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

Benjamin Williams*

September 29, 2016

Abstract

In this paper I study identification of a nonseparable model with endogeneity arising due to unobserved heterogeneity. Identification is not based on an instrumental variable that is excluded from the outcome equation. Instead, it relies on the availability of binary proxies for the unobserved heterogeneity. When the support of the unobserved heterogeneity is not finite, the model is identified only in the limit as the number of proxies increases. This allows consistent estimation 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, and Elie Tamer 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 on a previous version and seminar participants at the Center for Education Research at UW-Madison, the FDIC, George Washington University, Johns Hopkins University, the University of Chicago and the 2013 Latin American Workshop in Econometrics.

1 Introduction

This paper considers the problem of controlling for latent heterogeneity in a nonseparable model using multiple binary measurements or proxies. This empirical problem arises in many applications in economics. The binary proxies may be questions on an exam, results of a personality test or psychological assessment, or responses to opinion surveys. 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 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. The binary proxies may instead 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 Jr, 1997; Poole and Rosenthal, 1985, 1997), one may wish to separate the effect of observed legislator attributes on a key vote from the effect of latent legislator preferences. Votes on other bills can be used as proxies if they are all driven by the same latent preferences.

One solution to this problem is to avoid the need to control for the latent variable by instead using panel data to difference it out or finding an instrumental variable that is excluded from the outcome equation and independent of the latent variable. These strategies for identifying structural effects have been extended to nonseparable models as well. This paper, however, is motivated by the many economic applications where it is of interest to know how the structural effects vary across the distribution of the latent variable.

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 a another model – and is often not justified theoretically. One contribution of this paper is to provide conditions under which 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

[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).

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, or 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. In addition, for the identification results in Section 3, I allow for weak conditional dependence.

The identification problem consists of two separate parts. The first problem is that the proxies are binary while the latent variable is continuous. Mahajan (2006), Gawade (2007), and Hu (2008) restrict the support of the latent variable to be discrete for this reason. Carneiro et al. (2003) instead use an argument based on additive separability and support conditions, similar to the special regressor argument in a related class of models (Lewbel, 1998, 2014). The argument uses continuous variation in observed covariates to trace out the distribution of an additive latent index underlying each binary proxy. Without additive separability and a support condition on observed covariates, the model is not nonparametrically point identified. While this is a negative result concerning identification, I provide two positive results in this paper. First, while the model is not point identified, there are nontrivial bounds on parameters of interest. Second, under sufficient regularity conditions the width of these bounds converges to zero as a function of the number of binary proxies. Using this second result it is possible to construct estimators that are consistent as both the sample size and the number of proxies grow.

The second part of the identification problem arises only when the model contains observed covariates that are present in the structural outcome equation as well as the equations for the proxies. This is overcome in Carneiro et al. (2003) through the availability of an exogenous instrument. I instead show how restrictions in the model for the proxies can be used. In particular, for the main results in this paper I assume that one of the binary proxies is conditionally independent of the endogenous choice variable.[2] For example, suppose the

[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

binary proxies are the individual questions on a test of ability. Depending on the content of the test and how it is administered to survey participants, the response to items may not be independent of schooling conditional on ability (Hansen et al., 2004). However, a question that involves only content learned much earlier in school should not depend on whether the individual is currently in high school, a high school dropout, or a high school graduate. This exclusion restriction is fundamental to the main identification argument, in Section 3, but can be replaced with alternative restrictions as discussed in Section 5. For example, controlling separately for years of schooling at the time of the test and final years of schooling, as in Hansen et al. (2004), can be used to identify variation in latent ability.

I demonstrate the methods developed in this paper through two empirical illustrations. For the first illustration, 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 second illustration uses 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 (i) estimate the effect of education on responses to individual questions on this test and (ii) estimate the effect of education on earnings at age 30 while controlling for latent ability.

The remainder of the paper is organized as follows. In Section 2, the assumptions of the model are stated formally and discussed in the context of the two applications studied in this paper, and numerical examples of the identified set are presented. In Section 3, the main point identification result is presented in Section 3.1, an estimator is proposed and its asymptotic properties are investigated in Section 3.2, and a Monte Carlo study is presented in Section 3.3. In Section 4, I provide two empirical illustrations of the proposed methods. In Section 5, I discuss some extensions of the model. Section 6 concludes.

2 The model

In this section I outline the model. The outcome variable is Y, and X denotes a vector of observed covariates with finite support, $\mathcal{X}$. I assume that

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

for one proxy is excluded from the outcome equation and from the latent index for all other proxies.

where θ and U are both unobserved.

The identification results in this paper are fundamentally about how to obtain variation in θ from multiple binary proxies. While existing methods, such as Carneiro et al. (2003), rely on additivity in the outcome equation, it is straightforward to apply my basic identification results in the nonseparable model of (2.1). Moreover, the prevalence of unobserved heterogeneity in the effects of economic choices on outcomes has been widely recognized by economists (Heckman, 2001; Heckman et al., 2010), as well as in other areas of research (see, for example, Longford, 1999) and this is captured by the nonseparability in this model.

In this model, the average structural function (Blundell and Powell, 2003) is

$$\begin{aligned} ASF(x) &= E(g(x,\theta,U)) \\ &= \int g(x,t,u)dF_{\theta,U}(t,u) \end{aligned}$$

The function

$$\begin{aligned} CASF(x,t) &:= E(g(x,t,U)) \\ &= \int g(x,t,u)dF_U(u) \end{aligned}$$

can be thought of as a *conditional* average structural function that describes how the structural function varies with unobserved heterogeneity, θ, averaging out the other components of the unobserved heterogeneity, U. Alternatively, this is simply how the average structural function would be defined if θ were observed. In this paper I consider identification of these two objects.

First, I make the following assumptions, which would be sufficient for identification of either $ASF(x)$ or $CASF(x,t)$ if θ were observed (cf. Matzkin, 2004, 2003). I use the notation "$\perp\!\!\!\perp$" here and throughout the rest of the paper to denote independence.

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

Assumption 2.2. $support(\theta \mid X) = support(\theta)$

Under Assumption 2.1,

$$\begin{aligned} E(Y \mid X=x,\theta=t) &= \int g(x,t,u)dF_{U|X,\theta}(u \mid x,t) \\ &= \int g(x,t,u)dF_U(u) \\ &= CASF(x,t) \end{aligned}$$

For any $x \in \mathcal{X}$, under Assumption 2.2, $CASF(x,t)$ is defined for every $t \in \mathit{support}(\theta)$. Therefore, $ASF(x) = \int CASF(x,t)dF_\theta(t)$.

While identification of $ASF(x)$ would require only that X is conditionally independent of U given θ, in models that satisfy Assumption 2.1 the $CASF$ is identified as well. One advantage of the approach taken in this paper relative to an instrumental variable approach is the ability to identify heterogeneity in the structural function. For this reason, I maintain Assumption 2.1 rather than the weaker conditional independence.

Assumption 2.1 can be justified by an economic model where X is a choice made in order to maximize Y, based on partial information that includes θ but not U. When X denotes years of schooling and Y earnings, U can be thought of as an earnings shock that occurs after the schooling decision. In the other example used in this paper to illustrate the model, X denotes years of schooling as well but Y is a measure of voting behavior. While schooling is not chosen in order to attain optimal voting behavior, the independence assumption can still 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.

Next, let $M = (M_1, \dots, M_J)$ denote the binary proxies with

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

The system of equations, (2.1) and (2.2), is assumed to satisfy the following condition.

Assumption 2.3.

(i) $U \perp\!\!\!\perp (\epsilon_1, \dots, \epsilon_J) \mid (X, \theta)$

(ii) $\epsilon_1, \dots, \epsilon_J$ *are mutually independent conditional on* (X, θ)

Assumption 2.3 implies that the proxies are conditionally independent of Y given (X, θ) and mutually independent conditional on (X, θ). Responses to questions on a test satisfy the second part of this assumption if they are related only through latent ability; an individual that answers one question correctly is more likely to answer any other question correctly only to the extent that this reveals that she has higher than average ability. If Y denotes earnings then the first part of the assumption is satisfied as long as better responses to these questions do not lead to higher earnings, as may be the case if the test is used to determine college acceptance or job placement. In the civic returns application, the proxies for "civic-mindedness" must not be related to voting behavior except due to the fact that these behaviors are all a function of this underlying trait. For this reason, if Y is an indicator for whether the individual is registered to vote, then a variable indicating whether the individual voted in the last election cannot be used as a proxy since being registered to vote is a prerequisite for voting.

This type of conditional independence is common in models of measurement error (see Chen et al., 2011). Indeed, under Assumptions 2.1-2.3 the model can be viewed as a non-standard measurement error problem where θ is the "mismeasured" covariate. Conditional independence is also satisfied in the model of Carneiro et al. (2003). In addition, condition (ii) is a standard assumption in item response theory (Sijtsma and Junker, 2006).

Nevertheless, this is a strong assumption that may be inappropriate in some applications. If, for example, $M_1, \ldots, M_J$ denote a series of questions asked on a test then there may be residual dependence among subsets of the proxies after conditioning on θ and X. This could happen if the test consists of multiple subtests each of which measures separate dimensions of ability. Additionally, there may be time series dependence among the M_j if they are measured sequentially over time. Fortunately, this assumption can be relaxed considerably when J is large, as discussed in the next section.

Before proceeding, I define, for each j, the reduced form conditional response functions, $p_j(x,t) := Pr(M_j = 1 \mid X = x, \theta = t) = F_{\varepsilon_j|X}(h_j(x,t) \mid x)$. Under Assumptions 2.1 and 2.3, for any $\mathcal{J} \subset \{1, \ldots, J\}$ and any $c \in \{0,1\}$,

$$E(Y^c \prod_{j \in \mathcal{J}} M_j \mid X = x) = \int CASF(x,t)^c p_{\mathcal{J}}(x,t) dF_{\theta|X}(t \mid x) \tag{2.3}$$

where $p_{\mathcal{J}} := \prod_{j \in \mathcal{J}} p_j$. One problem that is immediately evident is that one could define an observationally equivalent model using any monotonic transformation of θ. In this observationally equivalent model $CASF$ would be different, though ASF would not. In order to obtain results on identification of $CASF$, I impose the following normalization.

Assumption 2.4. $\theta \sim Uniform(0,1)$

In addition to normalizing the distribution of θ, this requires θ to be scalar. See Williams (2013) for a version of this model that allows multidimensional θ. While normalizing the distribution of θ is necessary as there is no information on the scale of the latent variable θ, alternative normalizations in the model are possible, as discussed in Section 5.

There is a more fundamental problem with identification in this model that may be less apparent and that remains after imposing Assumption 2.4. Namely, it is possible to define an observationally equivalent model based on $\tilde{\theta} = G(X, \theta)$ where $G(x, \cdot)$ 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 conditonal on X dramatically. For example, suppose $G(x, \cdot)$

is the conditional distribution function, $G(x,t) = F_{\theta|X}(t \mid x) = Pr(\theta \leq t \mid X = x)$. Then $\tilde{\theta} \mid X = x \sim Uniform(0,1)$. That is, any model satisfying Assumptions 2.1-2.4 with dependence between θ and X is observationally equivalent to another model satisfying these assumptions with θ independent of X. In order to resolve this problem I consider models that satisfy the following assumption.

Assumption 2.5. *For some $1 \leq j_0 \leq J$, for every $x, x' \in \mathcal{X}$, $p_{j_0}(x,t) = p_{j_0}(x',t)$ for all $t \in [0,1]$.*

Assumption 2.5 is an exclusion restriction. In equation (2.2), it is satisfied if $h_{j_0}(x,t) = h_{j_0}(x',t)$ 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 θ. Suppose, for example, that X denotes years of schooling. Years of schooling may be related to test scores if individuals with more schooling are more likely to have attained more years of schooling at the time the test was administered, as was the case in the administration of the AFQT to NLSY79 survey respondents (Hansen et al., 2004). Suppose, however, that everyone 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. 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.

