# Identification of a Nonseparable Model under Endogeneity using Binary Proxies for Unobserved Heterogeneity

Benjamin Williams*

August 2, 2018

**Abstract**

In this paper I study identification of a nonseparable model with endogeneity arising due to unobserved heterogeneity. Identification relies on the availability of binary proxies that can be used to control for the unobserved heterogeneity. I show that the model is identified in the limit as the number of proxies increases. The argument does not require an instrumental variable that is excluded from the outcome equation nor does it require the support of the unobserved heterogeneity to be finite. I then propose a nonparametric estimator that is consistent as the number of proxies increases with the sample size. I also show that, for a fixed number of proxies, nontrivial bounds on objects of interest can be obtained. Finally, I study two real data applications that illustrate computation of the bounds and estimation with a large number of items.

*George Washington University, Monroe Hall 309, 2115 G Street NW, Washington, DC 20052. bdwilliams@gwu.edu. This work was previously circulated under the title "A Measurement Model with Discrete Measurements and Continuous Latent Variables". It has benefited greatly from discussions with Jim Heckman, Susanne Schennach, Azeem Shaikh, Elie Tamer, and Matthew Wiswall, as well as comments and suggestions by the editor, Stephane Bonhomme, Martin Browning, Steve Durlauf, Florian Gunsilius, Yingyao Hu, Tim Moore, Bob Phillips, Tara Sinclair, Richard Spady, Ed Vytlacil, anonymous referees and seminar participants at the Center for Education Research at UW-Madison, the FDIC, George Washington University, Johns Hopkins University, the University of Chicago, the 2013 Latin American Workshop in Econometrics, and UW-Madison.

# 1 Introduction

This paper considers how multiple binary measurements or proxies can be used to control for latent heterogeneity in a nonseparable model. I assume a model $Y = g(X, \theta, U)$ where $\theta \in \mathbb{R}$ and $U$ are both unobserved, and I assume access to binary proxies of $\theta$, denoted $M_1, \ldots, M_{J+1}$. I study nonparametric identification of an average structural function, $E(g(x, t, U))$ without imposing restrictions on the dependence between $X$ and $\theta$.

This empirical problem arises in many applications in economics. Responses to individual questions on an exam are binary proxies for latent ability. Responses to each item on a personality test or psychological assessment are proxies for a particular latent trait. Responses to items on opinion surveys are proxies for an underlying attitude or belief. Heckman et al. (2006a) and Spady (2007) are typical examples of the use of such data in economics. See Almlund et al. (2011) on the role of the psychology of personality in economics. The binary proxies could also consist of other outcomes that are driven by the same latent variable. For example, in models of legislative roll call voting (Clinton et al., 2004; Heckman and Snyder, 1997; Poole and Rosenthal, 1985, 1997), separate votes are considered binary proxies of latent legislator preferences.

Binary proxies may also arise in measuring economic primitives that vary across economic agents. Bloom and Van Reenen (2007), for example, use discrete responses to survey items to measure the managerial productivity of firms. They aggregate these responses and use this to control for managerial productivity in estimating a production technology. In this context, $X$ represents observed inputs, $\theta$ represents the unobserved managerial productivity, and $U$ represents the residual variation in productivity. Three important features of this model are addressed in this paper. First, managerial productivity varies continuously while the survey items are discrete. Second, managerial productivity is likely correlated with observed inputs if the latter are chosen optimally. Third, responses to the questions on the survey may be affected by observed inputs conditional on managerial productivity.

There are several common approaches to measuring and controlling for latent variables when only binary proxies are available. One approach that is common is to control for the latent variable by conditioning on an average of the proxies or another aggregation of the proxies, such as an estimate from a parametric item response model.[1] This is typically done ad hoc – plugging estimates from one model into another model – and is often not justified theoretically. One contribution of this paper is to provide conditions under which

[1]Item response models are similar to random effects models for binary choice panel data. The binary responses to each item are modeled jointly as a function of the latent variable, item-specific parameters, and idiosyncratic item-specific shocks. These are typically estimated using maximum likelihood or other likelihood-based methods. See van der Linden and Hambleton (2013) or Lord (1980).

this practice can be justified.

More formal approaches involve jointly modeling the economic outcome and the binary proxies, assuming that these are conditionally independent given the latent variable (and observed covariates). In some cases, the latent variable is restricted to have a finite support (Gawade, 2007; Hu, 2008; Mahajan, 2006). This is a restriction on the dependence between the latent variable and the observed covariates. Alternatively, parametric restrictions on the structure of the model can be sufficient to achieve identification without restricting the support of the latent variable. This approach is common in empirical work and is analogous to the correlated random effects model for panel data (see, for example, Junker et al., 2012).

The model studied in this paper does not impose a finite support for the latent variable, any other restrictions on the dependence between the latent variable and observed covariates, or any parametric structure in the model. Carneiro et al. (2003) provide an important identification result for this model. Their result uses exogenous variation in an instrumental variable that is excluded from the outcome equation to identify the distribution of choice-specific outcomes and a large support condition and additive separability to identify the joint distribution of outcomes and the latent variable. This paper provides an alternative identification strategy that does not require an instrument, additive separability, or large support conditions.

The identification problem consists of two parts. The first part deals with the fact that the proxies are binary while the latent variable is continuous. If the proxies are independent of observed covariates conditional on the latent variable and $\theta \sim Uniform(0,1)$ then the percentile of the average of the $J$ proxies converges to $\theta$ as $J \to \infty$. As a result, the model is point identified in the limit as $J \to \infty$ as variation in $\theta$ can be obtained from variation in this percentile score. Thus, the support of $\theta$ can be infinite because this percentile score varies continuously in the limit. The second part of the identification problem arises when observed covariates that are present in the structural outcome equation are also present in the equations for the proxies. As a result, the proxies are not independent of observed covariates conditional on the latent variable and, even when $J$ is large, the percentile of the average of the proxies is no longer a valid estimate of the latent variable. To solve this problem, I assume that one of the binary proxies, $M_{j_0}$, is independent of the observed covariates, $X$, conditional on $\theta$.[2] I show that under this exclusion restriction $E(M_{j_0} \mid \bar{M}_J, X) \approx E(M_{j_0} \mid \theta)$ for large $J$, where $\bar{M}_J = J^{-1} \sum_{j \neq j_0} M_j$. Thus, in the limit, variation in $\theta$ can be obtained by varying the percentiles of this conditional expectation if $E(M_{j_0} \mid \theta = t)$ is a strictly

[2]This is different than the usual exclusion restriction satisfied by an instrumental variable. It also differs from the type of restriction discussed by Carneiro et al. (2003) where a covariate that enters the latent index for one proxy is excluded from the outcome equation and from the latent index for all other proxies.

monotonic function and $\theta \sim Uniform(0,1)$. I also demonstrate how the proxies can be used to construct other estimates of $\theta$ that purge the proxies of $X$ under alternative restrictions on the model.

The exclusion restriction is satisfied in many common applications. For example, suppose the binary proxies are the individual questions on a test of ability. Responses to questions on the test are likely affected by the individual's educational level at the time of the test conditional on ability (Hansen et al., 2004). However, if one question requires only basic knowledge it is plausible that this item does not depend on education at the time of the test conditional on ability, provided that all individuals in the sample have obtained a minimal level of schooling.

I demonstrate the methods developed in this paper through two empirical illustrations. For the first illustration, I use recently released question-level data on the Armed Forces Qualifying Test from the National Longitudinal Survey of Youth (NLSY79). I use the methods developed in this paper to estimate the effect of education on responses to individual questions on the test. In a second empirical application, I revisit an influential paper on the civic returns to education (Dee, 2004). As argued by Dee (2004), schooling is determined in part by individual traits that are potentially correlated with another trait – "civic-mindedness" – that influences later behaviors such as whether the individual votes. Using the methods developed in this paper and data on civic-related behavior, I construct bounds on the effect of education on voting behavior at different points in the distribution of the latent trait.

The remainder of the paper is organized as follows. In Section 2, I lay out a general model and present the main identification and estimation results for large $J$. In Section 3, I present the empirical illustration of the large $J$ methods. In Section 4, I discuss some extensions of the model. In Section 5, I show that bounds on objects of interest can be constructed from moment conditions, study these bounds numerically in a few examples, and present an empirical illustration. Section 6 concludes.

# 2 Large $J$ identification and estimation

In this section, I first outline the general model and discuss the main assumptions. I then state and discuss the main identification result. Finally I propose an estimator and describe its asymptotic properties.

The outcome variable is $Y$ and $X$ denotes a vector of observed covariates. I assume that

$$Y = g(X, \theta, U), \tag{2.1}$$

where $\theta$ and $U$ are both unobserved and $g$ is an unknown function. No restrictions will be placed on the dimension of $U$ but $\theta$ is assumed to be scalar.[3] In addition, I assume the availability of binary proxies, $M = (M_1, \ldots, M_{J+1})$ such that, for each $j = 1, \ldots, J+1$,

$$M_j = \mathbf{1}(h_j(X, \theta) \geq \varepsilon_j). \tag{2.2}$$

where $\varepsilon_j$ is a scalar unobservable and the function $h_j$ is known.

While existing methods, such as Carneiro et al. (2003), impose additive separability in equation (2.1), I show identification without such a restriction. Nonseparability in equation (2.1) is important as it allows ceteris paribus effects of $X$ on $Y$ (e.g., $g(x', \theta, U) - g(x, \theta, U)$) to vary with the unobservables $\theta$ and $U$. The prevalence of unobserved heterogeneity in the effects of choices, actions, or treatments has been widely recognized by economists (e.g., Heckman, 2001; Heckman et al., 2010), as well as in other areas of research (see, for example, Longford, 1999).

The average structural function (ASF), defined by Blundell and Powell (2003) is given by $\int g(x, t, u) dF_{\theta,U}(t, u)$. I define the conditional average structural function (CASF) as the mean outcome averaging only over $U$, that is, $\int g(x, t, u) dF_U(u)$.[4] The CASF describes how the structural function varies with $\theta$, averaging out the other components of the unobserved heterogeneity, $U$. These two structural functions are the main objects of interest throughout this paper.

I maintain the following assumptions. I use the notation "$\perp\!\!\!\perp$" here and throughout the rest of the paper to denote independence. Let $\mathcal{X}$ denote the support of the distribution of $X$ and let $\Theta$ denote the support of the distribution of $\theta$ and, for each $x \in \mathcal{X}$, let $\Theta(x)$ denote the support of the conditional distribution of $\theta \mid X = x$.

**Assumption 2.1.** *$U \perp\!\!\!\perp (X, \theta)$.*

**Assumption 2.2.** *For each $x \in \mathcal{X}$, $\Theta(x) = \Theta$.*

These assumptions are sufficient for identification of both the ASF and the CASF from the distribution of $(Y, X, \theta)$ (cf. Matzkin, 2003, 2004). Indeed, under Assumption 2.1, the CASF is given by $G(x, t) := E(Y \mid X = x, \theta = t)$. Moreover, for any $x \in \mathcal{X}$, under Assumption 2.2, $G(x, t)$ is defined for every $t \in \Theta$. Therefore, the ASF is given by $\int G(x, t) dF_\theta(t)$.[5] Thus,

[3] See Williams (2013) for a version of this model that allows multidimensional $\theta$.

[4] This is similar to the definition of the CASF in Klein (2013).

[5] In many settings, $X = (X_1, X_2)$ where $X_1$ is a scalar regressor of interest and $X_2$ is a vector of "control variables." In this setting, $Y = g(X, \theta, U)$ and Assumption 2.1 would be replaced by the conditional independence assumption, $U \perp\!\!\!\perp (X_1, \theta) \mid X_2$. Then the object of interest would be $E(g(x_1, X_2, t, U))$, which would be identified from $\int E(Y \mid X = x, \theta = t) dF_{X_2}(x_2)$ if $\theta$ were observed. The results in this paper are

the identification problem is reduced to whether the conditional expectation function $G(x,t)$ and the distribution function $F_\theta$ can be identified from the distribution of $(Y,M,X)$.

Assumption 2.1 can be justified by an economic model where $X$ is a choice made under imperfect information that includes $\theta$ but not $U$. In the empirical application in Section 3, $Y$ will denote one of the items on the test while $M$ denotes the remaining items, $X$ denotes years of completed schooling at the time of the test, and $\theta$ denotes the underlying ability measured by the test. In this case, the assumption is satisfied provided that $U$ represents idiosyncratic, item-specific knowledge that does not influence the individual's educational choices. In the application at the end of Section 5, $X$ denotes years of schooling, $Y$ is a measure of voting behavior and $\theta$ denotes "civic-mindedness". The independence assumption can be justified by a model where $U$ denotes factors not determined by the time schooling decisions are made nor dependent on any relevant information that is available at that time.

One advantage of the approach taken in this paper relative to an instrumental variable approach is the ability to identify how structural effects vary with $\theta$. That is, I show identification of the CASF, not just the ASF. While identification of the ASF would require only that $X$ is conditionally independent of $U$ given $\theta$, I maintain Assumption 2.1 in order to show identification of the CASF as well.

Next, I normalize the distribution of $\theta$.

**Assumption 2.3.** $\theta \sim Uniform(0,1)$.

Because there is no observed information on the scale of $\theta$, some normalization in the model is necessary in order to identify $G(x,t)$. Otherwise any monotonic transformation of $\theta$ would define an observationally equivalent model. Alternative normalizations in the model are possible, as discussed in Section 4. This is analogous to the role of location and scale normalizations in traditional parametric models where $\theta$ enters linearly. On the other hand, no such normalization is necessary for identification of the ASF (see Corollary 2.1).

There is a more fundamental problem with identification in this model that may be less apparent and that remains after imposing Assumption 2.3. Namely, it is possible to define an observationally equivalent model based on $\tilde{\theta} = H(X,\theta)$ where $H(x,\cdot)$ is a monotonic transformation for each $x \in X$ and $\tilde{\theta}$ is uniformly distributed on the interval $[0,1]$. While the transformation does not change the marginal distribution, it can be done in such a way as to change the distribution conditional on $X$ dramatically. For example, suppose $H(x,\cdot)$ is the conditional distribution function, $H(x,t) = F_{\theta|X}(t \mid x)$. Then $\tilde{\theta} \mid X = x \sim Uniform(0,1)$. That is, any model satisfying Assumptions 2.1-2.3 with dependence between $\theta$ and $X$ is

still relevant for identification of the conditional expectation function $E(Y \mid X = x, \theta = t)$ in this case. However, I do not explore how additional restrictions on the role of the control variables, $X_2$, might improve identification.

observationally equivalent to another model satisfying these assumptions with $\theta$ independent of $X$.

In order to resolve this problem, I consider models that satisfy an exclusion restriction. Before stating the assumption, I define, for each $j$, the reduced form conditional response functions as $p_j(x,t) := Pr(M_j = 1 \mid X = x, \theta = t) = F_{\varepsilon_j|X,\theta}(h_j(x,t) \mid x,t)$. The following assumption states that one of these conditional response functions is invariant to $x$.

**Assumption 2.4.** *For some $1 \leq j_0 \leq J+1$, for every $x \in \mathcal{X}$, $p_{j_0}(x,t) = p_{j_0}(t)$ for all $t \in [0,1]$.*

This assumption is satisfied if $h_{j_0}(x,t) = h_{j_0}(t)$ and $\varepsilon_{j_0} \perp\!\!\!\perp X \mid \theta$. Under this restriction, one of the $J$ binary proxies does not depend on $X$ conditional on $\theta$. Suppose that $X$ denotes years of schooling at the time a test is administered, that everyone in the sample had attained a minimal level of schooling at the time of the test, and one particular question on the test pertains to knowledge that would have been accumulated before that minimal level of schooling. Then this question would satisfy the exclusion restriction.

