Submitted to Econometrica

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

Benjamin Williams

July 12, 2014
IDENTIFICATION OF A NONSEPARABLE MODEL UNDER ENDOGENEITY USING
BINARY PROXIES FOR UNOBSERVED HETEROGENEITY

Benjamin Williams

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 a large number of binary indicators that provide proxies for the unobserved heterogeneity. Identification is then viewed as a nonstandard measurement error problem. It is nonstandard because I allow for multiple dimensions of the unobserved heterogeneity, dependence between the measurement error and the endogenous explanatory variable, and binary measurements of the unobserved heterogeneity which is continuously distributed. The model is point identified in the limit as the number of indicators increases.


1. INTRODUCTION

The prevalence of unobserved heterogeneity is one of the most important empirical discoveries in economics and accounting for it is one of the greatest challenges of the field of microeconometrics (Heckman, 2001). Models that are additively nonseparable in the unobserved heterogeneity allow effects to vary across observably identical individuals. In economic models with unobserved heterogeneity of this form, endogeneity is often present because individuals act on knowledge, or at least partial knowledge, of these unobserved factors. This phenomenon, sometimes called essential heterogeneity (Heckman, Urzua, and Vytlacil, 2006), has been studied extensively in the recent econometrics literature.

Existing approaches to identification in nonseparable models either assume that the endogenous regressor is exogenous conditional on a set of additional controls or that an exogenous instrumental variable is available.1 Matzkin (2007) labels the two approaches the

1 This work was previously circulated as part of the paper “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 by Ed Vytlacil, Martin Browning, Steve Durlauf, Yingyao Hu, Richard Spady, Stephane Bonhomme, Bob Phillips, Tim Moore, anonymous referees on a previous version and seminar participants at the University of Chicago, George Washington University, Johns Hopkins University, the Center for Education Research at UW-Madison, the FDIC, and the 2013 Latin American Workshop in Econometrics. I can be contacted at bdwilliams@gwu.edu.

1 One exception is Matzkin (2004).
conditional independence approach and the marginal independence approach, respectively. Heckman and Navarro (2004) make a similar distinction between the matching approach and the (local) instrumental variable approach. In the context of a model with essential heterogeneity the conditional independence approach requires that the econometrician observes all relevant information observed by economic agents (Heckman and Navarro, 2004) and the marginal independence assumption requires that the economic agent’s decision is driven in part by an instrumental variable (for example, a cost factor) that is excluded from the outcome equation. Both types of conditions can be hard to justify.

In this paper I instead use additional variables that play a role distinct from the role of either additional regressors or an instrumental variable. These additional variables must be conditionally independent of the outcome and mutually conditionally independent given a vector of latent variables describing the unobserved heterogeneity. This will allow these observed variables to serve as proxies for the latent heterogeneity. Williams (2012) discusses this as a general approach to identification.

If these additional variables are continuous then an extension of results in Hu and Schennach (2008) could be applied; see, e.g., Cunha, Heckman, and Schennach (2010). Motivated in part by the use of item response data in Heckman, Stixrud, and Urzua (2006) and other studies of the role of personality traits in labor market outcomes (see the review article, Almlund, Duckworth, Heckman, and Kautz, 2011, for a discussion), this paper considers the case where the additional variables are all binary. In addition, although test performance data are often aggregated into a single score before analysis, this type of data can be viewed as binary responses to a series of individual items.

The binary variables that supplement the main model of this paper can also be additional outcomes and do not have to be interpreted as item responses intended to measure some individual characteristic. For example, Cunha, Heckman, and Schennach (2010) and Aizer and Cunha (2012) use multiple binary variables to proxy for parental investment in the human capital of children.

The main result of this paper provides general conditions under which this type of additional data can be used to solve the identification problem in a nonseparable model with

---

2Item response data, a term used primarily in the psychology literature Lord (1980), refers to a set of binary responses which are all noisy measures of the same vector of latent individual characteristics.