Since $p_{j_0}(x,t)$ does not vary with x, I drop x as an argument in the notation for the remainder of the paper, and write it as $p_{j_0}(t)$. In Section 5, I show how Assumption 2.5 can be replaced by alternative restrictions.

Lastly, I consider models that are monotonic in the latent variable.

Assumption 2.6. *For each $x \in \mathcal{X}$ and each $j = 1, \ldots, J$, $p_j(x, \cdot)$ 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 they can be reoriented. This assumption rules out the scenario where the correct orientation is neither known *a priori* nor prescribed by an economic model. It is possible that, given an assumption of monotonicity, the correct orientation is identified in the model though I do not pursue this here. The monotonicity of the response functions, however is essential, though this can be weakened somewhat when the number of proxies is large as described in the next section.[3]

Given Assumptions 2.1-2.6 nontrivial bounds on $CASF$ can be obtained. First, stack the response functions into the length J vector $\mathbf{p}(\cdot,\cdot)$. The triple

$$\gamma = (CASF(\cdot,\cdot), \mathbf{p}(\cdot,\cdot), F_{\theta|X}(\cdot \mid \cdot))$$

contains the "reduced form" parameters of the model. Let Γ denote the parameter space, that is, the space of all triples γ satisfying Assumptions 2.4, 2.5, and 2.6. For any γ, let

$$\mu_{c,\mathcal{J}}(x;\gamma) = \int CASF(x,t)^c p_{\mathcal{J}}(x,t) dF_{\theta|X}(t \mid x) \tag{2.4}$$

Let $m_{c,\mathcal{J}}(x)$ denote the observed moments, $E(Y^c \prod_{j\in\mathcal{J}} M_j \mid X = x)$. Then let $\mu(x;\gamma)$ and $m(x)$ denote the set of all possible $\mu_{c,\mathcal{J}}(x;\gamma)$ and $m_{c,\mathcal{J}}(x)$ as c and $\mathcal{J}$ vary. Under Assumptions 2.1-2.6 there exists $\gamma_0 \in \Gamma$ such that $m(x) = \mu(x;\gamma_0)$ for all $x \in \mathcal{X}$.

Then the space of parameter values consistent with the model is

$$\{\gamma \in \Gamma : \mu(x;\gamma_0) = \mu(x;\gamma) \ \forall x \in \mathcal{X}\}$$

and the identified set for $CASF$ is given by

$$\begin{aligned}\{CASF(\cdot,\cdot) : \exists \mathbf{p}(\cdot,\cdot), F_{\theta|X}(\cdot \mid \cdot) \text{ such that } (CASF(\cdot,\cdot), \mathbf{p}(\cdot,\cdot), F_{\theta|X}(\cdot \mid \cdot)) \in \Gamma \\ \text{and } m(x) = \mu(x; CASF(\cdot,\cdot), \mathbf{p}(\cdot,\cdot), F_{\theta|X}(\cdot \mid \cdot)) \ \forall x \in \mathcal{X}\}\end{aligned} \tag{2.5}$$

The identified set for other features of the $CASF$, such as the ASF or treatment effects $\Delta_{x,x'} := ASF(x') - ASF(x)$, can be obtained by projection. Also, additional restrictions can easily be applied by restricting the definition of Γ. In Section 4.1, I demonstrate how to construct bounds, building on Honore and Tamer (2006) and Chernozhukov et al. (2013). The construction is discussed in more detail in Appendix C in the supplementary material. Similarly to Chernozhukov et al. (2013), the method approximates the support of θ with a large number of support points.

In the remainder of this section I provide examples to gain some insight about identification in this model. For each of the models I impose the exclusion restriction of Assumption 2.5

[3] 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 $CASF(x,t) = \sum_{k=0}^{d} a_{jk}(x)\psi_k(t)$. 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 $CASF^*(x,t) := CASF(x,t) + \lambda(x)e_j(t)$ where $e_j(t) = \psi_{k^*}(t)$ for any $k > Jd$. But $|CASF(x,t) - CASF^*(x,t)| \geq |\lambda(x)|$. This works because of the orthogonality of the Legendre polynomials, which *requires* nonmonotonicity.

on all binary proxies, not just M_1. These additional restrictions are not required but improve the bounds for small J.

To provide examples to demonstrate the identified set I start by specifying a data generating process that can be used to derive the moments $m(x)$. I then use these moments to compute the set given by equation (2.5). First, $X = \mathbf{1}(s\Phi^{-1}(\theta) \geq \eta)$ where η is an independent standard normal, s is a constant that I set to either 0 or 1, and Φ is the standard normal distribution function. Then I use a probit model to generate a binary outcome and the binary proxies,

$$\begin{aligned} CASF(x,t) &= \Phi\left(\frac{\beta x + \mu_Y + \alpha_Y \Phi^{-1}(t)}{\sqrt{1-\alpha_Y^2}}\right) \\ p_j(x,t) &= \Phi\left(\frac{\mu + \alpha_Y \Phi^{-1}(t)}{\sqrt{1-\alpha_Y^2}}\right), \quad j = 1, \ldots, J \end{aligned}$$

where $\mu_Y = \Phi^{-1}(0.5)$ and $\mu = \Phi^{-1}(0.75)$ so that $E(M_j) = E(CASF(0,\theta)) = 0.5$ and $E(CASF(1,\theta)) = \Phi(\beta + \mu_Y)$.

I use $\alpha_2 = \alpha_3 = 0.99$ for each example and vary the values of s, β, and α_Y. I impose that $CASF(0,\cdot)$ and $CASF(1,\cdot)$ are weakly increasing as well as the item response functions, $p_j(x,\cdot)$. I compute bounds for $J = 1$ and $J = 2$ as well as the trivial case where $J = 0$. The results are displayed in Figures 1 and 2. Figure 1 displays bounds on the average treatment effects, $\Delta_{1,0}$, for six different models. Figure 2 displays pointwise bounds on the conditional treatment effects, $\Delta_{1,0}(t) := CASF(1,t) - CASF(0,t)$ for two of these models. Appendix C in the supplementary material contains further details on the computations.

It is apparent that the model can be far from identified. Even with J as small as 2 or 3, there can be identifying information, as the bounds are more narrow than the trivial bounds. It is also apparent that, at least in these examples, the bounds get smaller as J increases.

3 Large J identification and estimation

As the examples in the previous section demonstrate, the size of the identified set appears to decrease with J in most cases. In this section I provide conditions under which the identified set in fact shrinks to the true model as $J \to \infty$.

Throughout this section, I maintain Assumptions 2.1, 2.4, and 2.5. However, Assumption 2.3 is stronger than necessary when the number of proxies is large. Instead consider the following weak conditional dependence conditions.

Assumption 3.1.

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

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

(ii) For any $s \in (0, 1/2)$*, for each* J *there exists* $\mathcal{J}_Y^J(s) \subset \{1, \ldots, J\}$ *such that*

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

and $|\mathcal{J}_Y^J(s)| \to \infty$*.*

Condition (i) is a mixing condition on the sequence $M_1, \ldots, M_J$ conditional on (X, θ). Mixing conditions are a standard way to model (unconditional) dependence in time series data. 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. Condition (i) is natural in the 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. Also, if M_j represents the vote on the j^{th} bill in a session of Congress, then one would naturally expect some time series dependence conditional on legislator preferences. Condition (i) can also accomodate other types of conditional dependence such as the psychometric models of Jannarone (1997) and Bradlow et al. (1999).

Condition (ii) allows for various forms of dependence between Y and some of the binary proxies conditional on (X, θ). If Y is independent of only a subset of the proxies conditional on (X, θ) 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 $\mathcal{J}_Y^J(k)$ is $\{s : |j-s| > k\}$ so that $|\mathcal{J}_Y^J(\lfloor sJ \rfloor)| \geq J(1-2s) \to \infty$. Performance on a test may call on basic skills, represented by θ, 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). Lastly, note that the index set $\mathcal{J}_Y^J(k)$ in condition (ii) must be known. The discussion following the statement of the theorem below clarifies why this is important.

Both conditions in Assumption 3.1 could also be replaced by lower level conditions related to the structure in equations (2.1) and (2.2). For example, if (X, θ) is independent of $(\varepsilon_1, \ldots, \varepsilon_J)$ then condition (i) could be replaced by mixing condition on $\{\varepsilon_j : 1 \leq j \leq J\}$. And if U is independent of (X, θ) 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.

Assumption 2.6 also must be modified. In one sense this assumption is stronger than required when J is large because it is not necessary to require monotonicity for p_j for every j. But it is also *not* strong enough in that it only requires weak monotonicity. The model is assumed to satisfy the following assumption instead.

Assumption 3.2.

(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* $N_J := |\mathcal{J}_m^J| \to \infty$ *and, for each* $x \in \mathcal{X}$*,* $\sum_{j \in \mathcal{J}_m^J} p_j(x, \cdot)$ *is strictly increasing.*

In words, the assumption is 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 remaining items are strictly increasing *on average* and the number of remaining items is increasing in J. 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.

Define $\bar{p}(x, t) := N_J^{-1} \sum_{j \in \mathcal{J}_m^J} p_j(x, t)$. Under condition (ii) of Assumption 3.2, this function is strictly increasing in t for each x. Therefore, the inverse function $\bar{p}^{-1}(m; x)$ can be defined on the range of $\bar{p}(x, \cdot)$. That is, for each $x \in \mathcal{X}$ and each m in the range of $\bar{p}(x, \cdot)$ there is a unique t^*, which depends on x and m, such that $\bar{p}(x, t^*) = m$; I denote this t^* by $\bar{p}^{-1}(m; x)$. Likewise, under condition (i) of Assumption 3.2 $p_{j_0}^{-1}$ is uniquely defined on the range of p_{j_0}. Because p_{j_0} and $\bar{p}(x, \cdot)$ 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 $p_{j_0}^{-1}(u) = 1$. And for $m < \bar{p}(x, 0)$, let $\bar{p}^{-1}(m; x) = 0$ and for $m > \bar{p}(x, 1)$ let $\bar{p}^{-1}(m; x) = 1$.

Lastly I impose the following regularity conditions.

Assumption 3.3. *(i)* $CASF(x, t)$ *is continuous in* t *for each* $x \in \mathcal{X}$*, (ii)* p_{j_0} *is continuous, (iii)* $\bar{p}(x, t)$ *is continuous in* t *for each* J *and each* $x \in \mathcal{X}$*, and (iv) the quantile function,* $Q_{\theta|X}(\tau \mid x)$*, is defined for all* $\tau \in [0, 1]$ *and is uniformly continuous in* τ *for each* $x \in \mathcal{X}$*.*

Condition (i) here is the strongest as continuity on the interval $[0, 1]$ implies that the function is bounded. Condition (iv) implies that he support of $\theta \mid X = x$ is given by an interval, $[\underline{\theta}_x, \bar{\theta}_x] \subseteq [0,1]$. A slightly stronger version of these regularity conditions, stated as Assumption A.1 in Appendix A, is needed to to control the continuity of functions in the identified set as $J \to \infty$. Assumption 3.3 by itself does not prevent the limiting identified set from containing discontinuous functions, which would prevent identification in the limit.