Timing is also used to justify the exclusion restriction in the civic returns application. The population studied is a representative sample of high school sophomores in 1980. In the first survey these individuals were all asked a question related to their sense of civic responsibility. Responses to this question do not depend on whether they finished high school or attended college conditional on the underlying "civic-mindedness" trait. The other proxies used are measured later and hence may depend on whether the individual attended college. In Section 4, I show how Assumption 2.4 can be replaced by alternative restrictions.

I next impose the following conditional independence assumption.

**Assumption 2.5.** *$(U, \varepsilon_1, \ldots, \varepsilon_J)$ are mutually independent conditional on $(X, \theta)$.*

This implies that $(Y, M_1, \ldots, M_J)$ are mutually independent conditional on $(X, \theta)$. This assumption is imposed in Carneiro et al. (2003), as well as in many models of measurement error (Chen et al., 2011) and item response models (Sijtsma and Junker, 2006).[6] This assumption is relaxed by Assumption 2.8 below.

Next, I assume two monotonicity conditions.

**Assumption 2.6.**

*(i)* *$p_{j_0}(\cdot)$ is strictly increasing.*

*(ii)* *For each $x \in \mathcal{X}$, $\sum_{j=1}^{J+1} p_j(x,\cdot)$ is strictly increasing.*

[6]Note that the model of equations (2.1) and (2.2) can be viewed as a nonstandard measurement error problem where $\theta$ is the "mismeasured" covariate.

Condition (i) requires a monotonic relationship between $\theta$ and the probability of a positive response on the proxy, $j_0$, that satisfies the exclusion restriction, Assumption 2.4. Under condition (ii) the average of the $J+1$ reduced form conditional response functions is strictly increasing but individual response functions do not have to be strictly increasing. Thus, this assumption allows for limited nonmonotonicity in the response functions. As Sijtsma and Junker (2006) note, this is important in item response data, such as test scores. In the roll call voting example this may be important if, for example, the most liberal and the most conservative members vote together on a small fraction of bills. Condition (ii) is relaxed further by Assumption 2.9 below.

Let $\bar{p}_J(x,t) = J^{-1}\sum_{j\neq j_0} p_j(x,t)$. Under Assumption 2.6, this function is strictly increasing in $t$ for each $x$. Therefore, the inverse function $\bar{p}_J^{-1}(m;x)$ can be defined on the range of $\bar{p}_J(x,\cdot)$. That is, for each $x \in \mathcal{X}$ and each $m$ in the range of $\bar{p}_J(x,\cdot)$ there is a unique $t^*$, which depends on $x$ and $m$, such that $\bar{p}_J(x,t^*) = m$; I denote this $t^*$ by $\bar{p}_J^{-1}(m;x)$. Likewise, under condition (i) of Assumption 2.6 $p_{j_0}^{-1}$ is uniquely defined on the range of $p_{j_0}$. Because $p_{j_0}$ and $\bar{p}_J(x,\cdot)$ are each defined on the interval $[0,1]$, the inverse functions can naturally be extended to be defined on $[0,1]$. For $m < p_{j_0}(0)$, let $p_{j_0}^{-1}(m) = 0$ and for $m > p_{j_0}(1)$, let $p_{j_0}^{-1}(m) = 1$. Likewise, for $m < \bar{p}_J(x,0)$, let $\bar{p}_J^{-1}(m;x) = 0$ and for $m > \bar{p}_J(x,1)$, let $\bar{p}_J^{-1}(m;x) = 1$.

Lastly I impose the following regularity conditions.

**Assumption 2.7.** *There are constants $0 < c < C < \infty$ such that*

*(i)* $|G(x,t') - G(x,t)| \leq C|t'-t|$ *for all* $x \in \mathcal{X}$ *and* $t,t' \in \Theta(x)$,

*(ii)* $c|t'-t| \leq |p_{j_0}(t') - p_{j_0}(t)| \leq C|t'-t|$ *for all* $t,t' \in [0,1]$,

*(iii)* *for each* $J \geq 1$, $c|t'-t| \leq |\bar{p}_J(x,t') - \bar{p}_J(x,t)| \leq C|t'-t|$ *for all* $x \in \mathcal{X}$ *and* $t,t' \in \Theta(x)$,

*(iv)* $c|t'-t| \leq |F_{\theta|X}(t' \mid x) - F_{\theta|X}(t \mid x)|$, *for all* $x \in \mathcal{X}$ *and* $t,t' \in \Theta(x)$, *and*

*(v)* $\mathcal{X}$ *is finite and* $\inf_{x\in\mathcal{X}} Pr(X = x) \geq c$.

Conditions (ii) and (iii) imply that the inverse functions $p_{j_0}^{-1}$ and $\bar{p}_J^{-1}$ are Lipschitz continuous. Condition (iv) implies that the quantile function $Q_{\theta|X}(\cdot \mid x)$ is Lipschitz continuous. This assumption is relaxed in Section 2.4.

## 2.1 Identification

To formally define the notion of identification in the limit that is used, first let $\mathbb{P}_J^0$ denote the true population distribution of $(Y, M_1, \ldots, M_{J+1}) \mid X$. For $\gamma = (g, h_1, \ldots, h_{J+1}, F_{U,\varepsilon_1,\ldots,\varepsilon_{J+1}|X,\theta},$

$F_{\theta|X}$), let $\mathbb{P}_J(\gamma)$ denote the distribution given by

$$\int \mathbf{1}(g(x,t,u) \leq y) \prod_{j=1}^{J+1} \mathbf{1}(h_j(x,t) \geq \varepsilon_j)^{m_j} (h_j(x,t) < \varepsilon_j)^{1-m_j} dF_{U,\varepsilon_1,\ldots,\varepsilon_{J+1}|X,\theta} dF_{\theta|X}. \quad (2.3)$$

Let $\gamma_0 = (g_0, h_{1,0}, \ldots, h_{J+1,0}, F^0_{U,\varepsilon_1,\ldots,\varepsilon_{J+1}|X,\theta}, F^0_{\theta|X})$ denote the true parameter values so that $\mathbb{P}^0_J = \mathbb{P}_J(\gamma_0)$. A parameter value $\gamma$ is observationally equivalent to $\gamma_0$ if $\mathbb{P}_J(\gamma) = \mathbb{P}_J(\gamma_0)$.

Let $\Gamma_J$ denote the parameter space, which is restricted by the assumptions of the model. In particular, there are fixed constants $0 < c < C < \infty$ such that the conditions in Assumption 2.7 hold for all $\gamma \in \Gamma_J$, for all $J \geq 1$. Then, for any feature of the model defined by $\tau = \tau(\gamma)$ define the identified set, $\mathcal{I}_J(\gamma_0; \tau(\cdot))$, as

$$\mathcal{I}_J(\gamma_0; \tau(\cdot)) := \{\tau(\gamma) : \gamma \in \Gamma_J, \mathbb{P}_J(\gamma) = \mathbb{P}_J(\gamma_0)\}. \quad (2.4)$$

In Section 5, I consider computation of $\mathcal{I}_J(\gamma_0; \tau(\cdot))$ through an approximation of the integrals in (2.3). As $J$ grows, the identified set shrinks. Formally, $\tau_0 = \tau(\gamma_0)$ is said to be *large J identified* if

$$\lim_{J\to\infty} \sup_{\gamma_0 \in \Gamma_J} \sup_{\tau \in \mathcal{I}_J(\gamma_0;\tau(\cdot))} ||\tau - \tau_0|| = 0. \quad (2.5)$$

**Theorem 2.1.** *The CASF is large J identified under Assumptions 2.1 and 2.3-2.7.*

*Proof.* See Appendix A. □

Remark 1: *Under Assumption 2.1 the CASF is given by $G(x,t)$, which is a function on $\mathcal{X} \times \Theta$. In applying the definition (2.5) I use the sup norm, $||G - G_0|| = \sup_{x\in\mathcal{X}, t\in\Theta_0(x)} |G(x,t) - G(x,t)|$ where $\Theta_0(x)$ denotes the support of the distribution $F^0_{\theta|X}(\cdot \mid x)$.*

Remark 2: *In the proof of Theorem 2.1, a bound on the rate of convergence of the identified set is also derived. It is shown that the size of the identified set is bounded by $O(J^{-1/2+\epsilon})$ for all $\epsilon > 0$.*

Remark 3: *Note that Assumption 2.6 rules out a model with*

$$M_j = \mathbf{1}(\theta > c_j(X)) \quad (2.6)$$

*for each j because in that case $\bar{p}_J(x,t)$ is piecewise constant in t for each x. However, it can be shown that, under certain conditions on the thresholds $\{c_j(\cdot)\}$, the CASF is still large J identified (Williams, 2012). Essentially what is required is that $c_j(x)$ varies enough with j for each x.*

A fundamental idea behind Theorem 2.1 has been used in the nonparametric item response literature (Junker and Ellis, 1997) and has roots in earlier work in statistics (de Finetti, 1931; Diaconis and Freedman, 1980). The idea is that $\bar{M}_J := J^{-1} \sum_{j \neq j_0} M_j$ can serve as a sort of sufficient statistic for the latent heterogeneity. Douglas (2001) used this idea to formally prove nonparametric identification of the standard item response model.[7]

In Lemma A.1 in the appendix, I use Hoeffding's inequality to show that, under Assumption 2.5, $\bar{M}_J - \bar{p}_J(X, \theta) \to_p 0$ as $J \to \infty$. To provide some intuition for the identification result in Theorem 2.1, consider the limiting case where $\bar{M}_\infty = \bar{p}_\infty(X, \theta)$. Then, if the function $\bar{p}_\infty(x, \cdot)$ is invertible for each $x$,

$$\begin{aligned} T_\infty(m, x) &:= Pr(M_{j_0} = 1 \mid \bar{M}_\infty = m, X = x) \\ &= p_{j_0}(x, \bar{p}_\infty^{-1}(m; x)) \\ &= p_{j_0}(\bar{p}_\infty^{-1}(m; x)), \end{aligned} \tag{2.7}$$

where the final line follows from Assumption 2.4. Then this implies that $T_\infty(\bar{M}_\infty, X) = p_{j_0}(\theta)$ and

$$\begin{aligned} F_{T_\infty}(t) &:= Pr(T_\infty(\bar{M}, X) \leq t) \\ &= Pr(p_{j_0}(\theta) \leq t) \\ &= p_{j_0}^{-1}(t), \end{aligned} \tag{2.8}$$

where the final line follows from Assumptions 2.3 and 2.6(i). Combining these results, $F_{T_\infty}(T_\infty(\bar{M}_\infty, X)) = p_{j_0}^{-1}(p_{j_0}(\theta)) = \theta$. Thus, the joint distribution of $(Y, X', \theta)$ is pinned down by the joint distribution of $(Y, X', F_{T_\infty}(T_\infty(\bar{M}_\infty, X)))$. The CASF is then identified in this limiting case because, as argued after the statement of Assumptions 2.1 and 2.2 above, the CASF is given by $G(x, t) = E(Y \mid X, \theta)$.

The proof of Theorem 2.1 is more subtle than this for several reasons. First, $\bar{M}_\infty$ cannot simply be replaced by $\bar{p}_\infty(X, \theta)$; instead we must combine various limiting results. And, in order to do so, these limiting results must be uniform rather than pointwise. Second, the proof is based on $T_J(m, x) := Pr(M_{j_0} = 1 \mid |\bar{M}_J - m| < r_J, X = x)$ for a sequence $r_J \to 0$, rather than $Pr(M_{j_0} = 1 \mid \bar{M}_J = m, X = x)$, because the probability that $\bar{M}_J = m$ cannot be sufficiently bounded away from 0 as $J \to \infty$. Third, while $F_{T_\infty}(t)$ is a smooth function of $t$, the distribution function of $T_J(\bar{M}_J, X)$ is not smooth since $\bar{M}_J$ is a discrete random variable for a fixed $J$.

[7] Douglas (2001) formalizes an idea used in the psychometrics literature (Douglas, 1997; Ramsay, 1991) to nonparametrically estimate item response functions. This result has not previously received attention in the econometrics literature.

Nevertheless, this heuristic argument demonstrates the importance of the exclusion restriction and the monotonicity of $p_{j_0}$. Because $T_\infty(\bar{M}_\infty, X) = p_{j_0}(\theta)$, which does not depend on $X$, individuals can be ordered based on this "score", rather than on $\bar{M}_\infty$. In Section 4, I show how alternative restrictions can be used to derive different "score" functions.

This argument also suggests two corollaries to Theorem 2.1. First, under the common support condition, Assumption 2.2, the ASF, which is given by $\int G(x,t)dF_\theta(t)$, is identified without Assumption 2.3. That is, to identify the average (across the distribution of $\theta$) structural function it is not necessary to normalize the distribution of $\theta$.

Second, if the exclusion restriction (Assumption 2.4) holds for $\bar{p}_J$ rather than any particular $p_j$ then the CASF is large $J$ identified through a simpler argument. In this case, if $\bar{M}_\infty = \bar{p}_\infty(X,\theta)$ and if $F_{\bar{M}_\infty} := Pr(\bar{M}_\infty \leq m)$ then $F_{\bar{M}_\infty}(\bar{M}_\infty) = \theta$. The following two corollaries are proved in Appendix B in the supplementary material.

**Corollary 2.1.** *Suppose that (i) there is a constant $\bar{Y} < \infty$ such that $|Y| \leq \bar{Y}$ and (ii) $\theta$ has a distribution function $F_\theta$ such that $|F_\theta(t') - F_\theta(t)| \leq C|t' - t|$ for all $t, t' \in \mathbb{R}$. Then, under Assumptions 2.1, 2.2, 2.4 and 2.5-2.7, the ASF is large $J$ identified.*

**Corollary 2.2.** *If $\bar{p}_J(x,t) = \bar{p}_J(t)$ for all $x \in \mathcal{X}$ and $t \in [0,1]$ then the CASF is large $J$ identified under Assumptions 2.1, 2.3 and 2.5-2.7.*

## 2.2 Estimation

Consider an *i.i.d.* sample $(Y_i, X_i, M_{i,1}, \ldots, M_{i,J+1}), i = 1, \ldots, n$ from the model of equations (2.1) and (2.2). In this section I propose an estimator for the conditional average structural function, $CASF(x,t)$, that is consistent as $n, J \to \infty$ provided that $J$ is on the order of a power of $n$. The estimation strategy is to estimate $\theta_i$ for each $i = 1, \ldots, n$ in a first stage and to use these estimates, $\hat{\theta}_1, \ldots, \hat{\theta}_n$, in place of $\theta_1, \ldots, \theta_n$ in a second stage.

