provided the last year the respondent was enrolled was 2 years prior to last observed schooling state. The age restriction was placed to ensure that individuals who were actually censored were not mistakenly included in the sample; for example an individual who dropped out of the sample at age 18 with a high school degree may have gone on to attend some college or complete a 4-year degree and should not be coded as a high school graduate. In addition, 53 cases were discarded from the sample due to inconsistent schooling history or lack of sufficient information to conclude schooling status (final or at the test date, see below).

Schooling level and enrollment status at the test date were constructed as follows. The ASVAB was administered during July–October 1980. Respondents were interviewed during January–August 1980 and again in January–July 1981. Note also that the NLSY constructs a measure of schooling and enrollment status as of May 1 of each survey year. Since the academic year commonly ends in June (May for college), individuals typically advance to a higher completed grade level in May/June. We use highest grade completed and enrollment status as reported in the 1980 survey as measures of schooling and enrollment at the test date if the interview was conducted during July–August 1980, otherwise we use the variables reported in 1981 if the survey was conducted during January–April 1981. For those remaining we use the NLSY-constructed variables for May 1, 1981. We re-coded schooling state at the test date for 32 individuals to be compatible with final schooling state (mostly changing highest grade completed at the test date to 11 for high school dropouts). For the re-

maining 1% of the sample we used schooling and enrollment histories to come up with plausible categories for schooling at the test date.

In addition to schooling categories, measures of age-at-entry group were constructed. For those individuals who finished school before 1979 the survey asks the date at which they were last enrolled and the highest grade they had completed at that date. Recall that we assume continuous schooling profiles so that there are no skips or breaks in schooling from age at initial entry forward. We constructed our measure of age at initial entry date as follows. For those individuals enrolled in school in 1979, we let age at initial entry date equal years of schooling completed in 1979 minus age in 1979. For those individuals who had finished school prior to 1979 we made the same calculation using highest grade completed and age at last date of enrollment. In our empirical work we constructed two categories of endogenous entry status: "behind" if age at initial entry is greater than 6 years (the median age of entry), "normal" if age at initial entry is 6 years or less.

Appendix B. Estimation procedure for control function model

B.1. Random entry

Consider the case in which we assume no age effects and random entry into schooling. We group individuals into six categories of schooling at test date and four categories of completed schooling. The combinations of schooling at test date and final schooling are represented by the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ \sim & a_{42} & a_{43} & a_{44} \\ \sim & \sim & a_{53} & a_{54} \\ \sim & \sim & \sim & a_{64} \end{pmatrix}, \tag{B.1}$$

where a_{ij} represents the average test score for individuals with level i of schooling at the test date and level j of final schooling. A $\sim$ means that no observations are available for that cell.

Since we can only identify ratios of the loadings $\lambda(j)$, $j = 1,\ldots,6$, we normalize $\lambda(1) = 1$. Then we have six conditions identifying $\lambda(2)$:

$$\lambda(2) = \frac{a_{21} - a_{2j}}{a_{11} - a_{1j}}, \quad j = 2,3,4,$$

$$\lambda(2) = \frac{a_{22} - a_{2j}}{a_{12} - a_{1j}}, \quad j = 3,4,$$

$$\lambda(2) = \frac{a_{23} - a_{24}}{a_{13} - a_{14}}.$$

We impose these restrictions using a minimum distance framework. Let $Y_2(A)$ be the vector of the six unrestricted estimators. $\lambda(2)$ is estimated by minimizing

$$q(\lambda(2)) = (Y_2(A) - \imath\lambda(2))' W_2 (Y_2(A) - \imath\lambda(2)), \tag{B.2}$$

where W_2 is the inverse of an estimate of the asymptotic covariance matrix of $Y_2(A)$. The minimum is easily found to be

$$\hat{\lambda}(2) = (\imath' W_2 \imath)^{-1} \imath' W_2 Y_2(A). \tag{B.3}$$

For $\lambda(3)$ we have six similar conditions:

$$\lambda(3) = \frac{a_{31} - a_{3j}}{a_{11} - a_{1j}}, \quad j = 2,3,4,$$

$$\lambda(3) = \frac{a_{32} - a_{3j}}{a_{12} - a_{1j}}, \quad j = 3,4,$$

$$\lambda(3) = \frac{a_{33} - a_{34}}{a_{13} - a_{14}}$$

with a similar expression for the minimum distance estimator (but with a different weight matrix W_3).