3.1 Identification

Before stating the formal identification result, recall the definition of the identified set in the previous section in equation (2.5). Since Assumption 2.3 has now been replaced by Assumption 3.1, equation (2.3) no longer holds because the moments $m(x)$ are not uniquely determined by the triple $(CASF(\cdot,\cdot), \mathbf{p}(\cdot,\cdot), F_{\theta|X}(\cdot \mid \cdot))$ alone. Let γ^* denote the parameter that completes the model and for the remainder of this section let $\gamma = (CASF(\cdot,\cdot), \mathbf{p}(\cdot,\cdot), F_{\theta|X}(\cdot \mid \cdot), \gamma^*)$.[4] Then the moments $m(x)$ are uniquely characterized by the parameter γ and this mapping is given by $\tilde{\mu}(x;\gamma)$. The parameter space Γ is defined by Assumptions 2.4, 2.5, 3.1, and 3.2. There exists $\gamma_0 \in \Gamma$ such that $m(x) = \tilde{m}(x;\gamma_0)$. Then the identified set for $CASF(\cdot,\cdot)$ is given by

$$\begin{aligned}\mathcal{I}_J(\gamma_0) := \{CASF(\cdot,\cdot) : \exists \mathbf{p}(\cdot,\cdot), F_{\theta|X}(\cdot \mid \cdot), \gamma^* \text{ such that } (CASF(\cdot,\cdot), \mathbf{p}(\cdot,\cdot), F_{\theta|X}(\cdot \mid \cdot), \gamma^*) \in \Gamma \\ \text{and } \tilde{\mu}(x;\gamma_0) = \tilde{\mu}(x; (CASF(\cdot,\cdot), \mathbf{p}(\cdot,\cdot), F_{\theta|X}(\cdot \mid \cdot)), \gamma^*) \ \forall x \in \mathcal{X}\} \end{aligned} \tag{3.1}$$

The $CASF$ is said to be *large J identified* if

$$\lim_{J\to\infty} \sup_{\gamma_0 \in \Gamma} \sup_{CASF \in \mathcal{I}^J(\gamma_0)} ||CASF - CASF_0|| = 0 \tag{3.2}$$

where $||CASF - CASF_0|| = \sup_{x\in\mathcal{X}, t\in\Theta_0(x)} |||CASF(x,t) - CASF_0(x,t)|||$. The following theorem states the main identification result.

Theorem 3.1. *Under Assumptions 2.1, 2.4, 2.5, 3.1-3.3 and A.1, the CASF is large J identified.*

Proof. See Appendix A □

Remark 1: *In Williams (2013) a bound on the rate of convergence of the identified set is derived. The best possible convergence rate is bounded by $O(J^{-1/2+\epsilon})$ for all $\epsilon > 0$. This rate is obtained when the parameter space Γ is sufficiently smooth and $N_J^{-1} = O(J^{-1})$.*

[4] This γ^* specifies the nature of the conditional dependence among the M_j allowed by Assumption 3.1.

Remark 2: *Note that Assumption 3.2 rules out a model with*

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

for each j because in that case $\bar{p}(x,t)$ 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 3.1 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 $\bar{M} := J^{-1}\sum_{j=1}^{J} M_j$ 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.[5]

Let $\bar{M}$ and $\bar{p}$ be averages of M_j and p_j over $j \in \mathcal{J}_m^J$ as defined in Assumption 3.2. Lemma A.2 provides an exponential inequality (an extension of Hoeffding's inequality) under condition (ii) of Assumption 3.1, which implies for example that $\bar{M} - \bar{p}(x,\theta) \to_p 0$. Consider any $x \in \mathcal{X}$ and $t \in \Theta(x)$ and let $m = \bar{p}(x,t)$. This exponential inequality is also used in the proof of Theorem 3.1 to show that

$$\lim_{J\to\infty} E(Y \mid |\bar{M} - m| < r_J, X = x) - CASF(x,t) = 0 \tag{3.3}$$

for a sequence $r_J \to 0$. This is implicit in the result of Douglas (2001) though the observed covariates, X, are not present in that paper.[6]

Loosely, what (3.3) says is that

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

for large J. The left-hand side is observed while the right hand side is a combination of the object of interest and the unknown function $\bar{p}(x,t)$. Because $\bar{p}^{-1}$ varies with x, this does not isolate a causal effect of x on Y. However, by the same argument, and Assumption 2.5,

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

[5] 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.

[6] In addition, Douglas (2001) does not allow for weak conditional dependence, requires the response functions to have a derivative that is bounded and bounded away from 0, and assumes that the response functions equal 0 and 1 when $\theta = 0$ and $\theta = 1$, respectively. He also does not incorporate the covariates X.

for large J, using Assumption 2.5. From $CASF(x, \bar{p}^{-1}(m;x))$ and $p_{j_0}(\bar{p}^{-1}(m;x))$, if p_{j_0} is invertible, I can obtain $CASF(x, p_{j_0}^{-1}(m))$ for m in the range of p_{j_0}. This *does* identify a causal effect of x in the sense that this function only varies with x if $CASF(x,t)$ varies with x for some t. In other words, $CASF(x', p_{j_0}^{-1}(m)) - CASF(x, p_{j_0}^{-1}(m))$ represents a treatment effect at *some* location in the distribution of θ.

Next, because $\bar{M} \approx \bar{p}(X,\theta)$, it is also the case that $p_{j_0}(\bar{p}^{-1}(\bar{M};X)) \approx p_{j_0}(\theta)$. But $Pr(p_{j_0}(\theta) \leq m) = p_{j_0}^{-1}(m)$ by Assumptions 2.4 and 3.2. So $p_{j_0}(t)$ is approximately given by the t^{th} quantile of the distribution of $p_{j_0}(\bar{p}^{-1}(\bar{M};X))$ for large J. That is, $Q_{p_{j_0}(\bar{p}^{-1}(\bar{M};X))}(\tau) \approx p_{j_0}(\tau)$. Finally, if $CASF(x, p_{j_0}^{-1}(m))$ and $p_{j_0}(t)$ are known then $CASF(x,t)$ is as well.

This heuristic argument also suggests two corollaries to Theorem 3.1. First, under the common support condition, Assumption 2.2, $ASF(x)$ is identified without Assumption 2.4. That is, to identify the average (across the distribution of θ) structural function it is not necessary to normalize the distribution of θ. Second, if $E(\bar{M} \mid X, \theta) = E(\bar{M} \mid \theta)$ then the exclusion restriction (Assumption 2.5) is not required to identify $CASF(x,t)$ or $ASF(x)$.

Corollary 3.1. *Suppose $|Y| \leq \bar{Y}$ and θ has an absolutely continuous distribution function F_θ. Under Assumptions 2.1, 2.2, 2.5, 3.1-3.3 and A.1, the ASF is large J identified.*

Corollary 3.2. *If Assumption 2.5 is satisfied for all j then the CASF is large J identified under Assumptions 2.1, 2.4, 3.1-3.3 and A.1.*

Proofs of these results are in Appendix A.

3.2 Estimation

Consider an *i.i.d.* sample $(Y_i, X_i, M_{i,1}, \ldots, M_{i,J}), i = 1, \ldots, n$ 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 θ_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 3.1. As $J \to \infty$, the conditional probability function, $Pr(M_{i,j_0} = 1 \mid \bar{M}_i = m, X_i = x)$ converges to the function $q(x,m) := p_{j_0}(\bar{p}^{-1}(x,m))$ where $\bar{M}_i$ is the average of M_j over $j \in \mathcal{J}_m^J$ as defined in Assumption 3.2. Further, $q(X_i, \bar{M}_i) \approx p_{j_0}(\theta_i)$ for large J and, since $\theta_i \sim Uniform(0,1)$, $F_{q(X,\bar{M})}(q(X_i, \bar{M}_i)) \approx \theta_i$ where $F_{q(X,\bar{M})}$ denotes the distribution function of the random variable $q(X_i, \bar{M}_i)$. This argument suggests the estimator

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

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

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

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

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

To simplify notation let $h(x,t) := CASF(x,t)$. My proposed estimator for $h(x,t)$ is

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

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 3.2. *Under the assumptions maintained in Theorem 3.1 and Assumptions B.1-B.6 stated in Appendix B in the supplementary material,*

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

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

Proof. See Appendix B in the supplementary material. □

3.3 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 $\tilde{\theta}_i \mid X_i = x \sim N(x - .5, 1)$. 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 2 with $\theta_i = F_{\tilde{\theta}}(\tilde{\theta}_i)$.

In the simulations I calculate the estimator proposed above in Section 3.2, $\hat{h}(x, t)$, for $x = 0, 1$ and $t \in \mathcal{T} = \{.05, .1, \ldots, .95\}$. Since $\theta \sim Uniform(0, 1)$, $\frac{1}{10} \sum_{t \in \mathcal{T}} \hat{h}(1, t) - \hat{h}(0, t)$ provides an approximation of the $ATE = ASF(1) - ASF(0)$. 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 θ, 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 $\bar{M}_i$. 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 θ_i directly.

Overall the results suggest a substantial improvement over methods that do not properly control for θ_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.

4 Illustrations

4.1 Civic return to education

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 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 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),

$$Voted_i = g(SomeCollege_i, \theta_i, U_i)$$

where θ_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 θ_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 2.5) 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 2 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 estimates of estimated bounds on the ATE.

To estimate the bounds, I discretize the support of θ_i so that it has M support points. For each c and $\mathcal{J}$, replace $\mu_{c,\mathcal{J}}(x;\gamma)$ in equation (2.4) with

$$\mu^M_{c,\mathcal{J}}(x;\gamma^M) = \sum_{s=1}^{M} CASF(x,t_s)^c p_{\mathcal{J}}(x,t_s)\pi^M(x)$$

Then for each $a \in [-1, 1]$, I compute

$$Q^M(a) = min_{\gamma^M \in \Gamma^M} \sum_{c,\mathcal{J},x} (\mu^M_{c,\mathcal{J}}(x; \gamma^M) - m_{c,\mathcal{J}}(x))^2$$

$$\text{s.t.} \sum_{s=1}^{M} CASF(1, t_s) - CASF(0, t_s) = a$$

The estimated bounds are the interval $[a_{min}, a_{max}] = argmin_{a \in [-1,1]} Q^M(a)$. This procedure is discussed in the supplementary material in further detail. There I show that the interval produced by this procedure is the identified set in the limit as $M \to \infty$. This is challenging computationally as it requires minimizing a function of $2M$ parameters. The problem cannot be recast as a linear programming problem. For these reasons I use $M = 10$ for the results presented here.[7]

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 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 3 shows bounds on the conditional ATE. These bounds are not informative for model 1 and narrow only slightly for model 2. The bounds for the conditional ATE in model 3 are substantially more narrow. The lower bound for the conditional ATE at points above 0.5 is near 0.

4.2 Education, ability, and wage

In this section I use data from the National Longitudinal Survey of Youth (NLSY) to construct estimates of ability that separate out the effect of education. 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 only the arithmetic reasoning subcomponent of the Armed Forces Qualifying Test (AFQT), a test which consists of the verbal and math components of the ASVAB.[8] The method proposed in this paper is an alternative to the methods proposed by Hansen et al. (2004). For comparison purposes I use the same subsample of the NLSY used by Hansen et al. (2004), which consists of $1,927$ white non-Hispanic males. Hansen et al. (2004). Hansen et al. (2004) find that each additional year of education increases composite AFQT scores by $2 - 4\%$. They also find that these

[7] For the model with only one proxy I also computed bounds with $M = 20$. In that case the bounds were slightly wider.

[8] This is typically the component of the ASVAB that matters the most for wages.

effects are slightly larger for lower ability levels.

The arithmetic reasoning component of the ASVAB consists of 30 questions. Item-level responses, coded as correct or incorrect, have recently been made available. See Schofield (2014) for an early analysis of the item-level data. In my analysis, I assume that the second item satisfies Assumption 2.5. I then separately treat each item as the outcome Y while using the 29 other items as proxies to estimate the effect of education on responses to these questions. Figures 4 and 5 show a selection of the estimated item response functions, along with confidence bands computed via 100 bootstrap samples.

Next, in Figure 6, I show the CASF estimates for wages. I restrict the sample to $1,018$ full-time workers with non-missing average weekly wages. For comparison, the left panel shows estimates that use simply the percentile score. Note that if ability is not controlled for at the average difference in log wages is x.

Table 3 reports estimates of the average treatment effect for wages and for the individual test items. As in Hansen et al. (2004), I find that there is a substantial education effect on test scores. The effects found here are likely bigger because, unlike Hansen et al. (2004), I do not control for family background or demographic characteristics. The results here also demonstrate how the nature of the education effect can differ across items. For some items the effect is larger at lower ability levels and for some items the effect is larger at higher ability levels.