First, recall the intuition behind Theorem 2.1. Let $\bar{M}_{iJ} = J^{-1}\sum_{j \neq j_0} M_{i,j}$. As $J \to \infty$, the conditional probability function, $Pr(M_{i,j_0} = 1 \mid \bar{M}_{iJ} = m, X_i = x)$ converges to the function $q(x,m) := p_{j_0}(\bar{p}_J^{-1}(x,m))$. Further, $q(X_i, \bar{M}_{iJ}) \approx p_{j_0}(\theta_i)$ for large $J$ and, since $\theta_i \sim Uniform(0,1)$, $F_{q(X,\bar{M})}(q(X_i, \bar{M}_{iJ})) \approx \theta_i$ where $F_{q(X,\bar{M})}$ denotes the distribution function of the random variable $q(X_i, \bar{M}_{iJ})$. This argument suggests the estimator

$$\hat{\theta}_i = \hat{F}_{\hat{q}(X,\bar{M})}(\hat{q}(X_i, \bar{M}_{iJ})) \tag{2.9}$$

where $\hat{q}(x,m)$ is the following Nadaraya-Watson kernel estimator of $Pr(M_{i,j_0} = 1 \mid \bar{M}_{iJ} =$

$m, X_i = x)$

$$\hat{q}(x,m) = \frac{\sum_{i=1}^n M_{i,j_0} K_{h_1}(\bar{M}_{iJ} - m, X_i - x)}{\sum_{i=1}^n K_{h_1}(\bar{M}_{iJ} - m, X_i - x)}, \tag{2.10}$$

where $K_h(u,x) = h^{-1}K(h^{-1}u)\mathbf{1}(X_i = x)$ for a kernel function $K(\cdot)$ and bandwidth $h$, and $\hat{F}_{\hat{q}(X,\bar{M})}$ is the empirical distribution function

$$\hat{F}_{\hat{q}(X,\bar{M})}(p) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(\hat{q}(X_i, \bar{M}_{iJ}) \leq p). \tag{2.11}$$

My proposed estimator for $G(x,t)$ is

$$\hat{G}(x,t) = \frac{\sum_{i=1}^n Y_i L_{h_2}(\hat{\theta}_i - t, X_i - x)}{\sum_{i=1}^n L_{h_2}(\hat{\theta}_i - t, X_i - x)} \tag{2.12}$$

where $L_h(u,x) = h^{-1}L(h^{-1}u)\mathbf{1}(X_i = x)$ for a kernel function $L(\cdot)$ and bandwidth $h$.

To demonstrate the type of estimation results that can be obtained for the model, I derive a bound on the convergence rate of the estimator $\hat{h}(x,t)$. New results due to Mammen et al. (2012) on nonparametric estimation with regressors generated in a first stage suggest that the convergence rates derived here could be improved under certain smoothness conditions. However, because the conditions in that paper cannot be directly applied in the model of this paper, and since the primary focus of this paper is identification, I leave this to future research.

**Theorem 2.2.** *Under the assumptions maintained in Theorem 2.1 and Assumptions C.1-C.6 stated in Appendix C in the supplementary material,*

$$|\hat{G}(x,t) - G(x,t)| = O_p\left(\frac{1}{\sqrt{nh_{2n}}} + h_{2n}\right).$$

*If $J < \kappa n^{-2/3}$ for some $\kappa > 0$ then the bandwidths can be chosen so that $\hat{G}$ converges at a rate no slower than $n^{-1/3+\epsilon}$ for any $\epsilon > 0$.*

*Proof.* See Appendix C in the supplementary material. □

## 2.3 Monte Carlo

To demonstrate the performance of the proposed estimator I carried out a Monte Carlo exercise. The simulations were based on the model $Y_i = 0.5X_i + 0.5\tilde{\theta}_i + U_i$ where $U_i \sim N(0, \sigma_U^2)$, $\sigma_U = 0.1$. The observed covariate $X$ is binary with $Pr(X_i = 1) = 0.5$ and

Table 1. Monte Carlo simulations

| n | J | no controls | | score | | infeasible | | proposed method | |
|---|---|---|---|---|---|---|---|---|---|
| | | Bias | RMSE | Bias | RMSE | Bias | RMSE | Bias | RMSE |
| 1000 | 10 | 0.500 | 0.501 | 0.303 | 0.303 | 0.000 | 0.013 | 0.136 | 0.142 |
| | 30 | 0.502 | 0.504 | 0.280 | 0.280 | 0.001 | 0.013 | 0.082 | 0.096 |
| | 100 | 0.502 | 0.503 | 0.268 | 0.268 | 0.001 | 0.013 | 0.061 | 0.079 |
| 2000 | 10 | 0.499 | 0.499 | 0.295 | 0.296 | 0.000 | 0.009 | 0.120 | 0.123 |
| | 30 | 0.501 | 0.501 | 0.273 | 0.273 | 0.001 | 0.008 | 0.076 | 0.083 |
| | 100 | 0.498 | 0.498 | 0.260 | 0.260 | 0.002 | 0.010 | 0.058 | 0.067 |
| 5000 | 10 | 0.503 | 0.503 | 0.295 | 0.295 | 0.000 | 0.007 | 0.088 | 0.092 |
| | 30 | 0.500 | 0.500 | 0.273 | 0.273 | 0.001 | 0.006 | 0.049 | 0.053 |
| | 100 | 0.500 | 0.501 | 0.259 | 0.260 | 0.000 | 0.007 | 0.024 | 0.034 |

Notes: These results were obtained by simulating the model described in Section 3.3 100 times for each pair of n and J. The first column is the difference in sample means. The second column was obtained by conditioning nonparametrically on the percentile of the average of the proxies. The third was obtained by conditioning nonparametrically on the true latent variables. The fourth estimator is the estimator proposed in Section 3.2. All kernel regressions used the Epanechnikov kernel.

$\tilde{\theta}_i \mid X_i = x \sim N(x - .5, 1)$. The proxies are generated according to $M_1 = \mathbf{1}(\tilde{\theta}_i \geq \eta_{i1})$ and $M_j = \mathbf{1}(-0.5X + \tilde{\theta}_i \geq \eta_{ij})$ for $j > 1$ with $\eta_{ij} \sim^{iid} N(0, 1)$. This fits the model of Section 5 with $\theta_i = F_{\tilde{\theta}}(\tilde{\theta}_i)$.

In the simulations I calculate the estimator proposed above in Section 2.2, $\hat{G}(x, t)$, for $x = 0, 1$ and $t \in \mathcal{T} = \{.05, .1, \ldots, .95\}$. Since $\theta \sim Uniform(0, 1)$, $\frac{1}{10}\sum_{t\in\mathcal{T}} \hat{G}(1, t) - \hat{G}(0, t)$ provides an approximation of the average treatment effect, $ATE = \int G(1, t)dF_\theta(t) - \int G(0, t)dF_\theta(t)$. Further refining the grid did not change the overall results.

Table 1 reports the results of the simulations. I provide results from three other estimators for comparison. For the first column the $ATE$ was estimated, without controlling in any way for $\theta$, simply as $\hat{E}(Y_i \mid X_i = 1) - \hat{E}(Y_i \mid X_i = 0)$. For the second column I estimated a nonparametric kernel regression of $Y_i$ on the percentiles of $\bar{M}_{iJ}$. I computed these estimates on the grid $\mathcal{T}$ and averaged to get an estimate of the $ATE$. The third column shows results from the infeasible estimator that uses $\theta_i$ directly.

Overall the results suggest a substantial improvement over methods that do not properly control for $\theta_i$, even when $J = 10$. However, there is a non-negligible bias when $J$ is small. The simulation exercises also demonstrate that increasing $J$ leads to a bigger improvement in the MSE when $n$ is larger. And increasing $n$ leads to a bigger improvement in the MSE when $J$ is larger.

## 2.4 Large $J$ identification under weaker conditions

The CASF is large $J$ identified under conditions weaker than Assumptions 2.5-2.7, although the convergence is at a slower rate and the proof is considerably more complex. First, conditional independence is too strong in some applications. Consider instead the following

weak dependence assumptions.

**Assumption 2.8.**

*There exists a decreasing sequence* $\{\alpha_k : k \geq 1\}$ *with* $\lim_{k\to\infty} \alpha_k = 0$ *such that*

*(i)* $E\left(|E(M_j \mid X, \theta, \{M_s : |j-s| > k\}) - E(M_j \mid X, \theta)|\right) \leq \alpha_k$ *and*

*(ii) for any* $\eta \in (0, 1/2)$*, there exist* $\mathcal{J}_Y^J(\eta) \subset \{1, \ldots, J\}$ *for each* $J$ *such that* $|\mathcal{J}_Y^J(\eta)|^{-1} = O(J^{-1})$ *and*

$$E\left(|E(Y \mid X, \theta, \{M_j : j \in \mathcal{J}_Y^J(\eta)\}) - E(Y \mid X, \theta)|\right) \leq \alpha_{\lfloor \eta J \rfloor}.$$

Condition (i) is a mixing condition on the sequence $M_1, \ldots, M_J$ conditional on $(X, \theta)$. Mixing conditions are a standard way to model (unconditional) dependence in time series data.[8] Thus, this is a natural notion of dependence in a setting where the $M_j$ are realized consecutively. For example, if $M_j$ represents the response to the $j^{th}$ item on a test there may be dependence between consecutive questions due to factors other than the individual's ability level, such as learning from the test.

Condition (ii) allows for various forms of dependence between $Y$ and some of the binary proxies conditional on $(X, \theta)$. If $Y$ is independent of only a subset of the proxies conditional on $(X, \theta)$ this condition requires only that this subset grows with $J$. Alternatively it allows for $Y$ to be dependent on all of the proxies provided that the dependence is weak in a specific sense. It allows for the case, for example, where $Y$ is itself one of the proxies. In that case $\mathcal{J}_Y^J(\eta)$ is $\{s : |j-s| > \eta J\}$ so that $|\mathcal{J}_Y^J(\lfloor \eta J \rfloor)| \geq (1-2\eta)J$. Performance on a test may call on basic skills, represented by $\theta$, as well as various specific pieces of knowledge. Condition (i) specifies the sense in which these specific pieces of knowledge cannot dominate the test. Condition (ii) allows some of these individual factors to influence the outcome (wages, for example).

Both conditions in Assumption 2.5 could also be replaced by lower level conditions related to the structure in equations (2.1) and (2.2). For example, if $(X, \theta)$ is independent of $(\varepsilon_1, \ldots, \varepsilon_J)$ then condition (i) could be replaced by a mixing condition on $\{\varepsilon_j : 1 \leq j \leq J\}$. And if $U$ is independent of $(X, \theta)$ conditional on $(\varepsilon_1, \ldots, \varepsilon_J)$ then condition (ii) can be stated in terms of weak dependence between $U$ and $(\varepsilon_1, \ldots, \varepsilon_J)$. See de Jong and Woutersen (2011) for related results in a dynamic time series binary choice model.

Next, consider the following monotonicity conditions in place of Assumption 2.6.

[8] If the time series process $\{q_t\}$ is strongly mixing with mixing coefficients $\alpha_{t,k}$ then it can be shown that $E|E(q_t \mid \{q_s; |t-s| > k\}) - E(q_t)| \leq \alpha_{t,k}$ where $\sup_t \alpha_{t,k} \to 0$ (Dvoretzky et al., 1972; McLeish et al., 1975). Processes with $m$-dependence and ARMA processes are examples of processes that are strongly mixing.

**Assumption 2.9.**

*(i)* $p_{j_0}(\cdot)$ *is strictly increasing on* $[0,1]$.

*(ii) There exists a constant* $\eta > 0$*, and subsets* $\mathcal{J}_m^J \subset \mathcal{J}_Y^J(\eta) \cap \{1 \leq j \leq J : |j_0 - j| > \eta J\}$ *for each* $J$*, such that* $|\mathcal{J}_m^J|^{-1} = O(J^{-1})$ *and, for each* $x \in \mathcal{X}$*,* $\sum_{j \in \mathcal{J}_m^J} p_j(x, \cdot)$ *is strictly increasing.*

Condition (i) is the same as condition (i) of Assumption 2.6. Condition (ii) states that, once items near $j_0$, items not indexed by $j$ in $\mathcal{J}_Y^J(\eta)$, and a limited number of other items are excluded, the average of the remaining items is a strictly increasing function, and that the number of remaining items is proportional to $J$. Let $N_J := |\mathcal{J}_m^J|$ and redefine $\bar{p}_J(x,t) := N_J^{-1} \sum_{j \in \mathcal{J}_m^J} p_j(x,t)$. Under condition (ii) of Assumption 2.6, this function is strictly increasing in $t$ for each $x$. Therefore, the inverse function $\bar{p}_J^{-1}(m;x)$ can be defined on $[0,1]$ as before.

Lastly I impose the following regularity conditions, which weaker Lipschitz continuity to continuity.

**Assumption 2.10.** *(i)* $G(x,t)$ *is continuous in* $t$ *for each* $x \in \mathcal{X}$*, (ii)* $p_{j_0}$ *and* $p_{j_0}^{-1}$ *are both continuous, (iii)* $\bar{p}_J(x,t)$ *is continuous in* $t$ *for each* $J$ *and each* $x \in \mathcal{X}$ *and* $\bar{p}_J^{-1}(m;x)$ *is continuous in* $m$ *for each* $J$ *and each* $x \in \mathcal{X}$*, (iv) the quantile function,* $Q_{\theta|X}(\tau \mid x)$*, is defined for all* $\tau \in [0,1]$ *and is uniformly continuous in* $\tau$ *for each* $x \in \mathcal{X}$*, and (v)* $\mathcal{X}$ *is finite and* $\inf_{x \in \mathcal{X}} Pr(X = x) \geq c$.

A slightly stronger version of these regularity conditions, stated as Assumption B.1 in Appendix B in the supplementary material, is needed to control the continuity of functions in the identified set as $J \to \infty$. Assumption 2.10 by itself does not prevent the limiting identified set from containing discontinuous functions, which would prevent identification in the limit. Assumption B.1 additionally requires $\Gamma_J$ to be constructed from uniformly equicontinuous families of functions, just as it was assumed for Theorem 2.1 that for each $\gamma \in \Gamma_J$ the relevant functions were Lipschitz continuous with the same Lipschitz constants.

**Theorem 2.3.** *The CASF is large* $J$ *identified under Assumptions 2.1, 2.3-2.4, 2.8-2.10, and B.1.*

*Proof.* See Appendix B in the supplementary material. □

Remark 4: *The rate of convergence of the identified set is slower than the* $O(J^{-1/2+\epsilon})$ *attained under the assumptions of Theorem 2.1. The convergence rate depends on the rate of convergence of the mixing coefficients,* $\alpha_k$*, and the smoothness of the functions in the parameter space* $\Gamma_J$.

# 3 Education, ability, and test scores

There is substantial evidence that education can improve performance on tests of cognitive ability (see, e.g., Neal and Johnson, 1996; Winship and Korenman, 1997). Education, however, is endogenous if higher ability individuals achieve higher education levels on average. Hansen et al. (2004) propose two methods for dealing with this endogeneity, both of which are derived from a model that allows the mapping between latent ability and test scores to depend on the schooling level at the time of the test. One involves jointly modeling education and test scores in a parametric model and is closely related to Carneiro et al. (2003). The other is a semiparametric control function method that relies on the assumption that schooling at the time the test is taken is independent of latent ability conditional on the final education level obtained.