For $\lambda(4)$ we only have three conditions:

$$\lambda(4) = \frac{a_{42} - a_{4j}}{a_{12} - a_{1j}}, \quad j = 3,4,$$

$$\lambda(4) = \frac{a_{43} - a_{44}}{a_{13} - a_{14}}.$$

Finally, for $\lambda(5)$ there is only one condition:

$$\lambda(5) = \frac{a_{53} - a_{54}}{a_{13} - a_{14}}. \tag{B.4}$$

Here no minimum distance approach is needed. $\lambda(6)$ is not identified.

With all identified loadings estimated we can estimate intercepts $\mu(1),\ldots,\mu(5)$ and control functions $c_1(j) = E[f \mid S = j]$, $j = 1,\ldots,4$; and $c_2(j) = E[f \mid S_T = j]$, $j = 1,\ldots,6$. The model implies the following restrictions:

$$a_{1j} = \mu(1) + \lambda(1)c_1(j), \quad j = 1,2,3,4,$$

$$a_{2j} = \mu(2) + \lambda(2)c_1(j), \quad j = 1,2,3,4,$$

$$a_{3j} = \mu(3) + \lambda(3)c_1(j), \quad j = 1,2,3,4,$$

$$a_{4j} = \mu(4) + \lambda(4)c_1(j), \quad j = 2,3,4,$$

$$a_{5j} = \mu(5) + \lambda(5)c_1(j), \quad j = 3,4,$$

and

$$\frac{\sum_{j=1}^{4} n_{ij} a_{ij}}{\sum_{j=1}^{4} n_{ij}} = \mu(i) + \lambda(i)c_2(i), \quad i = 1,\ldots,5,$$

where n_{ij} is the sample size of cell (i,j). Finally the restrictions

$$\frac{\sum_{j=1}^{4} n_j c_1(j)}{n} = 0 \quad \text{and} \quad \frac{\sum_{j=1}^{6} \tilde{n}_j c_2(j)}{n} = 0$$

are imposed where $n_j = \sum_{i=1}^{6} n_{ij}$ is the count of individuals with final schooling level j and $\tilde{n}_j$ is the count of individuals with $S_T = j$. The second restriction identifies $c_2(6)$ and imposes no restrictions on $c_2(1),\ldots,c_2(5)$. These conditions imply 24 restrictions on the 15 parameters $\theta = (\mu(1),\ldots,\mu(5);\ c_1(1),\ldots,c_1(4);\ c_2(1),\ldots,c_2(6))$. The minimum distance problem is to minimize

$$q(\theta) = (Y(A) - H\theta)' W (Y(A) - H\theta), \tag{B.5}$$

where $Y(A)$ is a linear function of the A elements, H is a known matrix (given estimates of the loadings) and W is the inverse of an estimate of the covariance matrix of $Y(A)$. In forming $Y(A)$, it is convenient to drop the restriction $\frac{1}{n}\sum_{j=1}^{6} \tilde{n}_j c_2(j) = 0$ which identifies $c_2(6)$ and which contributes no further information to reduce the sampling variance of $c_2(1),\ldots,c_2(5)$, or any other parameters beyond $c_2(6)$. Then, in practice, we use 23 restrictions on the 14 parameters $\mu(1),\ldots,\mu(5);\ c_1(1),\ldots,c_1(4);\ c_2(1),\ldots,c_2(5)$ in the minimum distance estimation. The minimum distance estimate of θ is

$$\hat{\theta} = (H'WH)^{-1} H'WY(A). \tag{B.6}$$

Extending to the case controlling for endogenous entry into schooling is straightforward.