5 Extensions of the model

The basic model of Sections 2 and 3 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 θ to be $Uniform(0,1)$. Second I discuss alternatives to the exclusion restriction in Assumption 2.5. Williams (2013) considers some additional extensions of the model.

5.1 Alternative normalizations

Suppose the distribution of θ is not normalized, as imposed by Assumption 2.4 and instead assume that, for some $x_0 \in \mathcal{X}$ and some $j_1 \in \{1,\ldots,J\}$, $p_{j_1}(x_0,t) = \pi(t)$ where π is a known function. Recall from the discussion following the statement of Theorem 3.1 that for large J,

$$E(Y \mid \bar{M} \approx m, X = x) \approx CASF(x, \bar{p}^{-1}(m;x)) \qquad (5.1)$$

and

$$Pr(M_{j_0} = 1 \mid \bar{M} \approx m, X = x) \approx p_{j_0}(\bar{p}^{-1}(m; x)) \tag{5.2}$$

Likewise, under this alternative normalization,

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

Since π is known this implies that $\bar{p}^{-1}(m; x_0)$ is approximately identified for large J. Then, applying (5.2) for $x = x_0$, $p_{j_0}(t)$ is identified as well. Applying (5.2) again for any other value of x produces $\bar{p}^{-1}(m; x)$. And applying (5.1), $CASF(x, t)$ 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 3.1 was formalized in the proof of the theorem.

Suppose θ 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 θ. Then the distribution of θ 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 θ. 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.

5.2 Alternative restrictions

Next I consider two restrictions that can be used in place of Assumption 2.5.

5.2.1 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,

$$Pr(M_{j_0} = 1 \mid \bar{M} \approx m, X = x) \approx p_{j_0}(x_1, \bar{p}^{-1}(m; x)) \tag{5.3}$$

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

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

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 $\bar{p}^{-1}(m; x)$ is identified and hence $CASF$ 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 (θ) 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 θ 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.

5.2.2 Linking exclusion restrictions

According to Assumption 2.5 one item, M_{j_0}, must be independent of X conditional on θ. In some cases however, each item may be dependent on *some* components of X conditional on θ. 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 θ.

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,

$$Pr(M_{j_1} = 1 \mid \bar{M} \approx m, X = x) \approx p_{j_1}(x_1, \bar{p}^{-1}(m; x)) \tag{5.4}$$

and

$$Pr(M_{j_2} = 1 \mid \bar{M} \approx m, X = x) \approx p_{j_2}(x_2, \bar{p}^{-1}(m; x)) \tag{5.5}$$

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,

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

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, $\bar{p}^{-1}(m; x)$ 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 θ 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).

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 results, particularly those from the civic returns application in Section 4.1, 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.

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.

Bradlow, E. T., H. Wainer, and X. Wang (1999): "A Bayesian random effects model for testlets," *Psychometrika*, 64, 153–168.

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.

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 Jr (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.

Jannarone, R. J. (1997): "Models for locally dependent responses: Conjunctive item response theory," in *Handbook of modern item response theory*, Springer, 465–479.

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.

Lewbel, A. (1998): "Semiparametric latent variable model estimation with endogenous or mismeasured regressors," *Econometrica*, 105–121.

——— (2014): "An Overview of the Special Regressor Method," in *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, ed. by J. S. Racine, L. Su, A. Ullah, and A. Lewbel, Oxford University Press.

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.

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

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

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

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*, Springer Science & Business Media.

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.

Appendix

This appendix contains proofs of the main identification results – Theorems 3.1 and its corollaries. A separate supplementary appendix provides additional results on computation of the identified set and the proof of Theorem 3.2.

A Identification Proofs

This section presents the proof of Theorem 3.1. First, I formally state the restrictions imposed on the parameter space. Then I provide two lemmas. Then in A.3 I provide the proof of the theorem.

A.1 Uniform equicontinuity and modulus of continuity

For each J, let Γ_J denote the parameter space. It is assumed that Γ_J is defined so that Assumptions 2.4, 2.5, 3.1, and 3.2, are each satisfied for each J. To state the regularity condition precisely, for any J and any $\gamma_J \in \Gamma_J$ I will use the notation $CASF_{\gamma_J}$ to emphasize that this is the $CASF$ associated with the parameter γ_J, $\bar{p}_{\gamma_J}$ to emphasize that this is the $\bar{p}$ associated with the parameter γ_J, etc.

Assumption A.1. *The sequence of parameter spaces $\{\Gamma_J : J \geq 1\}$ are defined so that each of the following classes of functions is uniformly equicontinuous:*

(i) $\{CASF_{\gamma_J}(x, \cdot) : [0,1] \to \mathbb{R}; x \in \mathcal{X}, \gamma_J \in \Gamma_J, J \geq 1\}$

(ii) $\{\bar{p}_{\gamma_J}(x, \cdot) : [0,1] \to [0,1]; x \in \mathcal{X}, \gamma_J \in \Gamma_J, J \geq 1\}$

(iii) $\{\bar{p}^{-1}_{\gamma_J}(\cdot; x) : [0,1] \to [0,1]; x \in \mathcal{X}, \gamma_J \in \Gamma_J, J \geq 1\}$

(iv) $\{p_{j0,\gamma_J}(\cdot) : [0,1] \to [0,1]; \gamma_J \in \Gamma_J, J \geq 1\}$

(v) $\{p^{-1}_{j0,\gamma_J}(\cdot) : [0,1] \to [0,1]; \gamma_J \in \Gamma_J, J \geq 1\}$

(vi) $\{Q_{\theta|X;\gamma_J}(\cdot \mid x) : [0,1] \to [0,1]; x \in \mathcal{X}, \gamma_J \in \Gamma_J, J \geq 1\}$

This means, for example, that for all $\varepsilon > 0$ there exists $\delta > 0$ such that for all J, all $\gamma_J \in \Gamma_J$, all $x \in \mathcal{X}$ and any pair $t, t' \in [0,1]$, if $|t' - t| < \delta$ then $|\bar{p}(x,t') - \bar{p}(x,t)| < \varepsilon$. In the proof of Theorem 3.1, I make use of a convenient equivalent definition of uniform equicontinuity. Uniform equicontinuity of a class of functions, $\mathcal{H}$, on a space $\mathcal{Z}$, is equivalent to the existence of a real-valued function, $c_{\mathcal{H}}$, satisfying $\lim_{s\to 0} c_{\mathcal{H}}(s) = c_{\mathcal{H}}(0) = 0$, such that for every $h \in \mathcal{H}$ and every $z, z' \in \mathcal{Z}$, $|h(z') - h(z)| \leq c_{\mathcal{H}}(|z' - z|)$. The function $c_{\mathcal{H}}$ is called the modulus of continuity. Furthermore, $c_{\mathcal{H}}$ can be chosen to be monotonically increasing. Assumption A.1 supplies us with six the moduli of continuity which I will denote, respectively, c_C, $c_{\bar{p}}$, $c_{\bar{p}^{-1}}$, $c_{p_{j_0}}$, $c_{p_{j_0}^{-1}}$, and c_Q.

A.2 Lemmas

Lemma A.1. *(Azuma's inequality) Suppose $\mathcal{F}_j$ is a filtration and Z_j is a martingale difference with respect to $\mathcal{F}_j$. In addition, suppose $c_{1,J}, \ldots, c_{J,J}$ and $d_{1,J}, \ldots, d_{J,J}$ are constants such that $|c_{j,J} Z_j| \leq d_{j,J}$. Then*

$$Pr\left(\left|\sum_{j=1}^{J} c_{j,J} Z_j\right| > \varepsilon\right) \leq 2\exp\left(-\frac{1}{2}\varepsilon^2 / \sum_{j=1}^{J} d_{j,J}\right)$$

Lemma A.2. *Suppose $c_{1,J}, \ldots, c_{J,J}$ and $d_{1,J}, \ldots, d_{J,J}$ are constants such that $|c_{j,J}| \leq d_{j,J}$. Let $c_J = \sum_{j=1}^{J} c_{j,J}$ and $d_J = \sum_{j=1}^{J} d_{j,J}$. Under Assumption 3.1,*

$$Pr\left(c_J^{-1}|\sum_{j=1}^{J} c_{j,J}(M_j - E(M_j \mid X, \theta)| > \varepsilon\right) \leq 2k_J \exp\left(-\frac{1}{8}\frac{c_J^2 \varepsilon^2}{k_J d_J}\right) + \frac{2 d_J \alpha_{k_J}}{\varepsilon c_J}$$

Proof. First, let $\mathcal{F}_j$ denote the sigma algebra generated by $\{X, \theta, M_1, \ldots, M_j\}$ for $j \geq 1$ and let $\mathcal{F}$ denote the sigma algebra generate by $\{X, \theta\}$ for $j < 1$. Then

$$\begin{aligned}
&|M_j - E(M_j \mid X, \theta)| \\
&\leq \sum_{s=0}^{k_J - 1} |E(M_j \mid \mathcal{F}_{j-s}) - E(M_j \mid \mathcal{F}_{j-s-1})| + |E(M_j \mid \mathcal{F}_{j-k_J}) - E(M_j \mid X, \theta)|
\end{aligned}$$

so

$$\begin{aligned}
&Pr(c_J^{-1}|\sum_{j=1}^{J} c_{j,J}(M_j - E(M_j \mid X, \theta)| > \varepsilon) \qquad (A.1) \\
&\leq \sum_{s=0}^{k_J-1} Pr(c_J^{-1}|\sum_{j=1}^{J} c_{j,J}(E(M_j \mid \mathcal{F}_{j-s}) - E(M_j \mid \mathcal{F}_{j-s-1}))| > \varepsilon/(2k_J)) \\
&+ Pr(c_J^{-1}|\sum_{j=1}^{J} c_{j,J}(E(M_j \mid \mathcal{F}_{j-k_J}) - E(M_j \mid X, \theta))| > \varepsilon/2)
\end{aligned}$$

Clearly $E(M_j \mid \mathcal{F}_{j-s}) - E(M_j \mid \mathcal{F}_{j-s-1})$ is a martingale difference with respect to the filtration $\mathcal{F}_{j-s}$. Therefore, applying Azuma's inequality, the first term is bounded by

$$2k_J \exp\left(-\frac{1}{8}\frac{c_J^2\varepsilon^2}{k_J^2 d_J}\right) \tag{A.2}$$

Applying Markov's inequality to the second term of (A.1),

$$\begin{aligned} & Pr(c_J^{-1}|\sum_{j=1}^{J} c_{j,J}(E(M_j \mid \mathcal{F}_{j-k_J}) - E(M_j \mid X,\theta))| > \varepsilon/2) \\ \leq & \frac{2}{c_J\varepsilon}\sum_{j=1}^{J}|c_{j,J}|E|E(M_j \mid \mathcal{F}_{j-k_J}) - E(M_j \mid X,\theta))| \\ \leq & \frac{2}{c_J\varepsilon}\sum_{j=1}^{J} d_{j,J}\alpha_{j,k_J} \end{aligned} \tag{A.3}$$

By Assumption 3.1, $\sup_j \alpha_{j,k_J} \leq \alpha_{k_J}$. The desired result follows by combining (A.1)-(A.3). □

Lemma A.3. *For any constant $A > 0$, there exist sequences $\{k_N : N \geq 1\}$ and $\{r_N : N \geq 1\}$ such that $r_N \to 0$,*

$$\frac{\alpha_{k_N}}{r_N c_Q^{-1}(c_{\overline{p}}^{-1}(r_N))} \to 0$$

and

$$\frac{k_N}{c_Q^{-1}(c_{\overline{p}}^{-1}(r_N))}\exp\left(-A\frac{Nr_N^2}{k_N}\right) \to 0$$