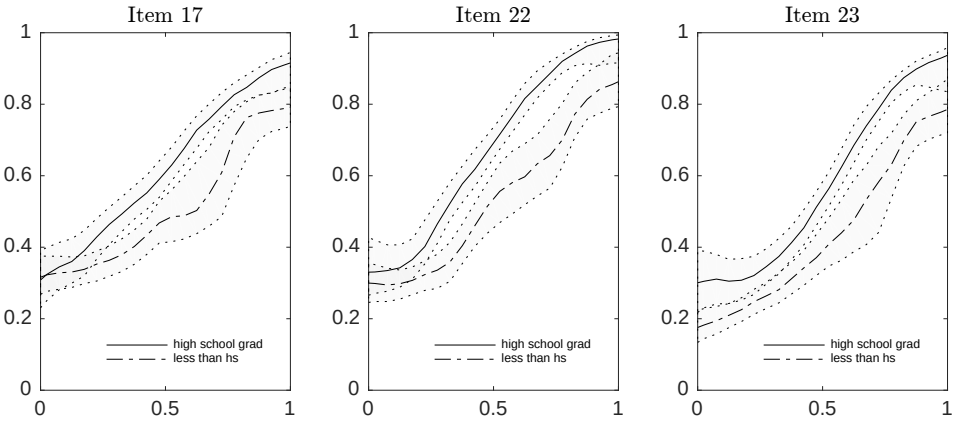
In this section, I use data from the National Longitudinal Survey of Youth (NLSY) to study the effect of education on test scores using the methods described in Section 2. The NLSY is a representative sample of individuals from the United States between the ages of 14 and 21 in 1979, when they were first interviewed. In 1980, these individuals were administered the Armed Services Vocational Aptitude Battery (ASVAB). As an illustration, I use the arithmetic reasoning subcomponent of the Armed Forces Qualifying Test (AFQT), a test which consists of the verbal and math components of the ASVAB. The arithmetic reasoning component of the ASVAB consists of 30 questions. Item-level responses, coded as correct or incorrect, have recently been made publicly available. See Schofield (2014) for an early analysis of the item-level data. I use the same subsample used in Hansen et al. (2004), which consists of $1,927$ white non-Hispanic males. Hansen et al. (2004) find that each additional year of education increases composite AFQT scores by $2-4\%$.

I study the effect of education on the test score by analyzing each item separately. Let $X_i$ denote a binary indicator of schooling level and let $\tilde{M}_i$ denote the full vector of 30 items from the arithmetic reasoning component of the ASVAB. For each item $s$, I apply the previous analysis in the paper with $Y_i = \tilde{M}_{is}$ and $M_i = \tilde{M}_{i,-s} = (\tilde{M}_{i1}, \ldots, \tilde{M}_{i,s-1}, \tilde{M}_{i,s+1}, \ldots, \tilde{M}_{i,30})$. So that the model is consistent as $s$ varies, I assume throughout the analysis that $\tilde{M}_{ij} = \mathbf{1}(h_j(X_i, \theta_i) \geq \tilde{\varepsilon}_{ij})$ and $\tilde{\varepsilon}_{ij} \perp\!\!\!\perp (X_i, \theta_i)$ for each $j = 1, \ldots, 30$. Then the CASF for each item $s$ is given by $G_s(x,t) = F_{\tilde{\varepsilon}_s}(h_s(x,t)) = p_s(x,t)$.[9]

The method described in Section 2 involves first using the individual items, $M_i$, to estimate $\hat{\theta}_i$ for each individual and then estimating $G_s(x,t)$ by a nonparametric regression of the outcome $Y_i$ on $\hat{\theta}_i$ and $X_i$. I use this method to compute estimates, $\widehat{G}_s(x,t)$, for each

[9]Imposing the assumption that $\tilde{\varepsilon}_{ij} \perp\!\!\!\perp (X_i, \theta_i)$ for all $j$ does not aid in identification of $G_s$ because it does not restrict the item response function $p_j(x,t)$ or the conditional dependence among the items given $(X_i, \theta_i)$.

Figure 1: The effect of schooling on individual items from the AR component of the ASVAB




Notes: The ASF estimates in the three panels are computed as described in the text treating responses to the $28^{th}$, $29^{th}$, and $30^{th}$ items on the Arithmetic Reasoning component of the ASVAB as the outcome. The Epanechnikov kernel was used in both steps. The shaded region is a 90% confidence interval computed using 200 bootstrap samples. Estimates are based on a sample of size $1,927$ from the NLSY79.

$1 \leq s \leq 30$ except $s = j_0$. In computing $\widehat{G}_s(x,t)$, I include the 28 items excluding item $s$ and item $j_0$ in $\mathcal{J}_m^J$, which means that the estimates of $\theta_i$ actually differ for each $s$.[10] To estimate the schooling effect on each item, I estimate a conditional average treatment effect as $\widehat{CATE}_s(t) = \widehat{G}_s(1,t) - \widehat{G}_s(0,t)$ for each $s \neq j_0$. By assumption $p_{j_0}(1,t) = p_{j_0}(0,t)$ for all $t$. Therefore,

$$\widehat{CATE}(t) = \frac{1}{J} \sum_{s \neq j_0} \widehat{CATE}_s(t)$$

provides an estimate of the conditional average causal effect of schooling on the test score.[11]

Figures 1 and 2 show $\widehat{G}_s(0,t)$ and $\widehat{G}_s(1,t)$ for a selection of the test items, along with 90% confidence bands computed via 200 bootstrap samples.[12] For these results I use $j_0 = 4$.[13]

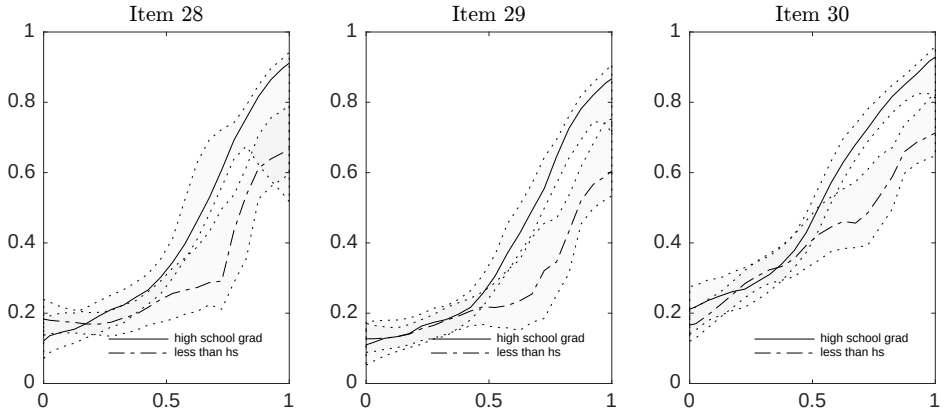
[10]The theoretical analysis in the paper requires that the outcome variable, $Y_i$, is not included in constructing the mean response $\bar{M}_{iJ}$. However, results using a single set of estimates $\hat{\theta}_i$ based on all 29 items excluding $j_0$ do not differ substantially from the reported results.

[11]"Test score" refers to the simple average, $\frac{1}{J}\sum_{s=1}^{30} \tilde{M}_{is}$. The estimator can be easily modified if the test score is a weighted average of the individual items.

[12]The validity of the bootstrap-based confidence intervals for kernel regression estimators has been established by Hall (1992), among others. These standard results do not apply immediately here because (a) the estimation procedure involves the use of a regressor generated in a first step and (b) the estimator is consistent only if $J \to \infty$ along with $n$. While formally establishing the validity of the bootstrap-based confidence intervals for the estimator $\widehat{G}_s(x,t)$ is beyond the scope of this paper, others have addressed these two issues separately. See Kapetanios (2008) regarding the validity of the (cross sectional) bootstrap in large $n$, large $T$ panel data models and see Mammen et al. (2016) for a result regarding the validity of the bootstrap for semiparametric estimators involving a generated regressor.

[13]See online appendix D where I describe a data-driven approach to this choice.

Figure 2: The effect of schooling on individual items from the AR component of the ASVAB

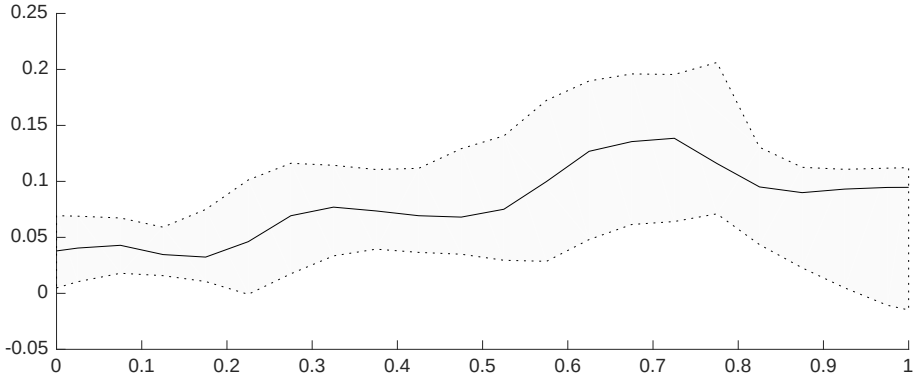



Notes: The ASF estimates in the three panels are computed as described in the text treating responses to the $28^{th}$, $29^{th}$, and $30^{th}$ items on the Arithmetic Reasoning component of the ASVAB as the outcome. The Epanechnikov kernel was used in both steps. The shaded region is a 90% confidence interval computed using 200 bootstrap samples. Estimates are based on a sample of size $1,927$ from the NLSY79.

Each of these items shows a clear effect of schooling on the probability of correctly answering the question. There is heterogeneity in this effect across items, however. For example, for the three items in Figure 1, there is a statistically significant effect at all ability levels. For the items in Figure 2, there is a large and significant effect only for individuals above the median ability level. Figure 3 plots $\widehat{CATE}(t)$. Here we see that the aggregate effect of schooling on the test score is increasing in ability and is statistically significant at all ability levels.

Since $\theta_i \sim Uniform(0,1)$, I can also estimate item-level average treatment effects as $\widehat{ATE}_s = \sum_{h=1}^{H} \widehat{CATE}_s(t_h)$ and the average treatment effect on the test score as $\widehat{ATE} = \sum_{h=1}^{H} \widehat{CATE}(t_h)$, where $\{t_h\}_{h=1}^{H}$ is an equidistant grid of 20 points. Table 2 reports estimates of these average treatment effects. As in Hansen et al. (2004), I find that there is a substantial education effect on test scores. There is a statistically significant average effect for all but four of the 29 items (excluding item 4). Having a high school education (or more) at the time of the test increases the test score by between 4.6 and 10 percentage points, on average. This effect is roughly $7-15\%$ of the average total score of 0.659. This is consistent with the findings of Hansen et al. (2004), though the effects found here are larger, potentially because I do not control for demographics and family background beyond restricting the sample to white males.

Figure 3: The effect of schooling on the AR component of the ASVAB




Notes: The ATE estimate is computed as described in the text. The Epanechnikov kernel was used in both steps. The shaded region is a 90% confidence interval computed using 200 bootstrap samples. Estimates are based on a sample of size $1,927$ from the NLSY79.

Table 2. Average treatment effect of schooling on the Arithmetic Reasoning component of the AFQT in the NLSY79

| item | ATE | 90% conf. int. | item | ATE | 90% conf. int. |
|---|---|---|---|---|---|
| 1 | 0.011 | [-0.008, 0.031] | 17 | 0.109 | [0.068, 0.155] |
| 2 | 0.029 | [0.008, 0.049] | 18 | 0.114 | [0.039, 0.164] |
| 3 | 0.032 | [-0.004, 0.058] | 19 | 0.069 | [0.025, 0.113] |
| 5 | 0.045 | [-0.002, 0.078] | 20 | 0.122 | [0.061, 0.172] |
| 6 | 0.042 | [0.001, 0.087] | 21 | 0.106 | [0.034, 0.151] |
| 7 | 0.027 | [-0.008, 0.061] | 22 | 0.141 | [0.076, 0.177] |
| 8 | 0.062 | [0.021, 0.104] | 23 | 0.148 | [0.095, 0.2] |
| 9 | 0.072 | [0.029, 0.109] | 24 | 0.142 | [0.064, 0.205] |
| 10 | 0.057 | [0.028, 0.097] | 25 | 0.085 | [0.025, 0.142] |
| 11 | 0.046 | [0.005, 0.086] | 26 | 0.071 | [0.009, 0.112] |
| 12 | 0.054 | [0.005, 0.095] | 27 | 0.119 | [0.05, 0.165] |
| 13 | 0.108 | [0.053, 0.16] | 28 | 0.114 | [0.046, 0.164] |
| 14 | 0.067 | [0.015, 0.119] | 29 | 0.113 | [0.034, 0.138] |
| 15 | 0.078 | [0.04, 0.119] | 30 | 0.106 | [0.054, 0.156] |
| 16 | 0.057 | [0.005, 0.105] | **avg.** | **0.081** | **[0.046, 0.104]** |

Notes: This table reports estimates of the the average treatment effect (ATE) of high school graduation on the probability of a correct response to each item. For each item, ATE was estimated as described in the text. The 90% confidence intervals were computed from simulating 200 bootstrap samples. The estimates are based on a sample of 1,927 white males from the NLSY. See the text for a further description of the sample.

# 4 Extensions of the model

The model of Section 2 can be extended in several ways. In this section I discuss two types of extensions. First, I introduce an alternative normalization that can be used to identify and estimation the model instead of normalizing $\theta$ to be $Uniform(0,1)$. Second I discuss alternatives to the exclusion restriction in Assumption 2.4. Williams (2013) considers some additional extensions of the model.

## 4.1 Alternative normalizations

Suppose the distribution of $\theta$ is not normalized, as imposed by Assumption 2.3 and instead assume that, for some $x_0 \in \mathcal{X}$ and some $j_1 \in \{1, \ldots, J+1\}$, $p_{j_1}(x_0, t) = \pi(t)$ where $\pi$ is a known function. Recall from the discussion following the statement of Theorem 2.1 that for large $J$,

$$E(Y \mid \bar{M}_J \approx m, X = x) \approx G(x, \bar{p}_J^{-1}(m; x)) \tag{4.1}$$

and

$$Pr(M_{j_0} = 1 \mid \bar{M}_J \approx m, X = x) \approx p_{j_0}(\bar{p}_J^{-1}(m; x)). \tag{4.2}$$

Likewise, under this alternative normalization,

$$Pr(M_{j_1} = 1 \mid \bar{M}_J \approx m, X = x) \approx \pi(x_0, \bar{p}_J^{-1}(m; x_0)). \tag{4.3}$$

Since $\pi$ is known this implies that $\bar{p}_J^{-1}(m; x_0)$ is approximately identified for large $J$. Then, applying (4.2) for $x = x_0$, $p_{j_0}(t)$ is identified as well. Applying (4.2) again for any other value of $x$ produces $\bar{p}_J^{-1}(m; x)$. And applying (4.1), $G(x, t)$ can then be obtained. While this is merely a heuristic argument the result can be proved formally under sufficient regularity conditions just as the discussion following the statement of Theorem 2.1 was formalized in the proof of the theorem.

Suppose $\theta$ represents ability and each $M_j$ is an item on a test. This shows how one item on the test can be used to set the scale of latent ability $\theta$. Then the distribution of $\theta$ can be identified and estimated rather than normalized, and, if the item satisfying the normalization is chosen carefully, then this can provide a more easily interpretable model. Alternatively, $M_{j_1}$ might represent a binary outcome rather than an item on the test, or a similar restriction on the function $g$, rather than on $p_{j_1}$, could be used to set the scale of $\theta$. In a model of the technology of skill formation, Cunha et al. (2010) emphasize the importance

of anchoring test scores in an interpretable metric in this way.

## 4.2 Alternative restrictions

Next I consider two restrictions that can be used in place of Assumption 2.4.

### 4.2.1 Conditional independence in the measurement

Suppose that $X_1$ is a subvector of $X$ such that $p_{j_0}$ varies only with $X_1$ and $\theta \perp\!\!\!\perp X_1 \mid X_{-1}$ where $X_{-1}$ denotes the components of the vector $X$ excluding the components of $X_1$.