B.2. Allowing for endogenous entry

The combinations of schooling at test date and final schooling are represented by the matrices

$$A = \begin{pmatrix} a_{11} & \sim & \sim & \sim \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ \sim & a_{42} & a_{43} & a_{44} \\ \sim & \sim & a_{53} & a_{54} \\ \sim & \sim & \sim & a_{64} \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \\ \sim & b_{42} & b_{43} & b_{44} \\ \sim & \sim & b_{53} & b_{54} \\ \sim & \sim & \sim & b_{64} \end{pmatrix}, \tag{B.7}$$

where a_{ij} represents average test score for individuals with level i of schooling at test date, level j of final schooling and who started school at a normal or early age ($N=0$). b_{ij} represents average test score for individuals with level i of schooling at test date, level j of final schooling and who started school late ($N=1$). A $\sim$ means that no observations are available for that cell.[45]

Since we can only identify ratios of the loadings $\lambda(j)$, $j=1,\ldots,6$, we normalize $\lambda(1)=1$.

[45] Due to the small size of our sample, we do not observe individuals in the cells $i=1$ and $j \geqslant 2$ of matrix A.

We now have seven conditions identifying $\lambda(2)$:

$$\lambda(2) = \frac{a_{21} - b_{21}}{a_{11} - b_{11}},$$

$$\lambda(2) = \frac{b_{21} - b_{2j}}{b_{11} - b_{1j}}, \quad j = 2,3,4,$$

$$\lambda(2) = \frac{b_{22} - b_{2j}}{b_{12} - b_{1j}}, \quad j = 3,4,$$

$$\lambda(2) = \frac{b_{23} - b_{24}}{b_{13} - b_{14}}.$$

As before the restrictions are imposed in a minimum distance framework.

There are seven conditions for $\lambda(3)$:

$$\lambda(3) = \frac{a_{31} - b_{31}}{a_{11} - b_{11}},$$

$$\lambda(3) = \frac{b_{31} - b_{3j}}{b_{11} - b_{1j}}, \quad j = 2,3,4,$$

$$\lambda(3) = \frac{b_{32} - b_{3j}}{b_{12} - b_{1j}}, \quad j = 3,4,$$

$$\lambda(3) = \frac{b_{33} - b_{34}}{b_{13} - b_{14}}.$$

$\lambda(4)$ and $\lambda(5)$ have the same conditions as for case 1 (this is because the cells $a(1,2)$ and $a(1,3)$ are empty).

Note that it is not possible to estimate the ratio $\lambda(6)/\lambda(1)$ as this would require observations in the $a(1,4)$ cell. However we can estimate the ratio $\lambda(6)/\lambda(2)$ from the condition:

$$\frac{\lambda(6)}{\lambda(2)} = \frac{a_{64} - b_{64}}{a_{24} - b_{24}}. \tag{B.8}$$

With this estimate in hand, we can obtain an estimate of the ratio $\lambda(6)/\lambda(1)$ by multiplying with the estimate of $\lambda(2)/\lambda(1)$.

With all loadings estimated we can estimate intercepts $\mu(1),\ldots,\mu(6)$ and control functions $c(1,0),\ldots,c(4,0)$, $c(1,1),\ldots,c(4,1)$. The model implies the following restrictions:

$$a_{1j} = \mu(1) + \lambda(1)c(j,0), \quad j = 1,2,3,4,$$

$$b_{11} = \mu(1) + \lambda(1)c(1,1),$$

$$a_{2j} = \mu(2) + \lambda(2)c(j,0), \quad j = 1,2,3,4,$$

$$b_{2j} = \mu(2) + \lambda(2)c(j,1), \quad j = 1,2,3,4,$$

$$a_{3j} = \mu(3) + \lambda(3)c(j,0), \quad j = 1,2,3,4,$$

$$b_{3j} = \mu(3) + \lambda(3)c(j,1), \quad j = 1,2,3,4,$$

$$a_{4j} = \mu(4) + \lambda(4)c(j,0), \quad j = 2,3,4,$$

$$b_{4j} = \mu(4) + \lambda(4)c(j,1), \quad j = 2,3,4,$$

$$a_{5j} = \mu(5) + \lambda(5)c(j,0), \quad j = 3,4,$$

$$b_{5j} = \mu(5) + \lambda(5)c(j,1), \quad j = 3,4,$$

$$a_{64} = \mu(6) + \lambda(6)c(4,0),$$

$$b_{64} = \mu(6) + \lambda(6)c(4,1),$$

and

$$\sum_{s=1}^{4}\sum_{n=0}^{1} c(s,n)P(s,n) = 0.$$

These conditions imply 34 restrictions on the 14 parameters $\theta = (\mu(1), \ldots, \mu(6); c(1,0), \ldots, c(4,1))$. The minimum distance estimator is used to impose the conditions.

Appendix C. Estimation procedure for structural model

Let $S \in \{1, \ldots, \bar{S}\}$ denote joint choice of completed schooling and age at entry. For clarity we create a special notation in this appendix and let $Z(s)$ be the set of Z variables with nonzero coefficients in the sth choice equation:

$$V(s) = z(s)\gamma(s) + \alpha(s)f + u(s), \quad s = 1, \ldots, \bar{S},$$

where $u(s) \sim \mathrm{N}(0,1)$. We observe $S = \mathrm{argmax}_s\{V(s)\}$.

Let $S_T \in \{1, \ldots, \bar{S}_T\}$ be observed schooling at test date. Let $T^*(S_T)$ be the vector of latent test scores conditional on schooling level s_T, where $T_k^*(S_T)$ denotes the kth test. For $k = 1, \ldots, K$, let

$$T_k^*(S_T) = X(S_T)\beta_k(S_T) + \lambda_k(S_T)f + \varepsilon_k(S_T), \quad S_T = 1, \ldots, \bar{S}_T,$$

where $\varepsilon_k(S_T) \sim \mathrm{N}(0, \sigma_k(S_T)^2)$. We observe

$$T_k = \begin{cases} T_k^*(s_T) & \text{if } T_k^*(s_T) < c_k, \\ c_k & \text{if } T_k^*(s_T) \geqslant c_k \end{cases}$$

if and only if $S_T = s_T$ where c_k is the known maximum value of test k.

Let

$$f \sim \sum_{i=1}^{I} p_i \mathrm{N}(\mu_i, \sigma_i^2).$$

We set $I = 3$.

Let θ denote the vector of model parameters $\{\gamma(s)\}_{s=1}^{\bar{S}}$, $\{\alpha(s)\}_{s=1}^{\bar{S}-1}$, $\{\beta_k(s), \lambda_k(s), \sigma_k(s)\}_{s=1,k=1}^{\bar{S}_T,K}$. We estimate the model parameters via Bayesian Markov Chain Monte

Carlo (MCMC) methods. The goal is to sample from the posterior distribution of the parameters θ and the parameters of the distribution of the factor f, conditional on observed outcomes, S and S_T, and covariates from a random sample of individuals indexed $i = 1, \ldots, n$. The posterior is only a computational device. We are doing maximum likelihood-based inference using MCMC as a computational tool. We impose a noninformative flat prior on all slope coefficients, γ and β. We put proper priors on the variance parameters from the Inverse Gamma family of distributions and on the factor loadings from the Normal distribution family. Under standard regularity conditions, the priors are asymptotically irrelevant.