Proof. First, let $\alpha(x) := \alpha_{\lfloor x \rfloor}$ for any positive real number x. I can assume without loss of generality that $\alpha(x)$ is a decreasing function. It is apparent that because $\alpha_k \to 0$ as $k \to \infty$ that there exists a sequence of positive numbers $x_N \to 0$ such that $\alpha(A^{1/2}N^{1/2}x_N) \leq x_N$ for all sufficiently large N. Because $c^*(r) := c_Q^{-1}(c_{\overline{p}}^{-1}(r))$ I can take r_N such that $r_N c^*(r_N) = x_N^{1/2}$. The condition that $r_N \to 0$ is met. Also, let $\xi_N = x_N^{1/2}$. Then

$$\alpha(A^{1/2}N^{1/2}r_N c^*(r_N)^{1/2}\xi_N) \leq \alpha(A^{1/2}N^{1/2}r_N c^*(r_N)\xi_N) \leq r_N c^*(r_N)\xi_N$$

Take $k_N = A^{1/2}N^{1/2}r_N c^*(r_N)^{1/2}\xi_N$. Then, $\frac{\alpha(k_N)}{r_N c^*(r_N)} \leq \xi_N \to 0$, as desired. And, for sufficiently large N,

$$k_N log \left(\frac{k_N}{c^*(r_N)\xi_N^2} \right) \leq \left(\frac{k_N^2}{c^*(r_N)\xi_N^2} \right) \leq ANr_N^2$$

Rearranging this inequality,

$$\frac{k_N}{c^*(r_N)} \exp \left(-A\frac{Nr_N^2}{k_N} \right) \leq \xi_N^2 \to 0$$

as desired. □

A.3 Proof of Theorem 3.1

Some preliminaries Though it is suppressed in the notation in the text, the parameter space should be indexed by J. For each J, consider an arbitrary $\gamma_{0J} \in \Gamma_J$ and $CASF_J \in \mathcal{I}^J(\gamma_{0J})$. To simplify the notation however I continue to suppress the J subscript. Let $\gamma_0 = (CASF^0, \mathbf{p}^0, F^0_{\theta|X}, \gamma_0^*)$. And let $\gamma \in \Gamma$ be such that $\gamma = (CASF, \mathbf{p}, F_{\theta|X}, \gamma^*)$ and $m(x) = \tilde{\mu}(x; \gamma_0) = \tilde{\mu}(x; \gamma)$ for all $x \in \mathcal{X}$.

I wish to show that there is a sequence such that $||CASF - CASF_0|| \leq a_J$ for all J and $a_J \to 0$ where a_J does not depend on γ_0 or γ. Since $||CASF - CASF_0|| = \sup_{x\in\mathcal{X}, t\in\Theta_0(x)} |CASF(x,t) - CASF_0(x,t)|$ it is sufficient to show that $|CASF(x,t) - CASF_0(x,t)| \leq a_J$ for all $x \in \mathcal{X}$ and $t \in \Theta_0(x)$.

Let $\tilde{k}_J = \alpha J$ for some $0 < \alpha < (1 - s_0)/4$. Let $N_J := |\mathcal{J}^J_m(\tilde{k}_J)|$ where $\mathcal{J}^J_m(\tilde{k}_J)$ is defined in Assumption 3.2 and let

$$\bar{M} := \frac{1}{N_J} \sum_{j \in \mathcal{J}^J_m(\tilde{k}_J)} M_j = \frac{\sum_{j=1}^J c_{j,J} M_j}{\sum_{j=1}^J c_{j,J}} \tag{A.4}$$

where $c_{j,J} = \mathbf{1}(j \in \mathcal{J}^J_m(\tilde{k}_J))$. Note that $N_J \leq J - 2k_J = \frac{1-2\alpha}{\alpha}\tilde{k}_J$. And $N_J \geq J - 4k_J - s_0 J = (1 - 4\alpha - s_0))J \to \infty$.

Fix $x \in \mathcal{X}$ and $t_0 \in \Theta_0(x)$ and let $m_0 = \bar{p}^0(x, t_0)$. Let $A = 1/128$ and take the corresponding sequences $\{k_N : N \geq 1\}$ and $\{r_N : N \geq 1\}$ guaranteed by Lemma A.3. Let $r_J := r_{N_J}$ and $k_J := k_{N_J}$.

Step 1. I first show that $E(Y \mid |\bar{M} - m_0| \leq r_J \mid X = x)$ is close to $CASF_0(x, t_0)$.

First, $\bar{M}$ is measurable with respect to $\{M_j : j \in \mathcal{J}^J_Y\}$ so that $E(Y \mid |\bar{M} - m_0| \leq r_J, X =$

$x) = E(E(Y \mid X, \theta, \{M_j : j \in \mathcal{J}_Y^J\})) \mid |\bar{M} - m_0| \leq r_J, X = x)$. Therefore,

$$\begin{aligned}
&E(Y \mid |\bar{M} - m_0| \leq r_J, X = x) \\
&= E(E(Y \mid X, \theta) \mid |\bar{M} - m_0| \leq r_J, X = x) \qquad\qquad (\text{A.5}) \\
&+ E\left(E(Y \mid X, \theta, \{M_j : j \in \mathcal{J}_Y^J\}) - E(Y \mid X, \theta) \mid |\bar{M} - m_0| \leq r_J, X = x\right)
\end{aligned}$$

Note that for any random variable Z and event A, $E(|Z| \mid A) = (E(|Z|) - E(|Z| \mid A^c)(1 - Pr(A)))/Pr(A) \leq E(|Z|)/Pr(A)$. Therefore,

$$\begin{aligned}
&\left|E\left(E(Y \mid X, \theta, \{M_j : j \in \mathcal{J}_Y^J\}) - E(Y \mid X, \theta) \mid |\bar{M} - m_0| \leq r_J, X = x\right)\right| \\
&\leq \frac{E\left|E(Y \mid X, \theta, \{M_j : j \in \mathcal{J}_Y^J\}) - E(Y \mid X, \theta)\right|}{Pr(|\bar{M} - m_0| \leq r_J, X = x)} \\
&\leq \frac{\alpha_{\tilde{k}_J}}{Pr(|\bar{M} - m_0| \leq r_J, X = x)} \qquad\qquad (\text{A.6})
\end{aligned}$$

where the second inequality follows from Assumption 3.1 and $\tilde{k}_J$ is defined in Assumption 3.2.

Next, by Assumption 2.1(i), $E(Y \mid X = x, \theta = \tau) = CASF_0(x, \tau)$ and therefore

$$\begin{aligned}
&|E(E(Y \mid X, \theta) \mid |\bar{M} - m_0| \leq r_J, X = x) - CASF_0(x, t_0)| \\
&= \left|\int (CASF_0(x, \tau) - CASF_0(x, t_0))\, dF^0_{\theta||\bar{M}-m_0|\leq r_J, X=x}(\tau)\right| \\
&\leq \int_{\tau:|\bar{p}^0(x,\tau)-\bar{p}^0(x,t_0)|\leq 3r_J} |CASF_0(x, \tau) - CASF_0(x, t_0)| dF^0_{\theta||\bar{M}-m_0|\leq r_J, X=x}(\tau) \\
&+ \int_{\tau:|\bar{p}^0(x,\tau)-\bar{p}^0(x,t_0)|> 3r_J} |CASF_0(x, \tau) - CASF_0(x, t_0)| dF^0_{\theta||\bar{M}-m_0|\leq r_J, X=x}(\tau) \\
&\leq c_C(c_{\bar{p}^{-1}}(3r_J)) + 2 \sup_{x \in \mathcal{X}, t \in [0,1]} |CASF_0(x, t)| \qquad\qquad (\text{A.7}) \\
&\qquad \times Pr(|\bar{p}^0(x, \theta) - \bar{p}^0(x, t_0)| > 3r_J \mid |\bar{M} - m_0| \leq r_J, X = x)
\end{aligned}$$

The first term in the final line follows because $|CASF_0(x, \tau) - CASF_0(x, t_0)| \leq c_C(|\tau - t_0|)$ and because

$$\begin{aligned}
|\tau - t_0| &= |\bar{p}_0^{-1}(\bar{p}_0(x, \tau); x) - \bar{p}_0^{-1}(\bar{p}_0(x, t_0); x)| \\
&\leq c_{\bar{p}^{-1}}(|\bar{p}_0(x, \tau) - \bar{p}_0(x, t_0)|)
\end{aligned}$$

The second term in the final line of (A.7) follows because $CASF_0(x, \cdot)$ 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 $\sup_{x \in \mathcal{X}, t \in [0,1]} |CASF_0(x, t)| \leq B/2$.

Next, if $|\bar{p}^0(x,\theta) - \bar{p}^0(x,t_0)| > 3r_J$ and $|\bar{M} - m_0| \leq r_J$ then $|\bar{M} - \bar{p}^0(x,\theta)| > r_J$. So

$$\begin{aligned} & Pr(|\bar{p}^0(x,\theta) - \bar{p}^0(x,t_0)| > 3r_J \mid |\bar{M} - m_0| \leq r_J, X = x) \\ & \leq Pr(|\bar{M} - \bar{p}^0(x,\theta)| > r_J \mid |\bar{M} - m_0| \leq r_J, X = x) \end{aligned} \tag{A.8}$$

But

$$\begin{aligned} & Pr(|\bar{M} - \bar{p}^0(x,\theta)| > r_J \mid |\bar{M} - m_0| \leq r_J, X = x) \\ & = \frac{Pr(|\bar{M} - \bar{p}^0(x,\theta)| > r_J \text{ and } |\bar{M} - m_0| \leq r_J \mid X = x)}{Pr(|\bar{M} - m_0| \leq r_J \mid X = x)} \\ & \leq \frac{Pr(|\bar{M} - \bar{p}^0(x,\theta)| > r_J \mid X = x)}{Pr(|\bar{M} - m_0| \leq r_J \mid X = x)} \end{aligned} \tag{A.9}$$

Using the definition of $\bar{M}$ in equation (A.4) and applying Lemma A.2,

$$Pr(|\bar{M} - \bar{p}^0(x,\theta)| > r_J \mid X = x) \leq 2k_J \exp(-\frac{1}{8}\frac{N_J r_J^2}{k_J}) + 2\frac{\alpha_{k_J}}{r_J}$$

Combining this with equations (A.5)-(A.9),

$$\begin{aligned} & E(Y \mid |\bar{M} - m_0| \leq r_J, X = x) \\ & \leq \frac{\alpha_{\tilde{k}_J} + 2Bk_J \exp(-\frac{1}{8}\frac{N_J r_J^2}{k_J}) + 2B\frac{\alpha_{k_J}}{r_J}}{Pr(|\bar{M} - m_0| \leq r_J, X = x)} + c_C(c_{\bar{p}^{-1}}(3r_J)) \end{aligned} \tag{A.10}$$

Lastly,

$$\begin{aligned} & Pr(|\bar{M} - m_0| \leq r_J \mid X = x) = Pr(|\bar{M} - \bar{p}^0(x,\theta)| \leq r_J \mid X = x) \\ & \geq \int Pr(|\bar{M} - \bar{p}_0(x,t_0)| < r_J \mid X = x, \theta = \tau) dF^0_{\theta|X=x}(\tau) d\tau \\ & \geq \int_{\tau : |\bar{p}_0(x,\tau) - \bar{p}_0(x,t_0)| < r_J/2} Pr(|\bar{M} - \bar{p}_0(x,t_0)| < r_J \mid X = x, \theta = \tau) dF^0_{\theta|X=x}(\tau) d\tau \end{aligned}$$

and

$$\begin{aligned} & Pr(|\bar{M} - \bar{p}_0(x,t_0)| < r_J \mid X = x, \theta = \tau) \\ & = 1 - Pr(|\bar{M} - \bar{p}_0(x,t_0)| \geq r_J \mid X = x, \theta = \tau) \\ & \geq 1 - Pr(|\bar{M} - \bar{p}_0(x,\tau)| \geq r_J/2 \mid X = x, \theta = \tau) \end{aligned}$$

where the inequality follows for τ such that $|\bar{p}_0(x,\tau) - \bar{p}_0(x,t_0)| < r_J/2$ because $|\bar{M} -$