Under this restriction,

$$Pr(M_{j_0} = 1 \mid \bar{M}_J \approx m, X = x) \approx p_{j_0}(x_1, \bar{p}_J^{-1}(m; x)) \tag{4.4}$$

and $p_{j_0}(X_1, \bar{p}_J^{-1}(\bar{M}_J; X)) \approx p_{j_0}(X_1, \theta)$. Then

$$\begin{aligned} Pr(p_{j_0}(X_1, \bar{p}_J^{-1}(\bar{M}_J; X)) \leq \pi \mid X = x) &\approx Pr(p_{j_0}(X_1, \theta) \leq \pi \mid X = x) \\ &= F_{\theta|X_{-1}}(p_{j_0}^{-1}(\pi; x_1) \mid x_{-1}). \end{aligned} \tag{4.5}$$

Then, since $\theta \sim Uniform(0, 1)$, averaging this over the distribution of $X_{-1}$ produces $p_{j_0}^{-1}(\pi; x_1)$. This implies that $\bar{p}_J^{-1}(m; x)$ is identified and hence the CASF is identified.

Hansen et al. (2004) use this type of normalization to estimate the effect of education on performance on a standardized test. This was extended to a model of the effect of education on economic and social outcomes as an adult ($Y$) by Heckman et al. (2006a). In these models, the score on a standardized test depends on the individual's education level at the time the test was taken ($X_1$). However, the individual's ability ($\theta$) is also correlated with $X_1$ because $X_1$ is dependent on the individual's final level of education, $X_2$, which is influenced by ability. So, for example, an individual with $X_1 = 12$ must have $X_2 \geq 12$ and therefore will have a higher $\theta$ on average than someone with $X_1 = 10$. If the problem of retention is ignored, conditional on $X_2$, $X_1$ is a deterministic function of the student's age at the time of the test. Because the age at which the test was administered is exogenous, $\theta \perp\!\!\!\perp X_1 \mid X_2$, and the identification strategy just described can be applied if $X$ contains both $X_1$ and $X_2$.

### 4.2.2 Linking exclusion restrictions

According to Assumption 2.4 one item, $M_{j_0}$, must be independent of $X$ conditional on $\theta$. In some cases however, each item may be dependent on *some* components of $X$ conditional on

$\theta$. In this case it is sufficient that, for each of the $K$ components of $X$, there is one item which is independent of that component conditional on $\theta$.

Here I provide a sketch of the argument for $X = (X_1, X_2)$. Suppose that $X_2$ is excluded from $p_{j_1}$ and $X_1$ is excluded from $p_{j_2}$. First,

$$Pr(M_{j_1} = 1 \mid \bar{M}_J \approx m, X = x) \approx p_{j_1}(x_1, \bar{p}_J^{-1}(m; x)) \tag{4.6}$$

and

$$Pr(M_{j_2} = 1 \mid \bar{M}_J \approx m, X = x) \approx p_{j_2}(x_2, \bar{p}_J^{-1}(m; x)). \tag{4.7}$$

From the right hand side of these two equations one can obtain $p_{j_1}(x_1, p_{j_2}^{-1}(\pi; x_2))$. Furthermore,

$$Pr(p_{j_1}(X_1, \bar{p}_J^{-1}(\bar{M}_J; X)) \leq \pi \mid X = x) \approx F_{\theta|X}(p_{j_1}^{-1}(\pi; x_1) \mid x).$$

Averaging this over the distribution of $X_2 \mid X_1 = x_1$ produces $F_{\theta|X}(p_{j_1}^{-1}(\pi; x_1) \mid x_1)$. Then, plugging in $p_{j_1}(x_1, p_{j_2}^{-1}(\pi; x_2))$ and averaging over the marginal distribution of $X_1$,

$$\int F_{\theta|X}(p_{j_2}^{-1}(\pi; x_2) \mid x_1) dF_{X_1}(x_1) = p_{j_2}^{-1}(\pi; x_2)$$

since $\theta \sim Uniform(0, 1)$. With $p_{j_2}^{-1}(\pi; x_2)$ identified, $\bar{p}_J^{-1}(m; x)$ and the rest of the model can be determined.

This argument can be extended to the case where every component of $X$ is excluded from the equation for at least one item. Suppose, for example, that $\theta$ represents a risk aversion parameter and the items $M_1, \ldots, M_J$ represent participation in different risky behaviors in a population of young adults. In estimating a causal effect of education on risky behaviors it is important to control for this latent risk aversion parameter, in addition to parents' income. The strategy discussed in this section can be used to identify such a model if at least one of the risky behaviors is not affected by education (perhaps because the risks involved are readily apparent) and at least one of the risky behaviors is not affected by parents' income (perhaps because there is no monetary cost of participation). The exclusion restriction is much weaker than the exclusions required by Carneiro et al. (2003).

# 5 The identified set when $J$ is small

In this section, I show that, even with a small number of proxies, this model has identifying power under the conditional independence of Assumption 2.5. I maintain equations

(2.1) and (2.2) and Assumptions 2.1-2.5 but I modify Assumptions 2.6 and 2.7.

Assumption 2.1 and 2.5 are sufficient to derive moment conditions that can be used to define the identified set for $G$. This is formalized in the following theorem, which is proved in Appendix A.

**Theorem 5.1.** *Under Assumption 2.1 the CASF is given by* $G(x,t) = E(Y \mid X = x, \theta = t)$ *and under Assumption 2.5, for any* $\mathcal{J} \subseteq \{1, \ldots, J+1\}$ *and any* $c \in \{0,1\}$,

$$E(Y^c \prod_{j \in \mathcal{J}} M_j \mid X = x) = \int G(x,t)^c p_{\mathcal{J}}(x,t) dF_{\theta|X}(t \mid x) \tag{5.1}$$

*where* $p_{\mathcal{J}} := \prod_{j \in \mathcal{J}} p_j$.

The identified set for the CASF is then the set of functions $G$ that are consistent with these $2^{J+1} - 1$ moment conditions. See supplementary appendix D for a careful definition and a description of how this set can be approximated.[14] In these examples I focus on identification of two scalar objects – the CATE at a fixed $t$, i.e., $G(x',t) - G(x,t)$, and the ATE, i.e., $\int_0^1 (G(x',t) - G(x,t))\, dt$. For the former, if not also for the latter, it is clear that some sort of shape restriction is essential.[15] Therefore, I consider models that are monotonic in the latent variable. A separate monotonicity assumption is used in the large $J$ analysis (see Assumptions 2.6 and 2.9 above). Here I assume the following.

**Assumption 5.1.** *For each* $x \in \mathcal{X}$, *each of the functions* $G(x,\cdot), p_1(x,\cdot), \ldots, p_{J+1}(x,\cdot)$ *is weakly increasing on the support of* $\theta \mid X = x$.

If it is known *a priori* that some of the proxies are positively related to the latent variable while others are negatively related then the latter can be redefined so that the assumption is still satisfied as stated. However, this assumption does rule out the scenario where the correct orientation is neither known *a priori* nor prescribed by an economic model. While it is possible that, given only an assumption of monotonicity in unknown direction, the correct orientation is identified in the model, I do not pursue this here.

Assumption 5.1 is still not sufficient for bounds to be nontrivial in all cases because it allows for the extreme case where each $p_j(x,\cdot)$ is a constant function. Indeed, in this case it

[14] I use an approximation method that extends methods implemented in Honore and Tamer (2006) and Chernozhukov et al. (2013) for a semiparametric panel data model.

[15] Consider the following example. For each $j$ let $p_j(x,t) = \frac{1}{2}\sum_{k=0}^{d} a_{jk}(x)\psi_k(t) + \frac{1}{2}$ where $\{\psi_k(\cdot)\}_{k=0}^{\infty}$ are the shifted Legendre polynomials defined on $[0,1]$ and $0 \leq \max_{x \in \mathcal{X}} \sum_{k=0}^{d} a_{jk}(x) < 1$. In addition suppose that $G(x,t) = \sum_{k=0}^{d} a_k^C(x)\psi_k(t)$. Since $p_{\mathcal{J}}(x,t)$ can then be written as a linear combination of the first $Jd$ shifted Legendre polynomials, an observationally equivalent model can be defined by taking $G^*(x,t) := G(x,t) + \lambda(x)e_j(t)$ where $e_j(t) = \psi_{k^*}(t)$ for any $k > Jd$. But $|G(x,t) - G^*(x,t)| \geq |\lambda(x)|$. This works because of the orthogonality of the Legendre polynomials, which *requires* nonmonotonicity.

can be shown that the identified set is the trivial set. However, computation of the bounds is not tractable if we instead require the functions to be strictly increasing because this would entail optimization over a non-compact parameter space.

I now illustrate features of the identified set by way of several examples. The details of the computations presented here are also contained in supplementary appendix D. In each case, I calculate the population moments according to the following data generating process.

$$\begin{aligned} X &= \mathbf{1}(s\theta \geq V), \\ Y &= \mathbf{1}(\beta X + \mu_Y + \alpha_Y \Phi^{-1}(\theta) + \sqrt{1-\alpha_Y^2} U \geq 0), \text{ and} \\ M_j &= \mathbf{1}(\beta_j X + \mu_j + \alpha_j \Phi^{-1}(\theta) + \sqrt{1-\alpha_j^2} \varepsilon_j \geq 0), \quad j = 1, \ldots, J \end{aligned} \tag{5.2}$$

where $V, U, \varepsilon_1, \ldots, \varepsilon_J$ are mutually independent, $V \sim Uniform(0,1)$ and $U, \varepsilon_1, \ldots, \varepsilon_J$ are each drawn from the standard normal distribution.

Figures 4 and 5 show bounds on the ATE, $\int_0^1 (G(1,t) - G(0,t))\, dt$.[16] For each example, I compute the trivial bounds, the bounds based on observing $(Y, M_1, X)$ and bounds based on $(Y, M_1, M_2, X)$. Several features of the identified set that are common to these examples are apparent. First, the bounds are generally nontrivial even when only one binary proxy is observed. Second, if a second binary proxy is observed, the bounds typically become more narrow. The relative benefit of this second proxy depends on the model though. Third, when two binary proxies are observed, the bounds tend to be the most narrow when $\alpha_Y = 0.01$.[17]

## 5.1 An empirical example

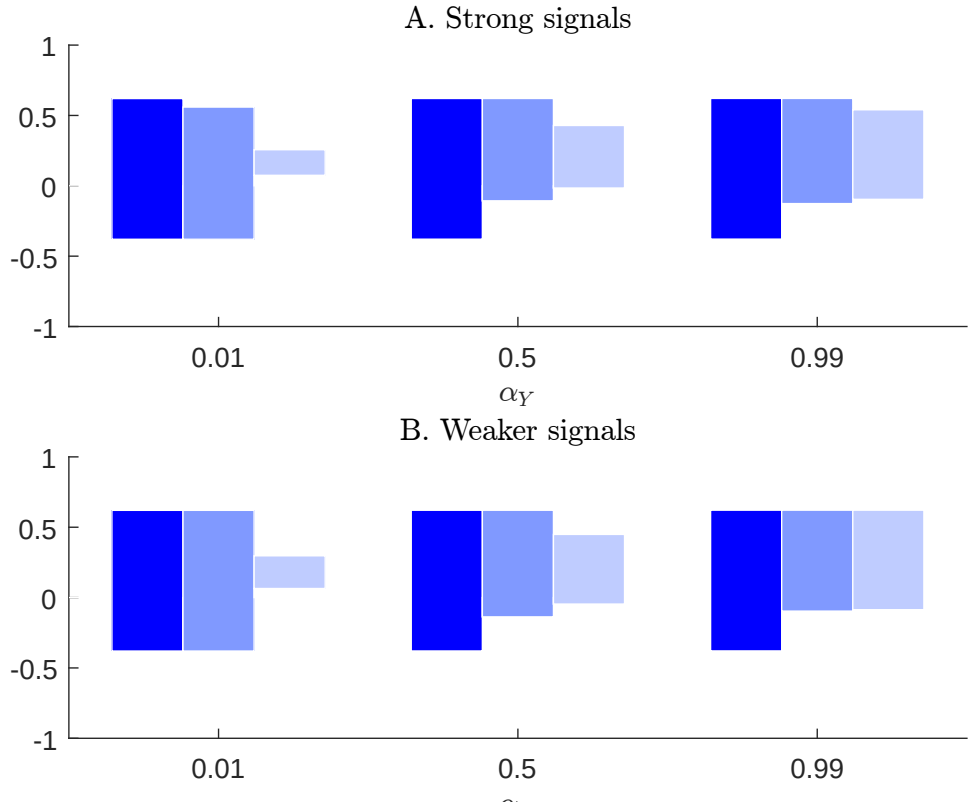
Dee (2004) provides evidence that college attendance substantially increases civic-related behavior. OLS estimates using data from the High School and Beyond (HSB) longitudinal study indicate that attending college by the age of 20 increases the probability of being registered to vote at age 28 by roughly 12 percentage points, for example. Dee (2004) also provides instrumental variables estimates that suggest a larger effect, an increase of roughly 22 percentage points. Identification is based on variation in college availability, which is assumed to be exogenous.

The instrumental variable analysis in Dee (2004) is motivated in part by the observation

[16] Rather than imposing the common support condition (Assumption 2.2), I instead impose that $G(x,t)$ is bounded between 0 and 1 to obtain nontrivial bounds on the ASF. The boundedness condition translates to a bound on the size of the effect outside of the common support. Neither condition is needed to obtain nontrivial bounds on $G(x,t)$ for $t \in \Theta(x)$.

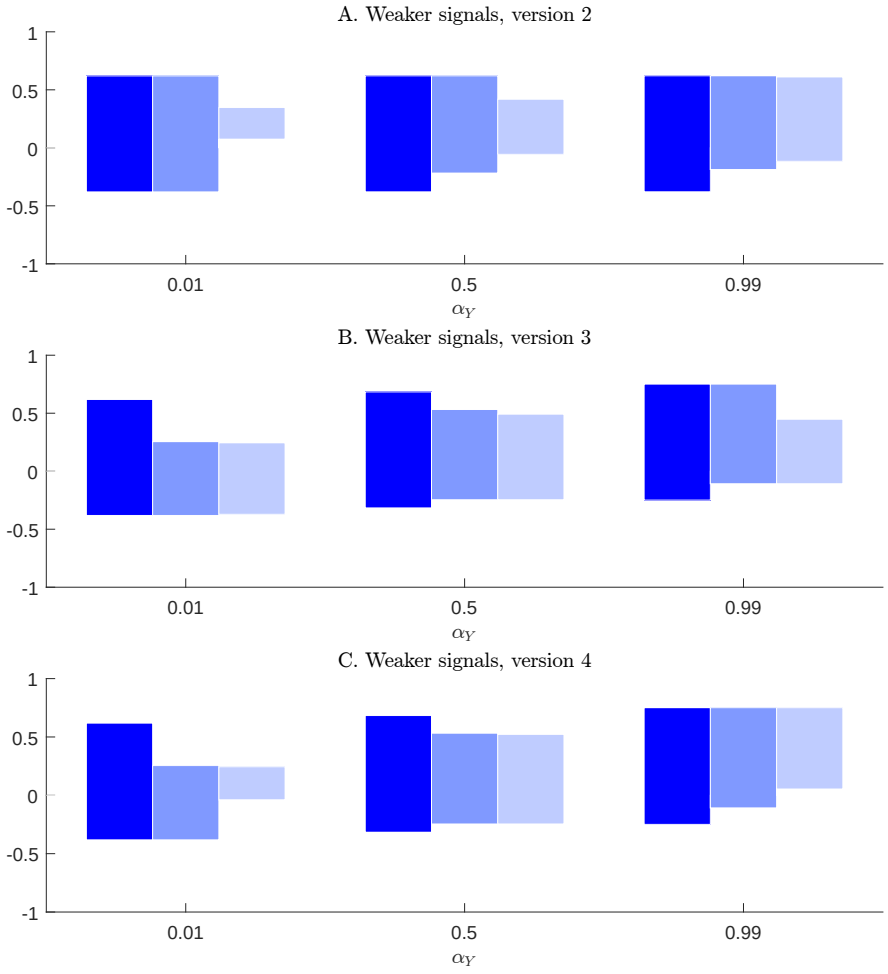
[17] In this case, $Y$ is nearly independent of $M_j$ conditional on $X$ and, hence, it can be inferred that $Y$ varies little, if at all, with $\theta$. From this it can further be inferred that dependence between $Y$ and $X$ must be largely due to the structural relationship.