The data for each individual are test scores, schooling at test date and completed schooling/entry age, $T(S_T)$, S_T, S, plus covariates, A, X, Z, where A is age at test date. The likelihood contribution for one individual is

$$\begin{aligned} & p(T(S_T) = t, S = s, S_T = s_T \mid A = a, X = x, Z = z) \\ & = \int p(t \mid S = s, S_T = s_T, A = a, X = x, Z = z, f) \\ & \quad \times Pr(S_T = s_T \mid S = s, A = a, X = x, Z = z, f) \\ & \quad \times Pr(S = s \mid f,\ Z = z)\, p(f)\, \mathrm{d}f. \end{aligned} \tag{C.1}$$

This likelihood simplifies due to the exact dependence between S, A, S_T described in the text. Associated with each S is a pair (Y, N) where Y is years of completed schooling and N is entry age.

$$\begin{aligned} & \text{If } A - N < Y, \quad S_T = A - N. \\ & \text{If } A - N \geqslant Y, \quad S_T = Y. \end{aligned} \tag{C.2}$$

The likelihood contribution for an individual who has not yet completed schooling is

$$\begin{aligned} & p(T(S_T) = t, S = s, S_T = s_T \mid A = a, X = x, Z = z) \\ & = \int p(t \mid S = s, S_T = s_T, A = a, X = x, f, s_T = a - n) \\ & \quad \times Pr(S = s \mid f,\ X = x,\ Z = z)\, p(f)\, \mathrm{d}f, \end{aligned}$$

while the likelihood contribution for an individual who has completed schooling is

$$\begin{aligned} & p(T(S_T) = t, S = s, S_T = s_T \mid A = a, X = x, Z = z) \\ & = \int p(t \mid S = s, S_T = s_T, A = a, X = x, f, s_T = y) \\ & \quad \times Pr(S = s \mid f,\ X = x,\ Z = z)\, p(f)\, \mathrm{d}f. \end{aligned}$$

The two likelihood contributions are functionally identical—the only difference is what value s_T is conditioned on.

To resolve the high dimensional integrals in these likelihood contributions we augment the likelihood with latent utilities V, determining choice of S, factors f and latent

Table 16
Specification of priors for reported structural model estimates

Parameter	Prior distribution	Prior specification
$\{\gamma(s)\}_{s=1}^{\bar{S}}$	Noninformative flat prior on nonzero coefficients. Degenerate prior with point mass at zero for restricted coefficients (see Table 6)	
$\{\alpha(s)\}_{s=1}^{\bar{S}-1}$	$N(\mu_1, \psi_1^2)$	$\mu_1 = 0,\ \psi_1^2 = 1$
$\{\beta_k(s)\}_{s=1,k=1}^{\bar{S}_T,K}$	Noninformative flat prior	
$\{\lambda_k(s)\}_{s=1,k=1}^{\bar{S}_T,K}$	$N(\mu_k(s), \psi_k(s))$	$\mu_k(s) = 0,\ \psi_k(s) = 1$
$\{\sigma_k(s)^2\}_{s=1,k=1}^{\bar{S}_T,K}$	$IG(a_s, b_s)$	$a_s = 2,\ b_s = 1$

test scores T^*.

$$p(T^*, V, f, S_T = s_T \mid A = a, X = x, Z = z)$$
$$= p(T^* \mid S_T = s_T, A = a, X = x, f)\, p(V \mid f,\ Z = z)\, p(f), \tag{C.3}$$

where s_T is either a or completed schooling depending on the individual having completed schooling or not at test date. Integration of (C.3) with respect to V, T^*, f leads us back to the original likelihood.

The (augmented) sample likelihood is defined as the product of (C.3) over all individuals. We can easily implement a Gibbs sampling algorithm which samples iteratively from the posterior distributions of the parameters and latent data conditional on the observed data. The stationary distribution of the Markov chain generated by this algorithm is the joint posterior distribution of the parameters.

The MCMC algorithm is implemented as follows. Given initial starting values for the parameters and V, f, T^* for $m=1,2,\ldots$ we can update the values of the other parameters and sample from the following conditional distributions (note that we implicitly are conditioning on the data as well as all other parameters). Table 16 summarizes the specifications of the prior distributions for the estimates reported in the paper.