$\bar{p}_0(x,\tau)| \geq |\bar{M} - \bar{p}_0(x,t_0)| - |\bar{p}_0(x,\tau) - \bar{p}_0(x,t_0)| \geq r_J/2$. So that

$$\begin{aligned}
&Pr(|\bar{M} - m_0| < r_J \mid X = x) && \text{(A.11)}\\
&\geq \int_{\tau: |\bar{p}_0(x,\tau) - \bar{p}_0(x,t_0)| < r_J/2} \big(1 - Pr(|\bar{M} - \bar{p}_0(x,\tau)| \geq r_J/2 \mid X = x, \theta = \tau)\big)\, dF^0_{\theta|X=x}(\tau)\\
&\geq Pr(|\bar{p}_0(x,\theta) - \bar{p}_0(x,t_0)| < r_J/2 \mid X = x) - Pr(|\bar{M} - \bar{p}_0(x,\tau)| \geq r_J/2 \mid X = x)
\end{aligned}$$

But by Assumption 3.3

$$\begin{aligned}
&Pr(|\bar{p}_0(x,\theta) - \bar{p}_0(x,t_0)| < r_J/2 \mid X = x)\\
&= F^0_{\theta|X=x}(\bar{p}_0^{-1}(\bar{p}_0(x,t_0) + r_J/2; x)) - F^0_{\theta|X=x}(\bar{p}_0^{-1}(\bar{p}_0(x,t_0) - r_J/2; x))\\
&\geq c_Q^{-1}(c_{\bar{p}}^{-1}(r_J)) && \text{(A.12)}
\end{aligned}$$

and applying Lemma A.2,

$$\begin{aligned}
Pr(|\bar{M} - \bar{p}^0(x,\theta)| > r_J/2 \mid X = x) &\leq 2k_J \exp(-\frac{1}{32}\frac{N_J r_J^2}{k_J}) + 4\frac{\alpha_{k_J}}{r_J} && \text{(A.13)}\\
&\leq \varepsilon c_Q^{-1}(c_{\bar{p}}^{-1}(r_J)) && \text{(A.14)}
\end{aligned}$$

where the second line holds for any $\epsilon > 0$, for all sufficiently large J by Lemma A.3, and because $N_J \to \infty$ as $J \to \infty$.

Finally, combining (A.10)-(A.13),

$$\begin{aligned}
&|E(Y \mid |\bar{M} - m_0| \leq r_J, X = x) - CASF_0(x,t_0)|\\
&\leq \frac{\alpha_{\tilde{k}_J} + 2Bk_J \exp(-\frac{1}{8}\frac{N_J r_J^2}{k_J}) + 2B\frac{\alpha_{k_J}}{r_J}}{c_Q^{-1}(c_{\bar{p}}^{-1}(r_J)) - 2k_J \exp(-\frac{1}{32}\frac{N_J r_J^2}{k_J}) - 4\frac{\alpha_{k_J}}{r_J}} + c_C(c_{\bar{p}^{-1}}(3r_J)) && \text{(A.15)}\\
&\leq ((1-\varepsilon)c_Q^{-1}(c_{\bar{p}}^{-1}(r_J)))^{-1}\left(\alpha_{\tilde{k}_J} + 2Bk_J \exp(-\frac{1}{8}\frac{N_J r_J^2}{k_J}) + 2B\frac{\alpha_{k_J}}{r_J}\right) + c_C(c_{\bar{p}^{-1}}(3r_J))
\end{aligned}$$

Step 2. Next, using many of the same calculations I can show that $\exists t \in \Theta(x)$ such that $m := \bar{p}(x,t)$ is close to m_0.

First, by (A.11)-(A.13) above,

$$\begin{aligned}
&Pr(|\bar{M} - \bar{p}_0(x,t_0)| < r_J/2 \mid X = x)\\
&\geq c_Q^{-1}(c_{\bar{p}}^{-1}(r_J/2)) - 2k_J \exp(-\frac{1}{128}\frac{N_J r_J^2}{k_J}) - 8\frac{\alpha_{k_J}}{r_J} && \text{(A.16)}\\
&\geq (1-\varepsilon)c_Q^{-1}(c_{\bar{p}}^{-1}(r_J))
\end{aligned}$$

for any $\varepsilon > 0$ and sufficiently large J. On the other hand, if $|\bar{M} - \bar{p}^0(x, t_0)| \leq r_J/2$ then for any τ either $|\bar{p}(x,\tau) - \bar{p}^0(x,t_0)| < r_J$ or $|\bar{M} - \bar{p}(x,\tau)| \geq r_J/2$. Therefore,

$$\begin{aligned}
&Pr(|\bar{M} - \bar{p}_0(x,t_0)| < r_J/2 \mid X = x) \\
&\leq Pr(|\bar{p}(x,\theta) - \bar{p}^0(x,t_0)| < r_J \mid X = x) + Pr(|\bar{M} - \bar{p}(x,\theta)| \geq r_J/2 \mid X = x)
\end{aligned}$$

The first term here is equal to 0 if $|\bar{p}(x,\tau) - \bar{p}^0(x,t_0)| > r_J$ for all $\tau \in \Theta(x)$. And applying Lemma A.2 and Lemma A.3 again

$$Pr(|\bar{M} - \bar{p}^0(x,\theta)| > r_J/2 \mid X = x) \leq 2k_J \exp(-\frac{1}{32}\frac{N_J r_J^2}{k_J}) + 4\frac{\alpha_{k_J}}{r_J} \leq \varepsilon c_Q^{-1}(c_{\bar{p}}^{-1}(r_J))$$

for any $\varepsilon > 0$ and sufficiently large J. Therefore, if $|\bar{p}(x,\tau) - \bar{p}^0(x,t_0)| > r_J$ for all $\tau \in \Theta(x)$ then

$$(1-\varepsilon)c_Q^{-1}(c_{\bar{p}}^{-1}(r_J)) \leq Pr(|\bar{M} - \bar{p}_0(x,t_0)| < r_J/2 \mid X = x) \leq \varepsilon c_Q^{-1}(c_{\bar{p}}^{-1}(r_J))$$

which is a contradiction if $\varepsilon < 1/2$.

I can conclude that for all sufficiently large J, $\exists t \in \Theta(x)$ such that $|m_0 - \bar{p}(x,t)| \leq r_J$.

Step 3. In Step 1 it was shown that

$$\begin{aligned}
&|E(Y \mid |\bar{M} - m_0| \leq r_J, X = x) - CASF_0(x,t_0)| \\
&\leq ((1-\varepsilon)c_Q^{-1}(c_{\bar{p}}^{-1}(r_J)))^{-1}\left(\alpha_{\tilde{k}_J} + 2Bk_J\exp(-\frac{1}{8}\frac{N_J r_J^2}{k_J}) + 2B\frac{\alpha_{k_J}}{r_J}\right) + c_C(c_{\bar{p}^{-1}}(3r_J))
\end{aligned}$$

Let $m = \bar{p}(x,t)$ where t is as defined in Step 2. The same argument from Step 1 can be used to show that

$$\begin{aligned}
&|E(Y \mid |\bar{M} - m_0| \leq r_J, X = x) - CASF(x,t)| \\
&\leq ((1-\varepsilon)c_Q^{-1}(c_{\bar{p}}^{-1}(r_J)))^{-1}\left(\alpha_{\tilde{k}_J} + 2Bk_J\exp(-\frac{1}{8}\frac{N_J r_J^2}{k_J}) + 2B\frac{\alpha_{k_J}}{r_J}\right) + c_C(c_{\bar{p}^{-1}}(3r_J))
\end{aligned} \tag{A.17}$$

The only place where the argument needs adjusted is in (A.8). If $|\bar{p}(x,\theta) - \bar{p}(x,t)| > 3r_J$ and $|\bar{M} - m_0| \leq r_J$ then

$$|\bar{M} - \bar{p}(x,\theta)| \geq |\bar{p}(x,\theta) - \bar{p}(x,t)| - |\bar{M} - m_0| - |m_0 - \bar{p}(x,t)| \geq r_J$$

since, as shown in Step 2, t can be chosen so that $|m_0 - \bar{p}(x,t)| < r_J$. Therefore

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

and the remainder of the proof follows as before.

Step 4. It remains to show that t is close to t_0.

First, $\bar{M}$ is measurable with respect to $\mathcal{J}^J_{M_{j_0}} := \{M_j : |j - j_0| > \tilde{k}_J\}$ so that $E(M_{j_0} \mid |\bar{M} - m_0| \leq r_J, X = x) = E(E(M_{j_0} \mid X, \theta, \{M_j : |j - j_0| > \tilde{k}_J\}) \mid |\bar{M} - m_0| \leq r_J, X = x)$. Therefore,

$$\begin{aligned}&E(M_{j_0} \mid |\bar{M} - m_0| \leq r_J, X = x)\\&= E(p_{j_0,0}(\theta) \mid |\bar{M} - m_0| \leq r_J, X = x)\\&+ E\left(E(M_{j_0} \mid X, \theta, \{M_j : |j - j_0| > \tilde{k}_J\})) - p_{j_0,0}(\theta) \mid |\bar{M} - m_0| \leq r_J, X = x\right)\end{aligned}$$

where I have also used Assumption 2.5 to write $E(M_{j_0} \mid X, \theta)$ as $p_{j_0,0}(\theta)$. As in Step 1, under Assumption 3.1,

$$\begin{aligned}&\left|E\left(E(M_{j_0} \mid X, \theta, \{M_j : j \in \mathcal{J}^J_{M_{j_0}}\})) - p_{j_0,0}(\theta) \mid |\bar{M} - m_0| \leq r_J, X = x\right)\right|\\&\leq \frac{E\left|E(M_{j_0} \mid X, \theta, \{M_j : j \in \mathcal{J}^J_{M_{j_0}}\})) - p_{j_0,0}(\theta)\right|}{Pr(|\bar{M} - m_0| \leq r_J, X = x)}\\&\leq \frac{\alpha_{\tilde{k}_J}}{Pr(|\bar{M} - m_0| \leq r_J, X = x)}\end{aligned}$$

where the second inequality follows from Assumption 3.1. Repeating arguments in Step 1, I can conclude that

$$\begin{aligned}&|E(M_{j_0} \mid |\bar{M} - m_0| \leq r_J, X = x) - p_{j_0,0}(t_0)|\\&\leq ((1-\varepsilon)c_Q^{-1}(c_{\bar{p}}^{-1}(r_J)))^{-1}\left(\alpha_{\tilde{k}_J} + 2Bk_J \exp(-\frac{1}{8}\frac{N_J r_J^2}{k_J}) + 2B\frac{\alpha_{k_J}}{r_J}\right) + c_C(c_{\bar{p}^{-1}}(3r_J))\end{aligned}$$

Let the right-hand side of this inequality be denoted by δ_J. The same line of argument, and the argument from step 3 imply that $|E(M_{j_0} \mid |\bar{M} - m_0| \leq r_J, X = x) - p_{j_0}(t)| \leq \delta_J$ so that by the triangle inequality

$$|p_{j_0,0}(t_0) - p_{j_0}(t)| \leq 2\delta_J \tag{A.18}$$

In fact, these arguments also imply that for any $x' \in \mathcal{X}$, $t_0' \in \Theta_0(x')$ and $m_0' \in [0,1]$, if $|\bar{p}_0(x', t_0') - m_0'| < r_J$ then $|E(M_{j_0} \mid |\bar{M} - m_0'| \leq r_J, X = x') - p_{j_0,0}(t_0')| \leq \delta_J$. Therefore, if $|E(M_{j_0} \mid |\bar{M} - m_0'| \leq r_J, X = x') - p_{j_0,0}(t_0')| \geq \delta_J$ for some $x' \in \mathcal{X}$, $t_0' \in \Theta_0(x')$ and $m_0' \in [0,1]$ it must be that $|\bar{p}_0(x', t_0') - m_0'| \geq r_J$. Let $T(m_0', x'; r_J) := E(M_{j_0} \mid |\bar{M} - m_0'| \leq r_J, X = x')$. Then this implies that

$$Pr(|T(\bar{M}, X; r_J) - p_{j_0,0}(\theta)| \geq \delta_J) \leq Pr(|\bar{p}_0(X,\theta) - \bar{M}| \geq r_J)$$