Figure 4: Bounds on the ATE, examples 1 and 2




Notes: Each bar represents bounds on the ATE. In each grouping, the first bar represents the trivial bounds, the second represents bounds based on only $M_1$, and the third bar represents bounds based on $M_1$ and $M_2$. In the model used in panel A, $\beta = 1$, $s = 0$, $\beta_j = 0$, $\alpha_1 = \alpha_2 = 0.99$, and $\mu_Y, \mu_1, \mu_2$ are set so that $E(M_1) = 0.5$, $E(M_2) = 0.9$, and $ASF(0) = 0.7$. The model used in panel B is the same except that $\alpha_1 = \alpha_2 = 0.5$. In all models, $j_0 = 1$. I used the approximation method described in Appendix C with a uniformly spaced partition of $S = 40$ points and set $\epsilon_S = 10^{-6}$. The ATE in the data generating process is 0.24.

Figure 5: Bounds on the ATE, examples 3-5




Notes: Each bar represents bounds on the ATE. In each grouping, the first bar represents the trivial bounds, the second represents bounds based on only $M_1$, and the third bar represents bounds based on $M_1$ and $M_2$. In the model used in panel A, $\beta = 1$, $s = 0$, $\beta_j = 0$, $\alpha_1 = \alpha_2 = 0.5$, and $\mu_Y, \mu_1, \mu_2$ are set so that $E(M_1) = 0.9$, $E(M_2) = 0.5$, and $ASF(0) = 0.7$. The model used in panel B is the same except that $s = 1$ so that $E(\theta \mid X = 1) - E(\theta \mid X = 0) = 0.33$. The model used in panel C is the same except that $s = 1$ and $\beta_2 = 1$. In all models, $j_0 = 1$. I used the approximation method described in Appendix C with a uniformly spaced partition of $S = 40$ points and set $\epsilon_S = 10^{-6}$. The ATE in the data generating process is 0.24.

that regressions using civic-related behaviors that preceded the college attendance decision as the dependent variable produce positive and significant college attendance effects. This is what we would expect if civic-related behaviors are driven by a latent "civic-mindedness" trait that is formed in high school. If there is heterogeneity in the civic returns to education then the IV estimates in Dee (2004) are estimates of the effect for those who would be induced to attend college by a reduction in the distance to nearby colleges as IV estimates are a weighted average of marginal treatment effects (Heckman et al., 2006b).

I analyze similar data from the same High School and Beyond (HSB) longitudinal study. The data consists of a sample of high school sophomores in 1980. Individuals in this sample who reported having attended college by 1984 (when the majority were 20 years old) were 23 percentage points more likely to have voted in an election between March of 1984 and February of 1986 than those who did not report having attended college. I consider the nonseparable model of equation (2.1),

$$Voted_i = g(SomeCollege_i, \theta_i, U_i) \tag{5.3}$$

where $\theta_i$ is the individual's latent "civic-mindedness". In addition I use data from the HSB on other civic-related behaviors. Specifically I find three proxies that are appropriate to measure $\theta_i$. The first proxy ($M_{i1}$) is whether the individual answered that correcting social and economic inequalities was very important, as opposed to somewhat or not important, in the baseline survey in 1980. The second proxy ($M_{i2}$) indicates whether the individual participated in service organizations, political clubs, neighborhood groups, or other volunteer work in the 1986 follow-up. The third proxy ($M_{i3}$) indicates whether the individual reported at least sometimes discussing public problems in the country or their own community with others in the 1986 follow-up. Identification relies on the exclusion restriction (Assumption 2.4) which requires that college attendance does not enter the equation for the first proxy. This assumption is satisfied because $M_{i1}$ was measured when the entire sample was still enrolled in high school.

I consider three models that use the first proxy only, the first two proxies only, and all three proxies. For comparison, the first column in Table 3 reports the results of a regression of $Voted_i$ on $SomeCollege_i$ that also controls for each of these sets of the three proxies. Controlling for these measures of civic behaviors and attitudes reduces the coefficient slightly, from 0.23 to 0.2. The second column reports estimated bounds on the ATE. I estimate the bounds using the method describe in supplementary appendix D.

Overall, while the bounds narrow as more proxies are included in the model, the bounds are quite wide. The bounds on the ATE do not exclude 0 in any of the three models and the

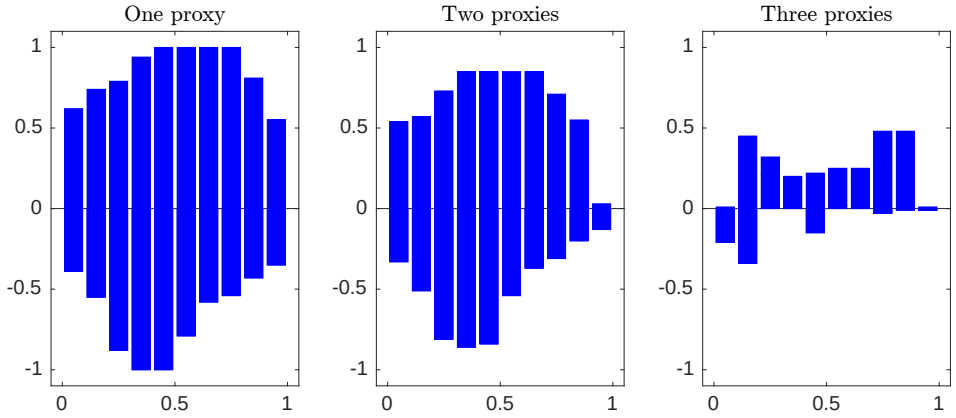
Table 3. Civic returns to education

| model | OLS | ATE bounds | |
|---|---|---|---|
| no proxies | 0.23 | | |
| M1 only | 0.23 | -0.34 | 0.56 |
| M1 and M2 | 0.21 | -0.26 | 0.22 |
| M1, M2, and M3 | 0.20 | -0.06 | 0.21 |

Notes: The first column in this table reports the coefficient on SomeCollege in an OLS regression that uses the proxies as controls. The outcome is an indicator for whether the individual has voted in and election in the past two years. The sample size is 10,515.

width of the bounds narrows from 0.9 to 0.27. The OLS estimates are at the upper end of the identified set for models 2 and 3. Figure 6 shows pointwise bounds on the CATE. These bounds are not informative for model 1 and narrow only slightly for model 2. The bounds for the CATE in model 3 are substantially more narrow. The lower bound for the CATE at points above 0.5 is near 0.

Figure 6: Bounds on the conditional ATE of education on the probability of voting




Notes: The bounds are computed as described in Section 4.1. Estimates are based on a sample of size 10,515 from the HSB longitudinal survey.

# 6 Conclusion

This paper introduces new results that demonstrate how binary proxies can be used to obtain identification in a nonseparable model with endogeneity. It provides an approach

that assumes neither exogeneity conditional on a vector of observed covariates nor requires an instrument that is excluded from the outcome equation. Nor does this approach require any covariates with large support. The model has identifying power, in the sense that the identified set is nontrivial, with even a few binary proxies. However, the empirical results in Section 5.1 suggest that the identifying power in the model can be weak. This suggests that, in these cases, identification in the standard parametric models is primarily imposed by the parametric structure. The more positive result coming from this paper is that the model is identified in the limit so that it can be estimated consistently with a large number of proxies. The paper also shows how the model can be nonparametrically estimated as $n, J \to \infty$.

These results also suggest an alternative use of high-dimensional data in the context of an economic model with heterogeneity to current work (see Belloni et al., 2013, for a different approach). In a setting where big data can be quickly and inexpensively generated, the identification conditions provide a roadmap for how to produce data that will facilitate identification.

# Appendix

## A Identification Proofs

This appendix contains proofs of the main identification results. A separate supplementary appendix provides the remaining proofs and results on computation of the identified set.

### A.1 Preliminary results and a sketch of the proof of Theorem 2.1

**Lemma A.1.** *Under Assumption 2.5, for any $\epsilon > 0$,*

$$Pr\left(|\bar{M}_J - \bar{p}_J(X,\theta)| > \epsilon\right) \leq 2\exp\left(-2J\epsilon^2\right).$$

*Proof.* First,

$$\begin{aligned} &Pr(|\bar{M}_J - \bar{p}_J(x,\theta)| > \epsilon \mid X = x, \theta = t) \\ &= Pr(|\bar{M}_J - E(\bar{M}_J \mid X = x, \theta = t)| > \epsilon \mid X = x, \theta = t) \\ &\leq 2\exp\left(-2J\epsilon^2\right) \end{aligned} \tag{A.1}$$

where the equality follows from the definition of $\bar{p}_J$ and the inequality follows from Hoeffding's inequality since $\bar{M}_J = J^{-1}\sum_{j \neq j_0} M_j$ where the $M_j$ are independent random variables conditional on $(X, \theta)$ (by Assumption 2.5) and are bounded between 0 and 1.

Second, by the law of iterated expectations

$$\begin{aligned} Pr\left(|\bar{M}_J - \bar{p}_J(X,\theta)| > \epsilon\right) &= E\left(Pr(|\bar{M}_J - \bar{p}_J(x,\theta)| > \epsilon \mid X, \theta)\right) \\ &\leq E\left(2\exp\left(-2J\epsilon^2\right)\right) \\ &= 2\exp\left(-2J\epsilon^2\right). \end{aligned} \tag{A.2}$$

□

**Lemma A.2.** *Suppose that $A$ and $B$ are two random variables such that $Pr(|A-B| > x) \leq y$. Suppose the distribution function for $B$ is Lipschitz continuous with Lipschitz constant $\bar{f}_B$ (i.e., $|Pr(B \leq x') - Pr(B \leq x)| \leq \bar{f}_B|x' - x|$). Then for any $z, z'$ such that $|z' - z| \leq w$,*

$$|Pr(A \leq z) - Pr(B \leq z')| \leq y + (x + w)\bar{f}_B.$$

*Proof.* See Appendix B in the supplementary material. □

By Lemma A.1, $\bar{M}_J - \bar{p}_J(X,\theta) \to_p 0$ as $J \to \infty$. Consider any $x \in \mathcal{X}$ and $t \in \Theta(x)$. In Lemma A.4 below I also show that

$$|m - \bar{p}_J(x,t)| \leq r_J/2 \Longrightarrow \lim_{J\to\infty} E(Y \mid |\bar{M}_J - m| < r_J, X = x) - G(x,t) = 0 \tag{A.3}$$

for a sequence $r_J \to 0$. A similar result is shown in Douglas (2001) though the observed covariates, $X$, are not present in that paper.

Likewise, in the proof of Lemma A.3 below it is shown that

$$|m - \bar{p}_J(x,t)| \leq r_J/2 \Longrightarrow \lim_{J\to\infty} T_J(m,x) - p_{j_0}(t) = 0 \tag{A.4}$$

where $T_J(m,x) := Pr(M_{j_0} = 1 \mid |\bar{M}_J - m| < r_J, X = x)$ for a sequence $r_J \to 0$. This is then used to show that, for any $m$ in the range of $\bar{p}_J(x,\cdot)$,

$$\lim_{J\to\infty} Pr(T_J(\bar{M}_J, X) \leq T_J(m,x)) - \bar{p}_J^{-1}(m;x) = 0. \tag{A.5}$$

Intuitively, (A.5) follows because (A.4) implies that $T_J(m,x) \approx p_{j_0}(\bar{p}_J^{-1}(m;x))$ and hence

$$\begin{aligned} Pr(T_J(\bar{M}_J, X) \leq T_J(m,x)) &\approx Pr(p_{j_0}(\bar{p}_J^{-1}(\bar{M}_J;X)) \leq p_{j_0}(\bar{p}_J^{-1}(m;x))) \\ &\approx Pr(p_{j_0}(\bar{p}_J^{-1}(\bar{p}_J(\theta,X);X)) \leq p_{j_0}(\bar{p}_J^{-1}(m;x))) \\ &= Pr(p_{j_0}(\theta) \leq p_{j_0}(\bar{p}_J^{-1}(m;x))) \\ &= \bar{p}_J^{-1}(m;x). \end{aligned} \tag{A.6}$$

Now, let $S_1(m,x) = E(Y \mid |\bar{M}_J - m| < r_J, X = x)$ and $S_2(m,x) = Pr(T_J(\bar{M}_J, X) \leq T_J(m,x))$ and suppose, for the sake of argument, that $S_2$ is invertible in $m$. Then

$$\begin{aligned} |S_1(S_2^{-1}(t;x),x) - G(x,t)| \leq\ & |S_1(S_2^{-1}(t;x),x) - G(x,\bar{p}_J^{-1}(S_2^{-1}(t;x);x))| \\ &+ |G(x,\bar{p}_J^{-1}(S_2^{-1}(t;x);x)) - G(x,\bar{p}_J^{-1}(\bar{p}_J(x,t);x))|. \end{aligned} \tag{A.7}$$

But $S_2$ is not invertible since, for a fixed $J$, the support of $\bar{M}_J$ is finite. Thus the proof of Theorem 2.1 involves showing (a) that a similar expansion still holds (b) that the convergence in (A.3) holds uniformly so that the first term in this expansion converges to 0, and (c) that the convergence in (A.5) and continuity of $G(x,\bar{p}_J^{-1}(m;x))$, as a function of $m$, imply that the second term in the expansion converges to 0.

## A.2 Proof of Theorem 2.1

When considering two models $\gamma_0, \gamma \in \Gamma_J$ I will use notation $p_{j,0}$, $\bar{p}_{J0}$, $F^0_{\theta|X}$, $\Theta_0(x)$, etc. to denote the elements of the model corresponding to $\gamma_0$ and $p_j$, $\bar{p}_J$, $F_{\theta|X}$, $\Theta(x)$, etc. to denote the parameters of the model corresponding to $\gamma$.

**Proof of Theorem 2.1** Under Assumption 2.1, the CASF is given by

$$\begin{aligned} G(x,t) &:= E(Y \mid X = x, \theta = t) \\ &= \int g(x,t,u) dF_{U|X,\theta}(u \mid x,t) \\ &= \int g(x,t,u) dF_U(u). \end{aligned} \tag{A.8}$$

Given this mapping from the model parameter $\gamma$ to the object $G$ and given an arbitrary $\gamma_0 \in \Gamma_J$, the identified set $\mathcal{I}_J(\gamma_0; G(\cdot))$ is defined by equation (2.4). Let $G \in \mathcal{I}_J(\gamma_0; G(\cdot))$. Then there exists $\gamma \in \Gamma_J$ such that $\mathbb{P}_J(\gamma) = \mathbb{P}_J(\gamma_0)$ and $G(x,t) = \int g(x,t,u) dF_U(u)$ and $G_0(x,t) = \int g_0(x,t,u) dF^0_U(u)$.