3The recent release of individual item data for some components of the Armed Forces Qualifying Test in the 1979 National Longitudinal Survey of Youth makes this particularly relevant.
endogeneity. Speaking broadly there are three important identifying assumptions. First, these binary proxies must be mutually conditionally independent and conditionally independent of the outcome given the latent variables which are the source of endogeneity. Second, they must exhibit sufficient variation to uncover these latent variables. Because the proxies are binary and the latent variables are continuous, this will be satisfied only in a limiting sense as the number of these additional binary variables grows. And third, the endogenous regressor must be excluded from the equation for a small number of the proxies.\footnote{The number of exclusions required is equal to the dimension of the latent variable.} \footnote{This third condition is used for the main result in Section 3 but is not strictly needed. Section 4 discusses alternative strategies.}

Item response models and related factor models have been used extensively in structural econometric models. Much of the early work using these latent variables models was developed to account for the ability bias in estimates of the return to education (Chamberlain, 1977; Chamberlain and Griliches, 1975, et al.). See Aigner, Hsiao, Kapteyn, and Wansbeek (1984) for a review of the early literature. Latent variables models, more specifically the factor model, has been developed extensively as a tool in the study of counterfactual outcome models (Aakvik, Heckman, and Vytlacil, 2005; Abbring and Heckman, 2007; Carneiro, Hansen, and Heckman, 2003; Cunha, Heckman, and Navarro, 2005). More recent advances in the econometrics literature include Cunha and Heckman (2008) and Cunha, Heckman, and Schennach (2010). This recent work has used advances in the measurement error literature (Hu and Schennach, 2008) to obtain nonparametric identification results in these models.

This paper contributes to the existing literature by considering a model where (i) there is endogeneity in the outcome equation, (ii) the endogenous regressor is permitted to enter the measurement equations, (iii) both the main outcome equation and the equations for the binary indicators are nonseparable, (iv) the “measurements” are all binary, and (v) observed regressors are arbitrarily correlated with the latent variables. Moreover, the results here do not require a first stage model for the endogenous variable and do not call for an instrumental variable that is excluded from the outcome equation. While there are some existing identification results for models satisfying one or more of these properties this paper provides an identification result in a model satisfying them all. Some work (e.g. Carneiro, Hansen, and Heckman, 2003) considers similar models but relies on additional parametric structure, instrumental variables, and/or independence between some observed regressors.
and the latent variables. This paper complements the identification results of Carneiro, Hansen, and Heckman (2003) and similar results.

In the psychological and educational testing literature, similar models are used extensively but existing methods are primarily parametric. The available results there on nonparametric identification (Douglas, 1997, 2001; Ramsay, 1991) provide an important starting point for the results in this paper but these results are more limited than the results here as the focus there is primarily on the measurement of an underlying trait. Specifically this work does not typically consider identification with potentially endogenous explanatory variables or multiple dimensions of the latent variables. The existing work in psychometrics studying multidimensional models imposes stronger parametric or semiparametric assumptions than I do here (Peress, 2012; Reckase, 1997). See Sijtsma and Junker (2006) for a concise review of the current state of the literature.

The results in this paper also contribute to the measurement error literature in econometrics and statistics. While the motivation here is to solve the endogeneity problem and identify the heterogeneous effects of an endogenous, but observed, variable, the identification of the model presented in this paper also implies that the effect of variation in the latent variables is identified. In fact, as a special case of the main result, the model is identified even in the absence of any observed regressor $X$. Thus the paper also complements Mahajan (2006), Gawade (2007), Hu and Schennach (2008), and Hu (2008), among others, by providing identification results for a measurement error model with discrete measurements of continuous latent variables. Douglas (2001) is the only paper I am aware of that addresses this type of measurement error. The measurement error perspective is considered in greater detail in Williams (2013).