1. The conditional posterior distribution of the latent utilities V is just the product of the individual conditional posterior distributions of V_i by independence. Let S_i be observed final schooling for individual i and let $Z_i(s)$ be all covariates entering schooling alternative s. The elements of V_i are sampled from truncated normals (as in McCulloch and Rossi, 1994),

$$V_i(s) \sim \begin{cases} TN_{[\max_{l \neq s}\{V_i(l)\}, \infty)}(Z_i(s)\gamma(s) + \alpha(s) f_i, 1), & \text{if } s = S_i, \\ TN_{(-\infty, V_i(S_i)]}(Z(s)\gamma(s) + \alpha(s) f_i, 1), & \text{if } s \neq S_i. \end{cases}$$

2. Conditional on $V = \{V_i(s)\}_{i,s}$, the distribution of $\gamma(s)$ follows from a classical linear regression model with noninformative prior.

$$\gamma(s) \sim N(\widehat{\gamma(s)}, \widehat{\Omega(s)}), \quad s = 1, \ldots, \bar{S},$$

where

$$\widehat{\gamma(s)} = (Z(s)'Z(s))^{-1}Z(s)'(V(s) - \alpha(s)f),$$

$$\widehat{\Omega(s)} = (Z(s)'Z(s))^{-1},$$

where $V(s) = (V_1(s), \ldots, V_n(s))$ and $Z(s)' = (Z_1(s), \ldots, Z_n(s))$.

3. Assuming a normal $N(\mu_1, \psi_1^2)$ prior the conditional distribution of $\alpha(s)$ is:

$$\alpha(s) \sim N(\widehat{\alpha(s)}, \widehat{\Omega(s)}), \quad s = 1, \ldots, \bar{S} - 1.$$

where

$$\widehat{\alpha(s)} = \widehat{\Omega_1}\left(f'(V)(s) - Z(s)\gamma(s)) + \frac{\mu_1}{\psi_1^2}\right),$$

$$\widehat{\Omega_1} = \left(f'f + \frac{1}{\psi_1^2}\right)^{-1}.$$

$\alpha(\bar{S})$ is set to zero for identification. Similarly, coefficients for covariates common across alternatives are set to zero for $\gamma(\bar{S})$. We set $\mu_1 = 0$ and $\psi_1^2 = 1$.

4. For each test equation, $k = 1, \ldots, K$, at each schooling level, $s = 1, \ldots, \bar{S}_T$, we estimate the coefficients on the control variables as follows:

$$\beta_k(s) \sim N((X(s)'X(s))^{-1}X(s)'(T_k^*(s) - \lambda_k(s)f), \sigma_k(s)^2(X(s)'X(s))^{-1}),$$

where only those individuals who have completed schooling level s at the test date are included.

5. The factor loadings in the test equations are sampled as

$$\lambda_{k(}s) \sim N(\widehat{\lambda_k(s)}, \widehat{\Omega_k(s)}), \quad k = 1, \ldots, K; \ s = 1, \ldots, \bar{S}_T,$$

where

$$\widehat{\lambda_k(s)} = \widehat{\Omega_k(s)}\left(\frac{f'(T_k^*(s) - X(s)\beta_k(s))}{\sigma_k(s)^2} + \frac{\mu_k(s)}{\psi_k(s)}\right),$$

$$\widehat{\Omega^k(s)} = \left(\frac{f'f}{\sigma^k(s)^2} + \frac{1}{\psi_k(s)}\right)^{-1},$$

using only the individuals who have schooling level s at the test date and using a normal prior $N(\mu_k(s), \psi_k(s))$. We use $\mu_k(s) = 0$ and $\psi_k(s) = 1$.

6. Assuming an Inverse Gamma prior $IG(a_s, b_s)$ and letting $n(s)$ be the number of individuals in schooling group s at the test date, we have:

$$\sigma_k(s)^2 \sim IG\left(\frac{n(s)}{2} + a_s, \frac{e_k(s)'e_k(s)}{2} + b_s\right),$$

where $e_k(s) = T_k^*(s) - X(s)\beta_k(s) - \lambda_k(s)f$. We set $a_s = 2$ and $b_s = 1$.