Fix $x \in \mathcal{X}$ and $t_0 \in \Theta_0(x)$ and define $m_0 := \bar{p}_{J0}(x,t_0)$. By the triangle inequality, for any $r_J > 0$,

$$\begin{aligned} |G_0(x,t_0) - G(x,t_0)| \leq\ & |G_0(x,t_0) - E(Y \mid |\bar{M}_J - m_0| \leq r_J, X = x)| \\ & + |G(x,t) - E(Y \mid |\bar{M}_J - m_0| \leq r_J, X = x)| \end{aligned} \tag{A.9}$$

By Lemma A.3 there exists a constant $A > 0$ such that, for $J$ sufficiently large,

$$|m_0 - \bar{p}_J(x,t_0)| = |\bar{p}_{J0}(x,t_0) - \bar{p}_J(x,t_0)| \leq A \left(log(J)/J\right)^{1/2}.$$

Then for $r_J = 2 \max\{A, 1+\epsilon\} \left(log(J)/J\right)^{1/2}$, for $\epsilon > 0$,

$$\begin{aligned} & \sup_{\gamma_0 \in \Gamma_J, G \in \mathcal{I}_J(\gamma_0; G(\cdot))} ||G_0 - G|| \\ & \leq 2 \sup_{\mathbb{P}^0_J} \sup_{\gamma: \mathbb{P}_J(\gamma) = \mathbb{P}^0_J} \sup_{\substack{x \in \mathcal{X}, t \in \Theta(x) \\ m_0 \in [0,1]: |m_0 - \bar{p}_J(x,t)| \leq r_J/2}} \left| G(x,t) - E(Y \mid |\bar{M}_J - m_0| \leq r_J, X = x) \right| \\ & = O(r_J) \end{aligned} \tag{A.10}$$

by Lemma A.4. And, therefore, $\lim_{J \to \infty} \sup_{\gamma_0 \in \Gamma_J, G \in \mathcal{I}_J(\gamma_0; G(\cdot))} ||G_0 - G|| = 0$, as desired.

□

**Lemma A.3.** *Under the assumptions of Theorem 2.1*

$$\sup_{\gamma_0,\gamma\in\Gamma_J:\mathbb{P}_J(\gamma)=\mathbb{P}_J(\gamma_0)}\ \sup_{x\in\mathcal{X},t_0\in\Theta_0(x)}|\bar{p}_{J0}(x,t_0)-\bar{p}_J(x,t_0)|=O\left((log(J)/J)^{1/2}\right).$$

**Lemma A.4.** *Under the assumptions of Theorem 2.1, for* $r_J=A^*\left(log(J)/J\right)^{1/2}$*, for* $A^*>2$*,*

$$\sup_{\mathbb{P}_J^0}\ \sup_{\gamma:\mathbb{P}_J(\gamma)=\mathbb{P}_J^0}\ \sup_{\substack{x\in\mathcal{X},t\in\Theta(x)\\ m_0\in[0,1]:|m_0-\bar{p}_J(x,t)|\leq r_J/2}}\left|G(x,t)-E(Y\mid|\bar{M}_J-m_0|\leq r_J,X=x)\right|=O(r_J).$$

## A.3 Proof of Lemmas A.3 and A.4

**Proof of Lemma A.3**

**Step 1:** I will first show that there exists $J_0(c,C)$ such that, for all $x\in\mathcal{X}$ and $t_0\in\Theta_0(x)$ and all $J\geq J_0(c,C)$, $\exists t\in\Theta(x)$ such that

$$|\bar{p}_{J0}(x,t_0)-\bar{p}_J(x,t)|\leq 2\left(\frac{log(J)}{J}\right)^{1/2}. \tag{A.11}$$

Fix $x\in\mathcal{X}$ and $t_0\in\Theta_0(x)$ and let $a_J=\left(\frac{log(J)}{J}\right)^{1/2}$. I will first show that

$$Pr(|\bar{M}_J-\bar{p}_{J0}(x,t_0)|\leq a_J\mid X=x)\geq\frac{ca_J}{C}-2c^{-1}\exp(-\frac{1}{2}Ja_J^2). \tag{A.12}$$

First, by iterating expectations and then restricting the range of $\theta$,

$$\begin{aligned}
&Pr(|\bar{M}_J-\bar{p}_{J0}(x,t_0)|\leq a_J\mid X=x) \qquad (A.13)\\
&=\int Pr(|\bar{M}_J-\bar{p}_{J0}(x,t_0)|\leq a_J\mid X=x,\theta=\tau)dF^0_{\theta|X=x}(\tau)\\
&\geq\int_{\tau:|\bar{p}_{J0}(x,\tau)-\bar{p}_{J0}(x,t_0)|\leq a_J/2}Pr(|\bar{M}_J-\bar{p}_{J0}(x,t_0)|\leq a_J\mid X=x,\theta=\tau)dF^0_{\theta|X=x}(\tau)\\
&=\int_{\tau:|\bar{p}_{J0}(x,\tau)-\bar{p}_{J0}(x,t_0)|\leq a_J/2}\left(1-Pr(|\bar{M}_J-\bar{p}_{J0}(x,t_0)|>a_J\mid X=x,\theta=\tau)\right)dF^0_{\theta|X=x}(\tau)\\
&\geq\int_{\tau:|\bar{p}_{J0}(x,\tau)-\bar{p}_{J0}(x,t_0)|\leq a_J/2}\left(1-Pr(|\bar{M}_J-\bar{p}_{J0}(x,\tau)|>a_J/2\mid X=x,\theta=\tau)\right)dF^0_{\theta|X=x}(\tau)
\end{aligned}$$

where the last inequality follows because $|\bar{M}_J-\bar{p}_{J0}(x,\tau)|\geq|\bar{M}_J-\bar{p}_{J0}(x,t_0)|-|\bar{p}_{J0}(x,\tau)-\bar{p}_{J0}(x,t_0)|$.

Next, applying the law of iterated expectations in reverse,

$$\begin{aligned}
&\int_{\tau:|\bar{p}_{J0}(x,\tau)-\bar{p}_{J0}(x,t_0)|\leq a_J/2} \left(1 - Pr(|\bar{M}_J - \bar{p}_{J0}(x,\tau)| > a_J/2 \mid X = x, \theta = \tau)\right) dF^0_{\theta|X=x}(\tau) \\
&= \int_{\tau:|\bar{p}_{J0}(x,\tau)-\bar{p}_{J0}(x,t_0)|\leq a_J/2} dF^0_{\theta|X=x}(\tau) \qquad (A.14) \\
&- \int_{\tau:|\bar{p}_{J0}(x,\tau)-\bar{p}_{J0}(x,t_0)|\leq a_J/2} Pr(|\bar{M}_J - \bar{p}_{J0}(x,\tau)| > a_J/2 \mid X = x, \theta = \tau) dF^0_{\theta|X=x}(\tau) \\
&\geq Pr(|\bar{p}_{J0}(x,\theta) - \bar{p}_{J0}(x,t_0)| \leq a_J/2 \mid X = x) - Pr(|\bar{M}_J - \bar{p}_{J0}(x,\theta)| > a_J/2 \mid X = x).
\end{aligned}$$

By Assumption 2.7,

$$\begin{aligned}
Pr(|\bar{p}_{J0}(x,\theta) - \bar{p}_{J0}(x,t_0)| \leq a_J/2 \mid X = x) &\geq F^0_{\theta|X=x}(t_0 + \frac{a_J}{2C}) - F^0_{\theta|X=x}(t_0 - \frac{a_J}{2C}) \\
&\geq \frac{ca_J}{C}, \qquad (A.15)
\end{aligned}$$

and, applying Lemma A.1,

$$\begin{aligned}
Pr(|\bar{M}_J - \bar{p}_{J0}(x,\theta)| > a_J/2 \mid X = x) &\leq \frac{Pr(|\bar{M}_J - \bar{p}_{J0}(X,\theta)| > a_J/2, X = x)}{Pr(X = x)} \qquad (A.16) \\
&\leq 2c^{-1} \exp(-\frac{1}{2} J a_J^2).
\end{aligned}$$

Inequality (A.12) follows from (A.13)-(A.16).

On the other hand, consider the model parameterized by $\gamma \in \Gamma_J$. If $|\bar{M}_J - \bar{p}_{J0}(x,t_0)| \leq a_J$ then for any $\tau$ either $|\bar{p}_J(x,\tau) - \bar{p}_{J0}(x,t_0)| \leq 2a_J$ or $|\bar{M}_J - \bar{p}_J(x,\tau)| > a_J$. Therefore,

$$\begin{aligned}
&Pr(|\bar{M}_J - \bar{p}_{J0}(x,t_0)| \leq a_J \mid X = x) \qquad (A.17) \\
&= \int Pr(|\bar{M}_J - \bar{p}_{J0}(x,t_0)| \leq a_J \mid X = x, \theta = \tau) dF_{\theta|X=x}(\tau) \\
&\leq \int Pr(|\bar{p}_J(x,\tau) - \bar{p}_{J0}(x,t_0)| \leq 2a_J \mid X = x, \theta = \tau) dF_{\theta|X=x}(\tau) \\
&+ \int Pr(|\bar{M}_J - \bar{p}_J(x,\tau)| > a_J \mid X = x, \theta = \tau) dF_{\theta|X=x}(\tau) \\
&= Pr(|\bar{p}_J(x,\theta) - \bar{p}_{J0}(x,t_0)| \leq 2a_J \mid X = x) + Pr(|\bar{M}_J - \bar{p}_J(x,\theta)| > a_J \mid X = x).
\end{aligned}$$

Since $\gamma$ must also satisfy Assumption 2.5,

$$Pr(|\bar{M}_J - \bar{p}_J(x,\theta)| > a_J \mid X = x) \leq 2c^{-1} \exp(-2Ja_J^2). \qquad (A.18)$$

To prove the result by contradiction, suppose that $|\bar{p}_J(x,\tau) - \bar{p}_{J0}(x,t_0)| > 2a_J$ for all

$\tau \in \Theta(x)$. Then $Pr(|\bar{p}_J(x,\theta) - \bar{p}_{J0}(x,t_0)| < 2a_J \mid X = x) = 0$ and (A.12), (A.17), and (A.18) together imply that

$$\begin{aligned} \frac{c}{C}a_J - 2c^{-1}\exp(-\frac{1}{2}Ja_J^2) \leq Pr(|\bar{M}_J - \bar{p}_{J0}(x,t_0)| \leq a_J \mid X = x) \\ \leq 2c^{-1}\exp(-2Ja_J^2) \end{aligned} \tag{A.19}$$

which implies that

$$\frac{c}{C}\left(\frac{log(J)}{J}\right)^{1/2} \leq 2c^{-1}J^{-2} + 2c^{-1}J^{-1/2} \tag{A.20}$$

which implies a contradiction for large enough $J$. I can conclude that for all sufficiently large $J$, $\exists t \in \Theta(x)$ such that $|\bar{p}_{J0}(x,t_0) - \bar{p}_J(x,t)| \leq 2a_J$.

**Step 2:** I will next show that, because $|\bar{p}_{J0}(x,t_0) - \bar{p}_J(x,t)| \leq 2\left(\frac{log(J)}{J}\right)^{1/2}$, it follows that

$$|t - t_0| = O(r_J). \tag{A.21}$$

For any $x' \in \mathcal{X}$ and $m_0' \in [0,1]$, define $T_J(m_0', x'; r_J) := E(M_{j_0} \mid |\bar{M}_J - m_0'| \leq r_J, X = x')$. I will show below that for any $t_0' \in \Theta_0(x')$, if $|\bar{p}_{J0}(x', t_0') - m_0'| < r_J/2$ where $r_J = 4\left(\frac{log(J)}{J}\right)^{1/2}$ then

$$\begin{aligned} &|T_J(m_0', x'; r_J) - p_{j_0,0}(t_0')| \\ &\leq \delta_J := \frac{2B\exp(-2Jr_J^2)}{\frac{c^2}{2C}r_J - 2\exp(-\frac{1}{8}Jr_J^2)} + \frac{3C}{c}r_J. \end{aligned} \tag{A.22}$$

This implies that

$$\begin{aligned} Pr(|T_J(\bar{M}_J, X; r_J) - p_{j_0,0}(\theta)| > \delta_J) \leq Pr(|\bar{p}_{J0}(X,\theta) - \bar{M}_J| > r_J/2) \\ \leq \rho_J := 2\exp(-\frac{1}{2}Jr_J^2) \end{aligned} \tag{A.23}$$

where the second line follows from Lemma A.1.

Since $\gamma$ is observationally equivalent to $\gamma_0$, the same argument shows that for any $t' \in \Theta(x')$, if $|\bar{p}_J(x', t') - m_0'| \leq r_J/2$ then

$$|T_J(m_0', x'; r_J) - p_{j_0}(t')| \leq \delta_J \tag{A.24}$$

and hence

$$Pr(|T_J(\bar{M}_J, X; r_J) - p_{j_0}(\theta)| > \delta_J) \leq Pr(|\bar{p}_J(X, \theta) - \bar{M}_J| > r_J/2) \leq \rho_J. \tag{A.25}$$

It can also be concluded from (A.22) and (A.24) that for $m_0 = \bar{p}_{J0}(x, t_0)$, since, by assumption, $|m_0 - \bar{p}_J(x, t)| \leq r_J/2$,

$$\begin{aligned} |p_{j_0,0}(t_0) - p_{j_0}(t)| &\leq |T_J(m_0, x; r_J) - p_{j_0,0}(t_0)| + |T_J(m_0, x; r_J) - p_{j_0}(t)| \\ &\leq 2\delta_J. \end{aligned} \tag{A.26}$$

Then (A.23) implies that

$$|Pr(T_J(\bar{M}_J, X; r_J) \leq p_{j_0,0}(t_0)) - t_0| \leq \frac{\delta_J}{c} + \rho_J \tag{A.27}$$

by an application of Lemma A.2 with $A = T_J(\bar{M}_J, X; r_J)$, $B = p_{j_0,0}(\theta)$, and $z = z' = p_{j_0,0}(t_0)$ since $Pr(p_{j_0,0}(\theta) \leq p_{j_0,0}(t_0)) = t_0$ and the distribution function $Pr(p_{j_0,0}(\theta) \leq z)$ is Lipschitz continuous with $f_{\bar{B}} = 1/c$.

Similarly, (A.25) and (A.26) imply that

$$|Pr(T_J(\bar{M}_J, X; r_J) \leq p_{j_0,0}(t_0)) - t| \leq \frac{3\delta_J}{c} + \rho_J \tag{A.28}$$

by an application of Lemma A.2 with $A = T_J(\bar{M}_J, X; r_J)$, $B = p_{j_0}(\theta)$, $z = p_{j_0,0}(t_0)$ and $z' = p_{j_0}(t)$.

Then (A.27) and (A.28) imply that $|t - t_0| \leq \frac{4\delta_J}{c} + 2\rho_J$. The desired result follows since, plugging in $r_J = 4\left(\frac{log(J)}{J}\right)^{1/2}$,

$$\frac{4\delta_J}{c} + 2\rho_J = \frac{4}{c}\left(\frac{2BJ^{-32}}{\frac{2c^2}{C}(\log(J)/J)^{1/2} - 2J^{-2}} + \frac{3C}{c}r_J\right) + 4J^{-8} = O(r_J). \tag{A.29}$$

It remains to show that (A.22) holds for any $x' \in \mathcal{X}$, any $m_0' \in [0, 1]$ and any $t_0' \in \Theta_0(x')$ for which $|\bar{p}_{J0}(x', t_0') - m_0'| < r_J/2$. The proof of this is almost identical to the proof of Lemma A.4 so I will provide only a sketch.