The rest of the paper is organized as follows. Section 2 defines the model for the nonseparable outcome with endogeneity and the item response data and discusses the identifying assumptions. The main identification result, as well as a preliminary result on a toy model, is presented in Section 3. In Section 4 I consider several additional identification results that complement the main result. The paper concludes with Section 5.

---

6The latter condition is similar to the use of “special regressors” in some other recent identification results (Honore and Lewbel, 2002; Lewbel, 1998).

7Chernozhukov, Fernández-Val, Hahn, and Newey (2013) and Browning and Carro (2014), among others, consider panel models with this combination of binary observed outcomes and continuous latent variables (fixed effects) but the identification strategy is unique to the panel data setting.
2. THE MODEL

Let $Y$ denote an outcome and let $X$ be a vector of endogenous\(^8\) explanatory variables and suppose

$$Y = g(X, \eta)$$

where $\eta$ represents the unobserved heterogeneity. In some identification arguments that use the instrumental variable approach it is assumed that the unobservable in the outcome equation, $\eta$, is scalar and that $g$ is monotonic in $\eta$ (Chernozhukov and Hansen, 2005; d’Haultefoeuille and Fevrier, 2012; Newey and Powell, 2003; Torgovitsky, 2012). Other arguments based on instrumental variables avoid this but model $X$ in a first stage equation that is monotonic in a scalar unobservable (Chesher, 2003; Florens, Heckman, Meghir, and Vytlačil, 2008; Hoderlein and Sasaki, 2013; Imbens and Newey, 2009; Newey, Powell, and Vella, 1999) or monotonic in the instrument (Kasy, 2014). Kasy (2011) provides some results on the necessity of restricting the dimension of the unobserved heterogeneity in nonseparable models. A matching, or conditional independence, approach avoids restrictions on the dimensions of the model but does not identify any unobserved heterogeneity in effects.

Instead, I consider the outcome model

$$(2.1) \quad Y = g(X, \theta, U)$$

where $\theta$ and $U$ are two separate components of $\eta$ and I avoid modeling $X$ in a first stage equation. Chesher (2003) also explicitly decomposes the unobservable $\eta$ into two separate components, though the decomposition facilitates a different identification argument that relies on an excluded instrument and a first stage equation. I do not, however, entirely avoid monotonicity/dimensionality restrictions; see the invertibility condition in Assumption M below. I assume that the decomposition of $\eta$ satisfies two crucial properties.

First consider some common sources of endogeneity. Endogeneity may arise because $X$, or some components of $X$, are chosen to optimize an objective function that depends in some way on $\{Y_x\}_{x \in \mathcal{X}}$, where $\mathcal{X}$ denotes the support of $X$. In many cases this optimization is under imperfect information on the part of the decision-maker, in which case $X$ will only depend on components of $\eta$ which are known to the agent at the time of decision-making.

---

\(^8\)It will not be necessary to make a distinction between endogenous and exogenous explanatory variables in the model, and therefore I will treat all explanatory variables as potentially endogenous.
Endogeneity may also be due to an omitted variables bias where $X$ and some components of $\eta$ represent related individual characteristics, variables which are jointly determined. Also, $\theta$ may be chosen as a function of external variables, $X$; see Example 2 below.

The first property satisfied by the decomposition in equation (2.1) is that $\theta$ alone is the source of endogeneity in the sense of the following assumption.