But, as we have seen, by Lemma A.2,

$$Pr(|\bar{p}_0(X,\theta) - \bar{M}| \geq r_J) \leq \rho_J := 2k_J \exp(-\frac{1}{8}\frac{N_J r_J^2}{k_J}) + 2\frac{\alpha_{k_J}}{r_J}$$

Then note that if $T(\bar{M}, X; r_J) < p_{j_0,0}(t_0)$ then either $p_{j_0,0}(\theta) < p_{j_0,0}(t_0) + \delta_J$ or $|T(\bar{M}, X; r_J) - p_{j_0,0}(\theta)| \geq \delta_J$. Likewise, if $p_{j_0,0}(\theta) < p_{j_0,0}(t_0) - \delta_J$ then either $T(\bar{M}, X; r_J) < p_{j_0,0}(t_0)$ or $|T(\bar{M}, X; r_J) - p_{j_0,0}(\theta)| \geq \delta_J$. Therefore,

$$\begin{aligned}
&Pr(p_{j_0,0}(\theta) < p_{j_0,0}(t_0) - \delta_J) - Pr(|T(\bar{M}, X; r_J) - p_{j_0,0}(\theta)| \geq \delta_J) \\
&\leq Pr(T(\bar{M}, X; r_J) < p_{j_0,0}(t_0)) \\
&\leq Pr(p_{j_0,0}(\theta) < p_{j_0,0}(t_0) + \delta_J) + Pr(|T(\bar{M}, X; r_J) - p_{j_0,0}(\theta)| \geq \delta_J)
\end{aligned}$$

By Assumptions 2.4 and 3.2, $Pr(p_{j_0,0}(\theta) < p_{j_0,0}(t_0) \pm \delta_J) = p_{j_0,0}^{-1}(p_{j_0,0}(t_0) \pm \delta_J)$, and therefore,

$$|Pr(T(\bar{M}, X; r_J) < p_{j_0,0}(t_0)) - t_0| \leq c_{p_{j_0}^{-1}}(\delta_J) + \rho_J \tag{A.19}$$

It is also the case that

$$\begin{aligned}
&Pr(p_{j_0,0}(\theta) < p_{j_0}(t) - \delta_J + p_{j_0,0}(t_0) - p_{j_0}(t)) - \rho_J \\
&\leq Pr(T(\bar{M}, X; r_J) < p_{j_0,0}(t_0)) \\
&\leq Pr(p_{j_0,0}(\theta) < p_{j_0}(t) + \delta_J + p_{j_0,0}(t_0) - p_{j_0}(t)) + \rho_J
\end{aligned}$$

which implies that

$$|Pr(T(\bar{M}, X; r_J) < p_{j_0,0}(t_0)) - t| \leq c_{p_{j_0}^{-1}}(\delta_J + |p_{j_0,0}(t_0) - p_{j_0}(t)|) + \rho_J \tag{A.20}$$

Combining (A.18)-(A.20), by the triangle inequality,

$$|t - t_0| \leq 2c_{p_{j_0}^{-1}}(3\delta_J) + 2\rho_J \tag{A.21}$$

Step 5 Now, because $|CASF(x,t) - CASF(x,t_0)| \leq c_C(|t-t_0|)$, results (A.15), (A.17), and (A.21) imply that

$$\begin{aligned}
&|CASF_0(x,t_0) - CASF(x,t_0)| \\
&\leq |E(Y \mid |\bar{M} - m_0| \leq r_J, X = x) - CASF_0(x,t_0)| \\
&+ |E(Y \mid |\bar{M} - m_0| \leq r_J, X = x) - CASF(x,t)| \\
&+ |CASF(x,t) - CASF(x,t_0)| \\
&\leq 2((1-\varepsilon)c_Q^{-1}(c_{\bar{p}}^{-1}(r_J)))^{-1}\left(\alpha_{\tilde{k}_J} + 2Bk_J\exp(-\frac{1}{8}\frac{N_J r_J^2}{k_J}) + 2B\frac{\alpha_{k_J}}{r_J}\right) \\
&+ 2c_C(c_{\bar{p}^{-1}}(3r_J)) + c_C(2c_{p_{j_0}^{-1}}(3\delta_J) + 2\rho_J)
\end{aligned}$$

Recalling the definition of δ_J above, it is evident that it is sufficient to show that

$$(c_Q^{-1}(c_{\bar{p}}^{-1}(r_J)))^{-1}\left(\alpha_{\tilde{k}_J} + 2Bk_J\exp(-\frac{1}{8}\frac{N_J r_J^2}{k_J}) + 2B\frac{\alpha_{k_J}}{r_J}\right) \to 0$$

First, $\tilde{k_J} \geq \frac{1-5\alpha}{\alpha}N_J$ but $k_J < aN_J$ for any constant $a > 0$. Therefore, $\tilde{k_J} > k_J$ so $\alpha_{\tilde{k}_J} < \alpha_{k_J}$ so that

$$\begin{aligned}
&(c_Q^{-1}(c_{\bar{p}}^{-1}(r_J)))^{-1}\alpha_{\tilde{k}_J} \\
&\leq \frac{\alpha_{\tilde{k}_J}}{r_J c_Q^{-1}(c_{\bar{p}}^{-1}(r_J))} \\
&\leq \frac{\alpha_{k_J}}{r_J c_Q^{-1}(c_{\bar{p}}^{-1}(r_J))}
\end{aligned}$$

Next, because $N_J \to \infty$, r_J and k_J were chosen so that

$$\begin{aligned}
&\lim_{J\to\infty}(c_Q^{-1}(c_{\bar{p}}^{-1}(r_J)))^{-1}k_J\exp(-\frac{1}{8}\frac{N_J r_J^2}{k_J}) \\
&= \lim_{N\to\infty}(c_Q^{-1}(c_{\bar{p}}^{-1}(r_N)))^{-1}k_N\exp(-\frac{1}{8}\frac{N r_N^2}{k_N}) = 0
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{J\to\infty}(c_Q^{-1}(c_{\bar{p}}^{-1}(r_J)))^{-1}\frac{\alpha_{k_J}}{r_J} \\
&= \lim_{N\to\infty}(c_Q^{-1}(c_{\bar{p}}^{-1}(r_N)))^{-1}\frac{\alpha_{k_N}}{r_N} = 0
\end{aligned}$$

which Lemma A.3 shows is possible. □

A.4 Proofs of Corollaries 3.1 and 3.2

First, Corollary 3.1

Proof. Because F_θ is absolutely continuous there exists a density function f_θ such that $\int CASF(x,t)dF_\theta(t) = \int_{-\infty}^{\infty} CASF(x,t)f_\theta(t)dt$.

Divide the interval $[-\psi_J^{-1/2}, \psi_J^{-1/2}]$ into $S_J = \psi_J^{-1/2}$ intervals of size ψ_J, $[t_0^0, t_0^1], ..., [t_0^{S-1}, t_0^S]$. Then uniform continuity of $CASF(x,t)$ can be used to show that

$$\begin{aligned}
&\left| \int_{-\psi_J^{-1/2}}^{\psi_J^{-1/2}} CASF_0(x,t_0)f_\theta^0(t_0)dt_0 - \sum_{s=1}^{S_J} CASF(x,t_0^s)(F_\theta^0(t_0^s) - F_\theta^0(t_0^{s-1})) \right| \\
&\leq \int_{-\psi_J^{-1/2}}^{\psi_J^{-1/2}} |CASF_0(x,t_0) - CASF_0(x,t_0^s)| f_\theta^0(t_0)dt_0 \\
&\leq c_C(\psi_J)
\end{aligned}$$

In addition, because $||CASF_0(x,t_0)|| \leq \bar{Y}$,

$$\lim_{J\to\infty} \int_{-\infty}^{-\psi_J^{-1/2}} CASF_0(x,t_0)f_\theta^0(t_0)dt_0 = \lim_{J\to\infty} \int_{\psi_J^{-1/2}}^{\infty} CASF_0(x,t_0)f_\theta^0(t_0)dt_0 = 0$$

By Step 2 in the proof of Theorem 3.1, for each s there is a t^s such that $|\bar{p}(x,t^s) - \bar{p}_0(x,t_0^s)| \leq r_J$. This implies in turn that $|t^s - t^{s-1}| \leq c_{\bar{p}^{-1}}(r_J + c_{\bar{p}}(\psi_J))$.

The only thing left is to show that

$$\left| \sum_{s=1}^{S_J} CASF_0(x,t_0^s)(F_\theta^0(t_0^s) - F_\theta^0(t_0^{s-1})) - \sum_{s=1}^{S_J} CASF(x,t^s)(F_\theta(t^s) - F_\theta(t^{s-1})) \right| \to 0$$

Without Assumption 2.4, the arguments in Step 4 in the proof of Theorem 3.1 instead lead to

$$|F_\theta(t) - F_\theta^0(t_0)| \leq 2c_{p_{j_0}^{-1}}(3\delta_J) + 2\rho_J$$

Therefore, taking $\psi_J = 2c_{p_{j_0}^{-1}}(3\delta_J) + 2\rho_J$

$$\begin{aligned}
&\left|\sum_{s=1}^{S_J} CASF_0(x, t_0^s)(F_\theta^0(t_0^s) - F_\theta^0(t_0^{s-1})) - \sum_{s=1}^{S_J} CASF(x, t^s)(F_\theta(t^s) - F_\theta(t^{s-1}))\right| \\
&\leq \sum_{s=1}^{S_J} \left|CASF_0(x, t_0^s) - CASF(x, t^s)\right| (F_\theta^0(t_0^s) - F_\theta^0(t_0^{s-1})) \\
&+ \sum_{s=1}^{S_J} CASF(x, t^s) \left|(F_\theta^0(t_0^s) - F_\theta^0(t_0^{s-1})) - (F_\theta(t^s) - F_\theta(t^{s-1}))\right| \\
&\leq \max_{x,s} \left|CASF_0(x, t_0^s) - CASF(x, t^s)\right| + 2\bar{Y} S_J \psi_J \to 0
\end{aligned}$$

□

Next, the proof of Corollary 3.2.

Proof. Step 4 in the proof of Theorem 3.1 can be modified to show that

$$|Pr(\bar{M} < m_0) - t_0| \leq c_{\bar{p}^{-1}}(r_J) + \rho_J$$

and

$$|Pr(\bar{M} < m_0) - t| \leq c_{\bar{p}^{-1}}(2r_J) + \rho_J$$

without using Assumption 2.5. The rest of the proof is identical to that of Theorem 3.1. □

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.