7. The factors f and the parameters of the factor distribution are sampled as follows. Let $g_i \in \{1, \ldots, I\}$ denote the mixture component from which f_i is sampled. Note that g_i is unobserved. Conditional on g_i the conditional distribution of f_i is easily found to be

$$f_i \sim N(\widehat{f}_i, \widehat{\Gamma}_i),$$

where

$$\widehat{f}_i = \widehat{\Gamma}_i \begin{bmatrix} \sum_{s=1}^{\bar{S}} \alpha(s)(V_i(s) - Z_i(s)\gamma(s)) + \\ \sum_{k=1}^{K} (\lambda_k(s_{Ti})/\sigma_k(s_{Ti})^2)(T_{ik}^* - X_i(s_{Ti}\beta_k(s_{Ti})) + (1/\sigma_{gi}^2)\mu_{g_i} \end{bmatrix},$$

$$\widehat{\Gamma}_i = \left(\sum_{s=1}^{\bar{S}} \alpha(s)^2 + \sum_{k=1}^{K} \lambda_k(s_{Ti})^2/\sigma_k(s_{Ti})^2 + 1/\sigma_{g_i}^2 \right)^{-1},$$

where S_{Ti} is individual i's schooling at test date.

Conditional on f the mixture parameters are sampled by the usual trick of first updating the g_i indicators and then sampling the mixture parameters conditional on the g_i's, cf. Robert and Casella (1999). We impose the restriction $\sum_{i=1}^{I} p_i\mu_i = 0$ using the method in Richardson et al. (2002).

8. The test scores for individuals who hit the ceiling on a test are sampled from truncated normals, i.e.,

$$T_{ik}^* \sim \mathrm{TN}_{(c_k,\infty)}(X_i(s_{Ti})'\beta_k(s_{Ti}) + \lambda_k(s_{Ti})f_i, \sigma_k(s_{Ti})^2).$$

References

Aakvik, A., Heckman, J., Vytlacil, E., 1999. Semiparametric program evaluation lessons from an evaluation of a Norwegian training program. Unpublished manuscript, University of Chicago.

Altonji, J., Pierret, C., 2001. Employer learning and statistical discrimination. Quarterly Journal of Economics 116 (1), 313–350.

Anderson, T.W., Rubin, H., 1956. Statistical inference in factor analysis. In: Neyman, J. (Ed.), Proceedings of Third Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, CA, pp. 111–150.

Ashenfelter, O., Krueger, A., 1994. Estimates of the economic returns to schooling from a new sample of twins. American Economic Review 84 (5), 1157–1173.

Cameron, S., Heckman, J., 1993. The nonequivalence of high school equivalents. Journal of Labor Economics 11 (1), 1–47.

Cameron, S., Heckman, J., 1998. Life cycle schooling and dynamic selection bias: models and evidence for five cohorts of American males. Journal of Political Economy 106 (2), 262–333.

Cameron, S., Heckman, J., 2001. The dynamics of educational attainment for black, hispanic, and white males. Journal of Political Economy 109 (3), 455–499.

Carneiro, P., Hansen, K., Heckman, J., 2003. Estimating distribution of treatment effects counterfactuals with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. International Economic Review 44 (2), 361–422.

Ceci, S.J., 1991. How much does schooling influence general intelligence and its cognitive components?: a reassessment of the evidence. Developmental Psychology 27 (5), 703–722.

Florens, J., Mouchart, M., Rolin, J., 1990. Elements of Bayesian Statistics. Marcel Dekker, New York.

Hanushek, E., 2002. Publicly provided education. In: Auerbach, A., Feldstein, M. (Eds.), Handbook of Public Economics, Vol. 4. North-Holland, Amsterdam, pp. 2045–2141.