**Assumption 2.1** $U \perp\!\!\!\!\perp X, \theta$

If Assumption 2.1 fails but $U \perp\!\!\!\!\perp X \mid \theta$ then the model can be rewritten to satisfy the assumption since $g(X, \theta, U) = g(X, \theta, F_{U|\theta}^{-1}(\tilde{U} \mid \theta))$ where $\tilde{U} \perp\!\!\!\!\perp X, \theta$. Therefore it is possible to allow dependence between $U$ and $\theta$ at the expense of not being able to separately identifying the causal effect of $\theta$ on $Y$. If endogeneity is present because of selection then this assumption requires that $\theta$ captures the part of $\eta$ that is in the agent’s information set. If endogeneity arises due to an omitted variable as described above, this assumption requires that $\theta$ includes all unobserved individual characteristics that are related to $\theta$. Assumption 2.1 would be the main identifying assumption if $\theta$ were observed. In that case it is essentially the matching, or conditional independence, assumption. Suppose $X = (X_1, X_2)$ and the object of interest is the causal effect of $X_1$. Assumption 2.1 implies that $Y_{x_1} := g(x_1, X_2, \eta)$ is independent of $X_1$ given $X_2$ and $\theta$. This is the unconfoundedness assumption common in the treatment effect literature (Heckman, Ichimura, and Todd, 1998; Imbens, 2004). Equation (2.1) implicitly assumes, in addition, an invariance of $\eta$ to counterfactual manipulations of $X$, though this is not crucial to the arguments developed in this paper.

The second property satisfied by the decomposition in equation (2.1) is that there are $\tilde{J}$ observed binary variables satisfying

\begin{equation}
M_j = 1(h_j(X, \theta) \geq \epsilon_j), \quad j = 1, \ldots, \tilde{J}
\end{equation}

and

**Assumption 2.2**

(i) $\text{Supp}(\theta) \subset \mathbb{R}^K$

(ii) $U, \epsilon_1, \ldots, \epsilon_{\tilde{J}}$ are mutually independent conditional on $X, \theta$

(iii) there are $K$ distinct values of $j$ such that (a) $X$ is excluded from $h_j$, that is, $h_j(x, t) \equiv h_j(x', t)$ for all $x, x', t$ and (b) $\epsilon_j \perp\!\!\!\!\perp X \mid \theta$. 

These additional observed variables $M_j$ can be thought of as related outcomes that are independent of $Y$, and of each other, conditional on $X$ and $\theta$, or as binary proxies for the latent $\theta$. Rather than modeling $X$ explicitly, as in the common triangular models, equation (2.2) models other outcomes of which $\theta$ is a determinant, or measurements of $\theta$, and Assumption 2.2 requires that a small number of these are independent of $X$ conditional on $\theta$. Because these additional observed variables are binary, but $\theta$ is continuously distributed, the identification argument will require an asymptotic argument where $\tilde{J} \to \infty$.\footnote{That is, there must be continuous variation in the data to uncover the effect of continuous variation in $\theta$ on $Y$ or, alternatively, to uncover the continuous distribution of $\theta$.} I do not explicitly distinguish between elements of $X$ that are determinants of $M_j$ for different $j$. Such a delineation is possible with the caveat that the dimension of $X$ must remain fixed while the number of proxies increases.

Condition (i) of Assumption 2.2 enforces a dimensionality restriction on the unobserved heterogeneity in the outcome model. The dimension, $K$, is assumed to be known and fixed as $\tilde{J}$ increases. Condition (ii) implies that the outcomes, $Y, M_1, \ldots, M_{\tilde{J}}$, are conditionally independent given $X$ and $\theta$. If $\varepsilon_j$ is interpreted as primarily measurement error then this requires (conditionally) independent measurement error. If the $M_j$ represent additional outcomes then this condition requires these different outcomes to be determined independently, conditional on $X$ and $\theta$, and hence rules out feedback among the different outcomes.

Condition (iii) is an exclusion restriction that requires that some of the binary indicators do not depend on $X$ after conditioning on $\theta$. If $\theta$ is scalar this imposes a restriction on only one of the many items observed. Generally $K$, the dimension of $\theta$, will be much smaller than $\tilde{J}$ and hence most indicators are allowed to depend arbitrarily on both $X$ and $\theta$. The condition is easily satisfied when one indicator is available that is essentially a direct measurement of each relevant underlying trait. The exclusion is crucial to the identification argument as it enables treating $\theta$ and $X$ as correlated in the analysis. In Section 4 I discuss alternative solutions.