Table 2. 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

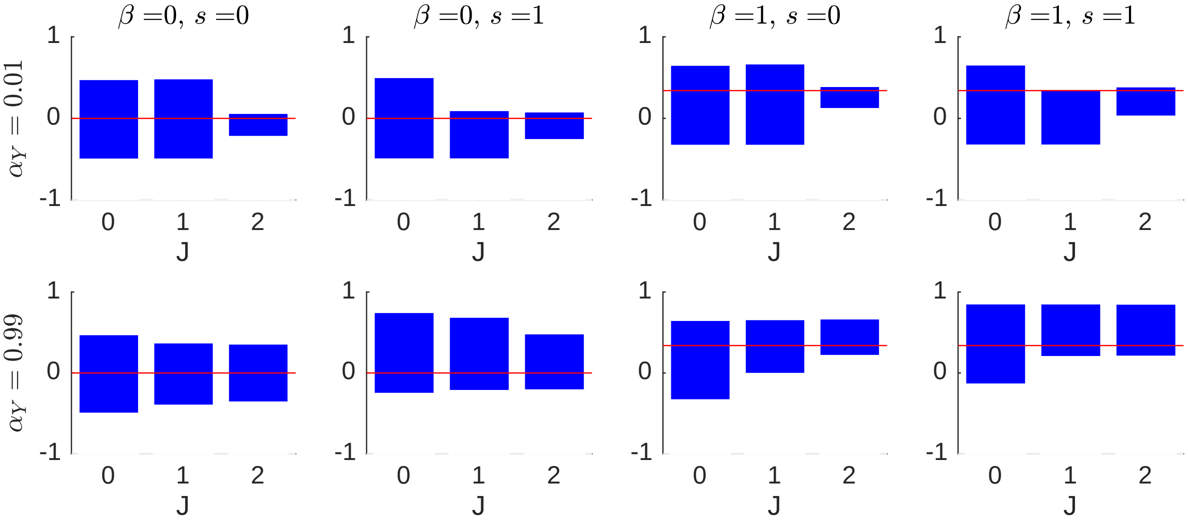
Notes: The first column in this table reports the coefficient on SomeCollege in an OLS regression that uses the proxies as controls. The outcome is an indicator for whether the individual has voted in and election in the past two years. The sample size is 10,515.

Table 3. Average treatment effect estimates

outcome	ATE	90% CI		outcome	ATE	90% CI	
wage	0.074	0.017	0.119	item 16	0.166	0.075	0.202
item 1	0.036	0.022	0.049	item 17	0.125	0.042	0.191
item 2	0.090	0.059	0.129	item 18	0.229	0.164	0.280
item 4	0.145	0.102	0.193	item 19	0.172	0.114	0.236
item 5	0.173	0.127	0.226	item 20	0.221	0.148	0.287
item 6	0.162	0.109	0.198	item 21	0.192	0.095	0.233
item 7	0.052	0.003	0.086	item 22	0.223	0.166	0.303
item 8	0.134	0.080	0.173	item 23	0.265	0.205	0.319
item 9	0.164	0.115	0.208	item 24	0.255	0.175	0.317
item 10	0.169	0.119	0.212	item 25	0.231	0.151	0.288
item 11	0.208	0.145	0.257	item 26	0.167	0.097	0.223
item 12	0.163	0.090	0.210	item 27	0.189	0.126	0.275
item 13	0.258	0.194	0.296	item 28	0.216	0.161	0.283
item 14	0.238	0.162	0.294	item 29	0.207	0.143	0.267
item 15	0.190	0.111	0.234	item 30	0.178	0.093	0.242

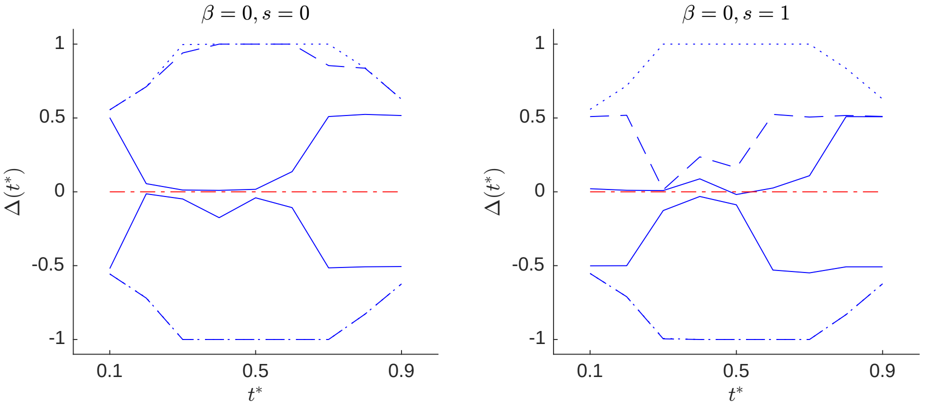
Notes: This table shows estimates of the average treatment effect (ATE), computed as described in the text, and 90% confidence intervals, obtained from simulating 100 boostrap samples. The sample size was 1,927 for all outcomes except wages. The sample used to estimate the wage ATE consisted of 1,018 observations. See the text for a description of the sample.

Figure 1: Bounds on the ATE as a function of J



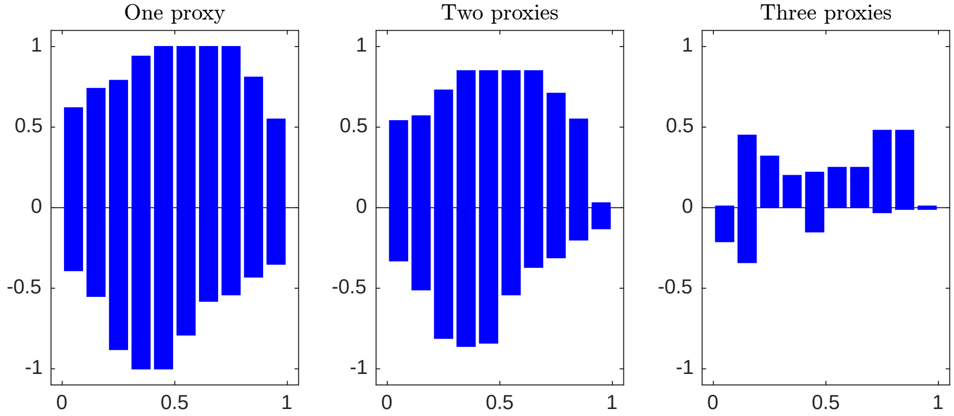
Notes: I used a uniformly spaced partition of $S = 20$ points and set $\epsilon_S = 10^{-8}$. For each graph the parameter values used to generate the model probabilities are given in the row and column headings. In each graph, the dashed line (----) marks the value of the ATE under the dgp and the solid bars denote the identified set for the ATE.

Figure 2: Pointwise bounds on $\Delta(t^*)$



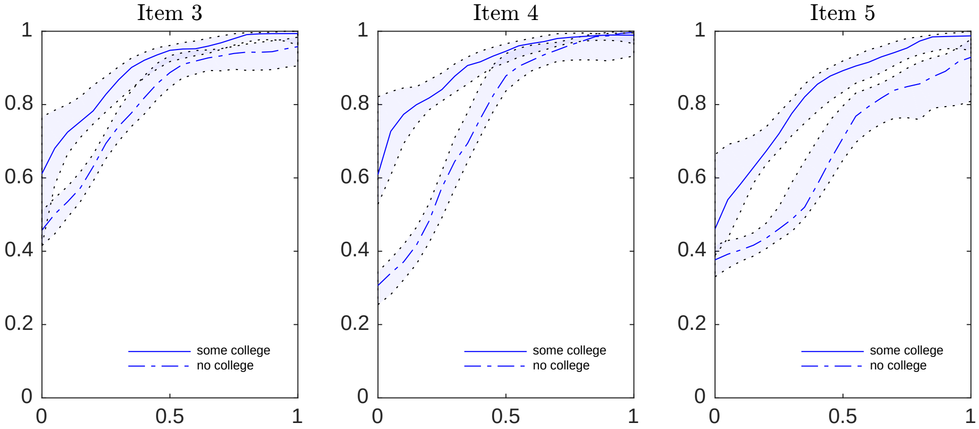
Notes: I used a uniformly spaced partition of $S = 20$ points and set $\epsilon_S = 10^{-8}$. For each graph the parameter values used to generate the model probabilities are given in the row and column headings. The dotted line (····) denotes bounds from the model with $J = 1$, the dashed line (-- --) denotes bounds from the model with $J = 2$ and the solid line (—) denotes bounds from the model with $J = 3$. A line (----) at 0 marks the value of the ATE in the data generating process.

Figure 3: Bounds on the conditional ATE of education on the probability of voting



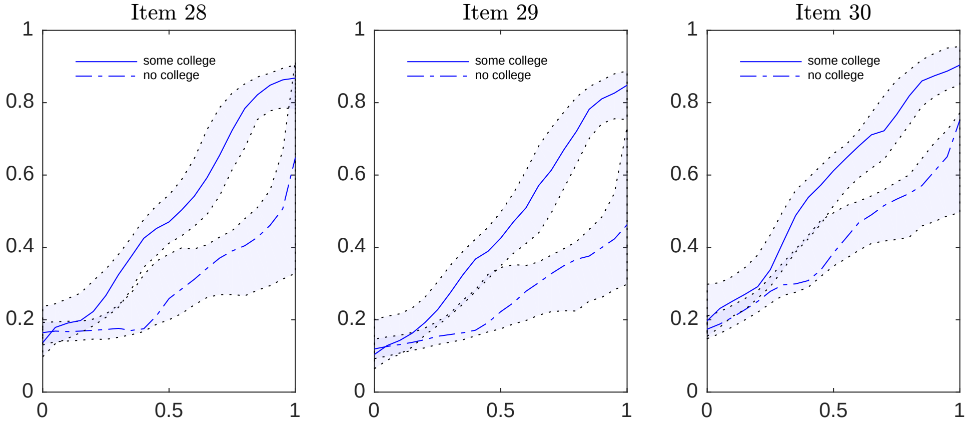
Notes: The bounds are computed as described in Section 4.1. Estimates are based on a sample of size $10,515$ from the HSB longitudinal survey.

Figure 4: ASF estimates – items 3-4 of the AR component of the ASVAB



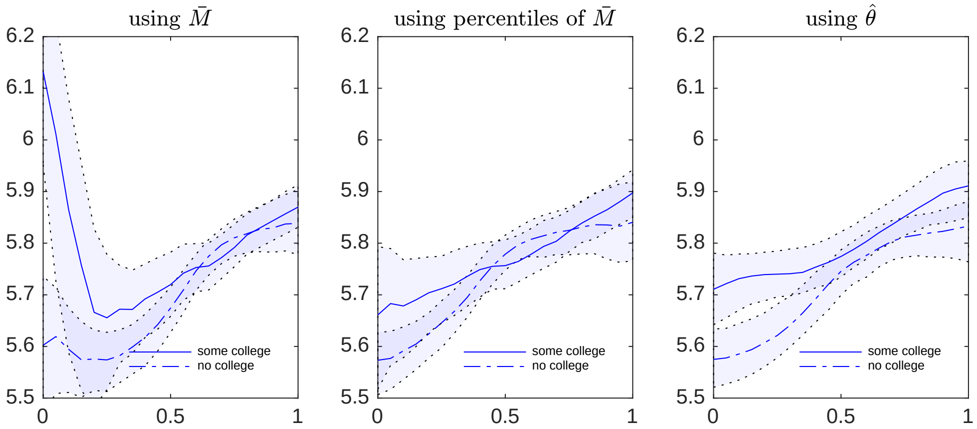
Notes: The ASF estimates in the three panels are computed as described in the text treating responses to the third, fourth, and fifth items on the Arithmetic Reasoning component of the ASVAB as the outcome. The Epanechnikov kernel was used in both steps. The shaded region is a 90% confidence interval computed using 100 bootstrap samples. Estimates are based on a sample of size $1,927$ from the NLSY79.

Figure 5: ASF estimates – items 28-30 of 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 100 bootstrap samples. Estimates are based on a sample of size $1,927$ from the NLSY79.

Figure 6: Three ASF estimates of wages



Notes: This figure displays three separate ASF estimates of log wages as a function of education in the three panels. The first panel shows separate kernel regression of log wages on $\bar{M}$ for those with some college and those without. The second panel shows separate kernel regression of log wages on $F_{\bar{M}}(\bar{M})$ for those with some college and those without. The third panel shows estimates of the ASF computed as proposed in Section 3.2 and described in the text. The Epanechnikov kernel was used in both steps. The shaded region is a 90% confidence interval computed using 100 bootstrap samples. Estimates are based on a sample of size $1,927$ from the NLSY79.