Heckman, J., 1974a. Effects of child-care programs on women's work effort. Journal of Political Economy 82 (2, Part II), S136–S163.

Heckman, J., 1974b. Shadow prices, market wages, and labor supply. Econometrica 42 (4), 679–694.

Heckman, J., 1976. Simultaneous Equation Models with both Continuous and Discrete Endogenous Variables With and Without Structural Shift in the Equations. Studies in Nonlinear Estimation. Ballinger, Cambridge, MA.

Heckman, J., 1980. Addendum to sample selection bias as a specification error. In: Stromsdorfer, E., Farkas, G. (Eds.), Evaluation Studies Review Annual, Vol. 5. Sage Publications, Beverley Hills, CA.

Heckman, J., 1981. Statistical models for discrete panel data. In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Data With Econometric Applications. MIT Press, Cambridge.

Heckman, J., Honoré, B., 1990. The empirical content of the Roy model. Econometrica 58 (5), 1121–1149.

Heckman, J., Robb, R., 1985. Alternative methods for evaluating the impact of interventions: an overview. Journal of Econometrics 30 (1–2), 239–267.

Heckman, J., Robb, R., 1986. Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In: Wainer, H. (Ed.), Drawing Inferences from Self-selected Samples. Lawrence Erlbaum, New Jersey, Reprinted 2000.

Heckman, J., Rubinstein, Y., 2001. The importance of noncognitive skills: lessons from the GED testing program. American Economic Review 91 (2), 145–149.

Heckman, J., Vytlacil, E., 2001. Identifying the role of cognitive ability in explaining the level of and change in the return to schooling. Review of Economics and Statistics 83 (1), 1–12.

Herrnstein, R., Murray, C., 1994. The Bell Curve. The Free Press, New York.

Jencks, C., 1972. Inequality: a Reassessment of the Effect of Family and Schooling in America. Basic Books, New York.

Jencks, C., Phillips, M., 1998. America's next achievement test: closing the black–white test score gap. American Prospect, Issue 40 (Sept.–Oct.), 44–53.

Kotlarski, I., 1967. On characterizing the gamma and normal distribution. Pacific Journal of Mathematics 20, 69–76.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. With contributions by Allan Birnbaum. Addison-Wesley, Reading, MA.

Manski, C., 1988. Identification of binary response models. Journal of the American Statistical Association 83 (403, September), 729–738.

Matzkin, R., 1993. Nonparametric identification and estimation of polychotomous choice models. Journal of Econometrics 58 (1-2), 137–168.

McCulloch, R.E., Rossi, P.E., 1994. An exact likelihood analysis of the multinomial probit model. Journal of Econometrics 64 (1-2), 207–240.

McFadden, D., 1984. Econometric analysis of qualitative response models. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, Vol. II. North-Holland, Amsterdam.

Neal, D., Johnson, W., 1996. The role of premarket factors in black–white wage differences. Journal of Political Economy 104 (5), 869–895.

Olley, G., Pakes, A., 1996. The dynamics of productivity in the telecommunications equipment industry. Econometrica 64 (6), 1263–1297.

Richardson, S., Leblond, L., Jaussent, I., Green, P.J., 2002. Mixture models in measurement error problems, with reference to epidemiological studies. Journal of the Royal Statistical Society: Series A (Statistics in Society), 165 (3), 549–566.

Robert, C.P., Casella, G., 1999. Monte Carlo Statistical Methods. Springer, New York.

Thompson, T.S., 1989. Identification of semiparametric discrete choice models. Discussion Paper 249, Department of Economics, University of Minnesota.

Winship, C., 2001. Does going to college make you smarter? Harvard University, unpublished manuscript.

Winship, C., Korenman, S., 1997. Does staying in school make you smarter? The effect of education on IQ in the bell curve. In: Devlin, B., Fienberg, S., Resnick, D., Roeder, K. (Eds.), Intelligence, Genes, and Success: Scientists respond to The Bell Curve. Copernicus Press, New York, pp. 215–234.