Example 1. This example is motivated by the popular empirical model of Aakvik, Heckman, and Vytlacil (2005); Carneiro, Hansen, and Heckman (2003); Heckman, Stixrud, and Urzua (2006). Let $Y$ denote an individual outcome such as wages at age 40 and let $X$ denote the
individual’s level of education. The nonseparable model of equation (2.1) generalizes the random coefficients model of the returns to education proposed by Becker and Chiswick (1966). This nonseparability is empirically relevant (Heckman, Schmiederer, and Urzua, 2010). The primary driver of heterogeneity in returns appears to be variation in abilities. Heckman, Stixrud, and Urzua (2006) find evidence of two important dimensions of ability – cognitive ability and noncognitive traits.\(^{10}\) Thus let \(\theta = (\theta^C, \theta^N)\) where \(\theta^C\) represents the individual’s latent cognitive ability and \(\theta^N\) represents the individual’s noncognitive ability. The binary indicators \(M_1, \ldots, M_J\) include responses to each item of an achievement test and responses to a short questionnaire on personality traits. Suppose these are administered after the individual’s schooling has been completed. The responses may depend on \(X\) conditional on \(\theta\) as there is compelling evidence that education can enhance latent ability (Hansen, Heckman, and Mullen, 2004; Winship and Korenman, 1997). Since \(K = 2\), Assumption 2.2(iii) is satisfied if responses to two test items do not depend on the individual’s education level. These items may be, for example, regarding basic enough tasks or skills that the ability to answer them correctly does not increase with more education.

\(\text{Example 2.}\) Suppose each \(M_j\) represents an input to parental investment in the human capital of a child, \(\theta\). Following Cunha and Heckman (2008), let \(M_j = 1(M_j^* \geq c_j)\) where \(M_1^*, \ldots, M_J^*\) is a subset of the vector of inputs that solves

\[
\min_{m_1, \ldots, m_S} \sum_{s=1}^{S} w_s m_s \\
\text{s.t. } \theta = f(m_1, \ldots, m_S; \gamma)
\]

where \(w_s\) represents the price of input \(m_s\) and \(\gamma\) represents parameters of the production function. The variable \(Y\) may then represent an academic or behavioral outcome of the child or the demand for one of the other inputs. The explanatory variables \(X\) are observed variables that induce variation in prices, \(w_s\), technology parameters, \(\gamma\), or measurement error, \(c_j\). If \(\theta\) is chosen optimally in a second stage utility maximization, it will typically be correlated with \(X\). This is the source of the endogeneity problem.

\(^{10}\)Recent research has shown the importance of additional dimensions such as health (Conti, Heckman, Fruhwirth-Schnatter, Lopes, and Piatek, 2012).
By condition (ii) of Assumption 2.2 demand for the inputs must be conditionally independent given $\theta$ and $X$. If there is unobserved heterogeneity in preferences then the dimension of $\theta$ must be increased to account for this. Condition (iii) of 2.2 requires that demand for one of the inputs is invariant to changes in $X$. The plausibility of this condition will depend on the role of $X$ but suppose, for example, that $X = w_1$. If $f(m_1, \ldots, m_S) = f_1(m_1, \ldots, m_{S_0}) + f_2(m_{S_0+1}, \ldots, m_S)$ where $f_1(m_1, \ldots, m_{S_0})$ is a Cobb-Douglas production technology then demand for inputs $m_2, \ldots, m_{S_0}$ will not depend on $X$ conditional on $\theta$.

Next, I consider three different structures for the outcome equation. These are motivated by models that would be identified if $\theta$ were observed. First, the function $\bar{g}(x,t) = E(g(x,t,U))$ would be identified from $E(Y \mid X = x, \theta = t)$ under Assumption 2.1 because $E(Y \mid X = x, \theta = t) = E(g(x,t,U) \mid X = x, \theta = t) = \bar{g}(x,t)$. Here, since $\theta$ is not observed, $\bar{g}(x,t)$ is an average over some but not all of the unobserved heterogeneity and thus still represents a heterogeneous return. Hoderlein and Mammen (2007) and Sasaki (2013) discuss alternative average effects that are identified from the distribution of $Y \mid X, \theta$ when the dimension of $U$ is unrestricted which I do not consider here.