First, by Assumption 2.5, $E(M_{j_0} \mid |\bar{M}_J - m_0'| \leq r_J, X = x') = E(E(M_{j_0} \mid X, \theta) \mid$

$|\bar{M}_J - m_0'| \leq r_J, X = x')$. Therefore, since, by Assumption 2.4, $E(M_{j_0} \mid X = x', \theta) = p_{j_0,0}(\theta)$,

$$\begin{aligned}
&\left|p_{j_0,0}(t_0') - E(M_{j_0} \mid |\bar{M}_J - m_0'| \leq r_J, X = x')\right| \\
&= |p_{j_0,0}(t_0') - E(E(M_{j_0} \mid X = x', \theta) \mid |\bar{M}_J - m_0'| \leq r_J, X = x')| \\
&\leq \left| \int (p_{j_0,0}(t_0') - p_{j_0,0}(\tau))\, dF^0_{\theta||\bar{M}_J - m_0'| \leq r_J, X = x'}(\tau) \right| \\
&\leq \frac{3C}{c} r_J + BPr(|\bar{p}_{J0}(x', \theta) - \bar{p}_{J0}(x', t_0')| > 3r_J \mid |\bar{M}_J - m_0'| \leq r_J, X = x') \qquad \text{(A.30)} \\
&\leq \frac{3C}{c} r_J + BPr(|\bar{M}_J - \bar{p}_{J0}(x', \theta)| > r_J \mid |\bar{M}_J - m_0'| \leq r_J, X = x') \\
&\leq \frac{3C}{c} r_J + \frac{2B \exp(-2Jr_J^2)}{\frac{c^2}{2C} r_J - 2\exp(-\frac{1}{8} J r_J^2)}
\end{aligned}$$

where the third inequality follows because $|\bar{p}_{J0}(x', t_0') - m_0'| < r_J/2$.

**Step 3:** Combining equations (A.11) and (A.21),

$$\begin{aligned}
|\bar{p}_{J0}(x, t_0) - \bar{p}_J(x, t_0)| &\leq |\bar{p}_{J0}(x, t_0) - \bar{p}_J(x, t)| + |\bar{p}_J(x, t) - \bar{p}_J(x, t_0)| \\
&\leq C|t_0 - t| + 2\left(\frac{log(J)}{J}\right)^{1/2} \qquad \text{(A.31)} \\
&= O\left(\left(\frac{log(J)}{J}\right)^{1/2}\right)
\end{aligned}$$

where the second line uses Assumption 2.7.

□

**Proof of Lemma A.4** Consider any $x \in \mathcal{X}$ and $t \in \Theta(x)$ and let $m_0$ be such that $|m_0 - \bar{p}_J(x, t)| \leq r_J/2$.

First, since $\mathbb{P}_J(\gamma) = \mathbb{P}_J^0$ and $\gamma$ satisfies Assumptions 2.1 and 2.5, $E(Y \mid |\bar{M}_J - m_0| \leq r_J, X = x) = \int G(x, \tau) dF_{\theta||\bar{M}_J - m_0| \leq r_J, X = x}(\tau)$. Therefore,

$$\begin{aligned}
&|G(x, t) - E(Y \mid |\bar{M}_J - m_0| \leq r_J, X = x)| \\
&= \left| \int (G(x, t) - G(x, \tau))\, dF_{\theta||\bar{M}_J - m_0| \leq r_J, X = x}(\tau) \right| \\
&\leq \int_{\tau: |\bar{p}_J(x,\tau) - \bar{p}_J(x,t)| \leq 3r_J} |G(x, \tau) - G(x, t)| dF_{\theta||\bar{M}_J - m_0| \leq r_J, X = x}(\tau) \qquad \text{(A.32)} \\
&+ \int_{\tau: |\bar{p}_J(x,\tau) - \bar{p}_J(x,t)| > 3r_J} |G(x, \tau) - G(x, t)| dF_{\theta||\bar{M}_J - m_0| \leq r_J, X = x}(\tau) \\
&\leq \frac{3C}{c} r_J + BPr(|\bar{p}_J(x, \theta) - \bar{p}_J(x, t)| > 3r_J \mid |\bar{M}_J - m_0| \leq r_J, X = x).
\end{aligned}$$

The first term in the final line follows because $|G(x,\tau) - G(x,t)| \leq C|\tau - t|$ and because $|\bar{p}_J(x,\tau) - \bar{p}_J(x,t)| \geq c|\tau - t|$. The second term in the final line of (A.35) follows because $G(x,\cdot)$ is uniformly continuous on a compact subset of $\mathbb{R}$ for each $x$ and $|\mathcal{X}|$ is finite and therefore there is some positive constant $B < \infty$ such that $\sup_{x\in\mathcal{X},t\in\Theta(x)} |G(x,t)| \leq B/2$.

Next, if $|m_0 - \bar{p}_J(x,t)| \leq r_J/2$, $|\bar{p}_J(x,\theta) - \bar{p}_J(x,t)| > 3r_J$, and $|\bar{M}_J - m_0| \leq r_J$ then

$$|\bar{M}_J - \bar{p}_J(x,\theta)| \geq |\bar{p}_J(x,\theta) - \bar{p}_J(x,t)| - |\bar{M}_J - m_0| - |m_0 - \bar{p}_J(x,t)| > r_J \tag{A.33}$$

and therefore

$$\begin{aligned} &Pr(|\bar{p}_J(x,\theta) - \bar{p}_J(x,t)| > 3r_J \mid |\bar{M}_J - m_0| \leq r_J, X = x) \\ &\leq Pr(|\bar{M}_J - \bar{p}_J(x,\theta)| > r_J \mid |\bar{M}_J - m_0| \leq r_J, X = x) \end{aligned} \tag{A.34}$$

Next,

$$\begin{aligned} &Pr(|\bar{M}_J - \bar{p}_J(x,\theta)| > r_J \mid |\bar{M}_J - m_0| \leq r_J, X = x) \\ &= \frac{Pr(|\bar{M}_J - \bar{p}_J(X,\theta)| > r_J, |\bar{M}_J - m_0| \leq r_J, X = x)}{Pr(|\bar{M}_J - m_0| \leq r_J, X = x)} \\ &\leq \frac{Pr(|\bar{M}_J - \bar{p}_J(X,\theta)| > r_J)}{Pr(|\bar{M}_J - m_0| \leq r_J, X = x)}. \end{aligned} \tag{A.35}$$

Applying Lemma A.1, since $\gamma$ must satisfy Assumption 2.5,

$$Pr(|\bar{M}_J - \bar{p}_J(X,\theta)| > r_J) \leq 2\exp(-2Jr_J^2) \tag{A.36}$$

Combining this with equations (A.32)-(A.35),

$$\begin{aligned} &\left|G(x,t) - E(Y \mid |\bar{M}_J - m_0| \leq r_J, X = x)\right| \\ &\leq \frac{2B\exp(-2Jr_J^2)}{Pr(|\bar{M}_J - m_0| \leq r_J, X = x)} + \frac{3C}{c} r_J \end{aligned} \tag{A.37}$$

The desired result follows because $Pr(|\bar{M}_J - m_0| \leq r_J, X = x) = Pr(|\bar{M}_J - m_0| \leq r_J \mid X = x)Pr(X = x)$, $Pr(X = x) \geq c$, and

$$Pr(|\bar{M}_J - m_0| \leq r_J \mid X = x) \geq \frac{cr_J}{2C} - 2c^{-1}\exp(-\frac{1}{8}Jr_J^2). \tag{A.38}$$

The proof is concluded by proving (A.38) since

$$\frac{2B\exp(-2Jr_J^2)}{\frac{c^2 r_J}{2C} - 2\exp(-\frac{1}{8}Jr_J^2)} = \frac{2BJ^{-2A^{*2}}}{\frac{A^* c^2}{2C}(\log(J)/J)^{1/2} - 2J^{-A^{*2}/8}} = O(r_J). \tag{A.39}$$

since $A^* > 2$.

Since $|m_0 - \bar{p}_J(x,t)| \leq r_J/2$, $Pr(|\bar{M}_J - m_0| \leq r_J \mid X = x) \geq Pr(|\bar{M}_J - \bar{p}_J(x,t)| \leq r_J/2 \mid X = x)$. Following the arguments in lines (A.13) and (A.14) of the proof of Lemma A.3,

$$\begin{aligned} &Pr(|\bar{M}_J - \bar{p}_J(x,t)| \leq r_J/2 \mid X = x) \\ &\geq Pr(\bar{p}_J(x,\theta) - \bar{p}_J(x,t)| \leq r_J/4 \mid X = x) - Pr(|\bar{M}_J - \bar{p}_J(x,\theta)| > r_J/4 \mid X = x). \end{aligned} \tag{A.40}$$

By Assumption 2.7,

$$\begin{aligned} &Pr(|\bar{p}_J(x,\theta) - \bar{p}_J(x,t)| \leq r_J/4 \mid X = x) \\ &\geq F_{\theta|X=x}(\bar{p}_J(x,t) + \frac{r_J}{4C}) - F_{\theta|X=x}(\bar{p}_J(x,t) - \frac{r_J}{4C}) \\ &\geq \frac{c r_J}{2C}, \end{aligned} \tag{A.41}$$

and, applying Lemma A.1,

$$Pr(|\bar{M}_J - \bar{p}_J(x,\theta)| > r_J/4 \mid X = x) \leq 2c^{-1}\exp(-\frac{1}{8}Jr_J^2). \tag{A.42}$$

This proves inequality (A.38) and completes the proof. □

# References


Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011): "Personality Psychology and Economics," *Handbook of the economics of education*, 4.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2013): "Program evaluation with high-dimensional data," *arXiv preprint arXiv:1311.2645*.

Bloom, N. and J. Van Reenen (2007): "Measuring and Explaining Management Practices across Firms and Countries," *The Quarterly Journal of Economics*, 1351–1408.

Blundell, R. and J. Powell (2003): "Endogeneity in semiparametric and nonparametric regression models," in *Advances in Economics and Econometrics: Theory and Applications*, ed. by L. Dewatripont, M. abd Hansen and S. Turnovsky, Cambridge University Press, 312 – 357.

Carneiro, P., K. T. Hansen, and J. J. Heckman (2003): "2001 Lawrence R. Klein Lecture: Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice," *International Economic Review*, 44, 361–422.

Chen, X., H. Hong, and D. Nekipelov (2011): "Nonlinear models of measurement errors," *Journal of Economic Literature*, 49, 901–937.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013): "Average and quantile effects in nonseparable panel models," *Econometrica*, 81, 535–580.

Clinton, J., S. Jackman, and D. Rivers (2004): "The statistical analysis of roll call data," *American Political Science Review*, 98, 355–370.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the technology of cognitive and noncognitive skill formation," *Econometrica*, 78, 883–931.

de Finetti, B. (1931): "Funzione Caratteristica di un fenomeno allatoria," *Atti della R. Accademia Nazionale dei Lincii Ser. 6*, 4, 251 – 299.

de Jong, R. M. and T. Woutersen (2011): "Dynamic time series binary choice," *Econometric Theory*, 27, 673–702.

Dee, T. S. (2004): "Are there civic returns to education?" *Journal of Public Economics*, 88, 1697–1720.

Diaconis, P. and D. Freedman (1980): "Finite exchangeable sequences," *The Annals of Probability*, 745–764.

Douglas, J. (1997): "Joint consistency of nonparametric item characteristic curve and ability estimation," *Psychometrika*, 62, 7–28.

——— (2001): "Asymptotic identifiability of nonparametric item response models," *Psychometrika*, 66, 531–540.

Dvoretzky, A. et al. (1972): "Asymptotic normality for sums of dependent random variables," in *Proc. 6th Berkeley Symp. Math. Statist. Probab*, vol. 2, 513–535.

Gawade, N. G. (2007): "Measurement Error in Discrete Explanatory Variables: Implications of Conditional Independence," working paper, Princeton University.

Hall, P. (1992): "On bootstrap confidence intervals in nonparametric regression," *The Annals of Statistics*, 695–711.

Hansen, K. T., J. J. Heckman, and K. J. Mullen (2004): "The effect of schooling and ability on achievement test scores," *Journal of Econometrics*, 121, 39 – 98.

Heckman, J. J. (2001): "Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture," *Journal of Political Economy*, 109, 673–748.

Heckman, J. J., D. Schmierer, and S. Urzua (2010): "Testing the correlated random coefficient model," *Journal of Econometrics*, 158, 177–203.

Heckman, J. J. and J. M. Snyder (1997): "Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators," *The Rand Journal of Economics*, 28, S142.

Heckman, J. J., J. Stixrud, and S. Urzua (2006a): "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior," *Journal of Labor Economics*, 24, 411–482.

Heckman, J. J., S. Urzua, and E. Vytlacil (2006b): "Understanding instrumental variables in models with essential heterogeneity," *The Review of Economics and Statistics*, 88, 389–432.

Honore, B. E. and E. Tamer (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74, 611–629.

Hu, Y. (2008): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution," *Journal of Econometrics*, 144, 27–61.

Junker, B., L. S. Schofield, and L. J. Taylor (2012): "The use of cognitive ability measures as explanatory variables in regression analysis," *IZA Journal of Labor Economics*, 1, 1–19.

Junker, B. W. and J. L. Ellis (1997): "A characterization of monotone unidimensional latent variable models," *The Annals of Statistics*, 1327–1343.

Kapetanios, G. (2008): "A bootstrap procedure for panel data sets with many cross-sectional units," *Econometrics Journal*, 11, 377–395.

Klein, T. J. (2013): "College education and wages in the UK: estimating conditional average structural functions in nonadditive models with binary endogenous variables," *Empirical Economics*, 44, 135–161.

Longford, N. T. (1999): "Selection bias and treatment heterogeneity in clinical trials," *Statistics in medicine*, 18, 1467–1474.

Lord, F. M. (1980): *Applications of item response theory to practical testing problems*, Routledge.

Mahajan, A. (2006): "Identification and Estimation of Single Index Models with Misclassified Regressor," *Econometrica*, 74, 631–665.

Mammen, E., C. Rothe, and M. Schienle (2012): "Nonparametric regression with nonparametrically generated covariates," *The Annals of Statistics*, 40, 1132–1170.

——— (2016): "Semiparametric estimation with generated covariates," *Econometric Theory*, 32, 1140–1177.

Matzkin, R. L. (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339–1375.

——— (2004): "Unobservable Instruments," unpublished mimeo, Northwestern University.

McLeish, D. L. et al. (1975): "A maximal inequality and dependent strong laws," *The Annals of probability*, 3, 829–839.

Neal, D. A. and W. R. Johnson (1996): "The Role of Premarket Factors in Black-White Wage Differences," *Journal of Political Economy*, 104, 869–895.

Poole, K. T. and H. Rosenthal (1985): "A spatial model for legislative roll call analysis," *American Journal of Political Science*, 357–384.

——— (1997): "Congress," *A Political-Economic History of Roll Call Voting. New York*.

Ramsay, J. (1991): "Kernel smoothing approaches to nonparametric item characteristic curve estimation," *Psychometrika*, 56, 611–630.

Schofield, L. S. (2014): "Measurement error in the AFQT in the NLSY79," *Economics letters*, 123, 262–265.

Sijtsma, K. and B. W. Junker (2006): "Item response theory: Past performance, present developments, and future expectations," *Behaviormetrika*, 33, 75–102.

Spady, R. H. (2007): "Semiparametric methods for the measurement of latent attitudes and the estimation of their behavioral consequences," Working Paper CWP26/07, CEMMAP.

van der Linden, W. J. and R. K. Hambleton (2013): *Handbook of modern item response theory*, New York, NY: Springer.

Williams, B. (2012): "Latent Variables Models," Ph.D. thesis, University of Chicago, Department of Economics.

——— (2013): "A Measurement Model with Discrete Measurements and Continuous Latent Variables," unpublished manuscript, George Washington University.

Winship, C. and S. Korenman (1997): "Does staying in school make you smarter? The effect of education on IQ in The Bell Curve," in *Intelligence, Genes, and Success*, Springer, 215–234.