If more structure is imposed then the function $g$ can be identified rather than just average effects. Indeed if $\theta$ were observed the function $g$ would be identified under either of the following two sets of conditions.

**Assumption 2.3** $U \sim \text{Uniform}(0,1)$, the random variable $Y$ is absolutely continuous with respect to Lebesgue measure on $Y$, and for every $(x,t) \in \text{Supp}(X,\theta)$, $g(x,t,\cdot)$ is continuous and strictly increasing.

Under Assumptions 2.1 and 2.3, for each $(x,t) \in \text{Supp}(X,\theta)$ and $u \in [0,1]$, $Pr(Y \leq g(x,t,u) \mid X = x, \theta = t) = u$. In other words, $g(x,t,u)$ is identified from the conditional quantile function, $g(x,t,u) = Q_{Y \mid X,\theta}(u \mid x,t)$. The results in this paper however do not rely on the outcome $Y$ being continuously distributed. The main result also provides identification under the following binary outcome structure.

**Assumption 2.4** $U \sim \text{Uniform}(0,1)$, and $Y$ is binary with $g(x,t,u) = 1(h(x,t) \geq u)$.

Under Assumptions 2.1 and 2.4, $Pr(Y = 1 \mid X = x, \theta = t) = h(x,t)$. This is more general than the semiparametric model, $g(x,t,u) = 1(\alpha + \beta'x + \gamma'\theta \geq u)$, that also puts quantile
and scale restrictions on $u$, but does not specify its distribution, because this semiparametric model implies that the marginal effects of components of $X$ and $\theta$ are proportional. On the other hand, the semiparametric model can be identified under weaker conditions (Peress, 2012). I consider identification under all three setups to emphasize that the identification strategy here does not depend on monotonicity in a scalar unobservable in the outcome equation.

Since $\theta$ is not observed, these conditions alone are not sufficient. As I demonstrate in the next section identification relies on the ability of the binary variables $M_j$ to serve as proxies for $\theta$ as $\bar{J} \to \infty$, which in part requires exclusion restrictions and normalizations. One set of exclusions has already been stated in Assumption 2.2(iii). In addition to that restriction the identification argument uses the following conditions.

**Assumption 2.5**

(i) $\text{Supp}(X,\theta) = \mathcal{X} \times [0,1]^K$

(ii) each component of $\theta$ satisfies $\theta_k \sim \text{Uniform}(0,1)$

(iii) There are $K$ distinct values of $j$, denoted $j_1, \ldots, j_K$, such that for all $x,t,t'$

$$h_{jk}(x,t) = h_{jk}(x,t'_1,\ldots,t'_{k-1},t_k,\ldots,t'_{k+1},\ldots,t_K)$$

and $\varepsilon_{jk} \perp \perp \theta_1,\ldots,\theta_{k-1},\theta_{k+1},\ldots,\theta_K | X,\theta_k$.

Condition (i) restricts the support of $\theta | X = x$ to be constant in $x$. While this does not appear to be necessary for identification it is used in explicitly in the proof of the main result and it would also be necessary to identify certain average effects of $X$ on $Y$. The remaining conditions in Assumption 2.5 impose conditions on the model that are extensions of standard normalizations in linear factor models. They could be replaced by different normalizations, also discussed below in Section 4.

If $\varepsilon_j \perp \perp X,\theta$ for each $j$ then Assumptions 2.2(iii) and 2.5(iii) can be stated solely as exclusion restrictions on the $h_j$ functions, though this assumption is not imposed as it is not necessary for identification. In fact, what the identification argument uses are restrictions on the conditional probabilities, $p_j(x,t) := Pr(M_j = 1 | X = x, \theta = t)$. Let $M_1,\ldots,M_J$ denote the binary indicators not used to satisfy Assumption 2.2(iii) or Assumption 2.5(iii), after relabeling. Then let $M_{E1},\ldots,M_{EK}$ and $M_{N1},\ldots,M_{NK}$ and denote those used to satisfy
Assumptions 2.2(iii) and 2.5(iii), respectively. Note that these latter two subsets of the $J$ binary variables may overlap but may not overlap with the other $J$ indicators. Define $p_{Nk}(x, t_k) := P(M_{Nk} = 1 \mid X = x, \theta_k = t_k)$ and $p_{Ek}(t) := P(M_{Ek} = 1 \mid X = x)$, and let $p_{E}(x, t)$ denote the vector $(p_{E1}(x, t), \ldots, p_{EK}(x, t))$. With this notation, condition 2.5(iii) implies that $E(M_{Nk} \mid X = x, \theta = t) = p_{Nk}(x, t_k)$ for each $k$ and condition 2.2(iii) implies that $E(M_{Ek} \mid X = x, \theta = t) = p_{Ek}(t)$ for each $k$.

Several remarks about Assumption 2.5 are in order. First, conditions (ii) and (iii) represent a natural analog to normalizations imposed in a linear factor model. Suppose $M_j = \alpha_j \theta + \varepsilon_j$. In this model normalizations are required because $\theta$ and $\alpha_j$ can be redefined as $\tilde{\theta} = P\theta$ and $\tilde{\alpha}_j = \alpha_j P^{-1}$ for any invertible matrix $P$ without changing the joint distribution of the $M_j$. One solution to this non-identification result is to impose that for the first $K$ values of the index $j$, $\alpha_j = e_j'$ where $e_j$ is a vector with a one in the $j^{th}$ entry and zeros elsewhere and that $Var(\theta_k) = 1$ for each $k$ (Anderson and Rubin, 1956). This normalization, and the generalization imposed here in Assumption 2.5 for a nonparametric binary model, are appealing in many cases because they do not require independence between the different dimensions of $\theta$. An alternative normalization common in linear factor models is to assume that $Var(\theta) = I_k$, so that the $\theta_k$ are uncorrelated across $k$, and that the matrix $\alpha = (\alpha'_1, \alpha'_2, \ldots)'$ is lower triangular. An extension of these normalizations for the model of this paper is discussed in Section 4. Choosing between alternative normalizations can be aided in some cases by economic theory, as in the discussion of Example 2 below. In an additively separable model, Cunha and Heckman (2008) discuss the use of outcomes to “anchor” the scale of $\theta$. This idea is also appealing but we do not pursue it here.

It should also be noted that the number of restrictions is relatively small. There are restrictions on at most $2K$ response functions (conditional expectation functions of the binary indicators) but the number of distinct functions must diverge in the identification argument while $K$ stays fixed. I do not require an increasing number of restrictions as $J$ diverges.

Example 1, cont’d. In Example 1, introduced above, the support condition, 2.5(i), could be problematic if very high or very low ability levels perfectly predict education status. Condition (ii) is innocuous as intelligence and personality are abstract constructs with no natural scale. Condition (iii) requires more discussion. This condition is satisfied if one item on the
personality survey does not depend on cognitive ability and one item on the achievement test is not affected by differences in noncognitive traits. Generally, performance on IQ tests or achievement tests such as the Armed Forces Qualifying Test (AFQT) depends on test motivation and other noncognitive traits (see, e.g., Duckworth, Quinn, Lynam, Loeber, and Stouthamer-Loeber, 2011; Heckman and Kautz, 2012), and it is possible that responses to a personality survey may depend on cognitive ability even conditional on the noncognitive ability. Assumption 2.5(iii) assumes that only one item on the test, or one additional item collected separately, does not depend on test motivation. Therefore, it is only necessary to prevent unwanted effects of test motivation on one question, not the whole test.

Example 2, cont’d. In Example 2 above, the conditions of Assumption 2.5 can be assessed in terms of the economic model of investment in children’s human capital. Suppose parents choose $\tilde{\theta}$ and consumption $q$ to maximize $U(q, \tilde{\theta}; \psi_U)$ subject to $q + c(\tilde{\theta}; X, \psi_c) = \omega$ where $\omega$ is the household income, $c(\tilde{\theta}; X, \psi_c)$ is the cost function produced by the first stage cost minimization, and $\psi_U$ and $\psi_C$ are preference parameters and cost factors, some of which may vary across individuals. This induces variation in investment, $\tilde{\theta} = \tilde{\theta}^*(X, \psi_c, \omega, \psi_U)$. I can define $\theta = F_{\tilde{\theta}}(\tilde{\theta})$ to satisfy 2.5(ii). The actual scale of investment cannot be inferred, only the rank. Condition (i) of Assumption 2.5 will often be satisfied because for a wide class of technologies and utility functions the support of $\tilde{\theta} \mid X = x$ will be $[0, \infty)$ if the support of $\omega$ is $[0, \infty)$, regardless of the value of $x$. Introduction of additional dimensions of $\theta$ to account for heterogeneity in $\psi_C$ will require exclusions described in 2.5(iii). These can similarly be justified based on properties of the production function.

In addition to Assumptions 2.2(iii) and Assumption 2.5, the ability of the binary variables, $M_j$, to serve as proxies for $\theta$ depends on the following invertibility condition. Divide $M_1, \ldots, M_J$ into $K$ distinct subsets and define $\bar{M}_k$ as the average of the $J_k$ variables in the $k^{th}$ subset. Define $\bar{p}_k(x, t) := E(\bar{M}_k \mid X = x, \theta = t)$ and $\bar{p}(x, t) = (\bar{p}_1(x, t), \ldots, \bar{p}_K(x, t))$. The identification argument will use the fact that $\bar{M} := (\bar{M}_1, \ldots, \bar{M}_K)$ converges in probability to $\bar{p}(X, \theta)$ as $J \to \infty$. In order to take advantage of this result I impose the following invertibility conditions.

**Assumption 2.6**
(i) For each $x \in \mathcal{X}$, $\bar{p}(x, \cdot)$ is injective.

(ii) $p_E(x, \cdot)$ is injective.

(iii) For each $k$ and each $x \in \mathcal{X}$ the function $p_{Nk}(x, \cdot)$ is strictly increasing.

If Assumption 2.6(i) is not satisfied then variation in $\bar{M}$ will not be rich enough to fully uncover variation in $\theta$, even as $J \to \infty$. Another function, $f(M_1, \ldots, M_J)$ where $f : \{0, 1\}^J \to [0, 1]^K$, could potentially serve as a proxy as well. For example, $f$ could be produced by a (possibly misspecified) likelihood optimization, $\arg \min \ell_J(\theta)$. While this could offer improved efficiency, it does not facilitate the identification argument in any way.

The advantage of using $\bar{M}$, on the other hand, is that $E(\bar{M} \mid X, \theta) = \bar{p}(X, \theta)$ and properties of the function $\bar{p}(x, t)$ can be linked to the structural model in equation (2.2).

Using condition 2.2(iii) to separate variation in $X$ from variation in $\theta$ will only work if 2.6(ii) holds. Similarly condition (iii) of Assumption 2.6 is needed to facilitate the separation of different components of $\theta$ enabled by Assumption 2.5(iii). The role of these conditions will become apparent in the next section.

The invertibility conditions in 2.6(i) and 2.6(ii) can be motivated by lower level conditions. For example, conditions 2.5(iii) and 2.6(iii) together impose invertibility on the system $p_N = (p_{N1}, \ldots, p_{NK})$ so the vector-valued functions $\bar{p}$ and $p_E$ could be restricted in the same way to satisfy Assumptions 2.6(i) and 2.6(ii). Typically those restrictions would be difficult to justify, especially for the function $\bar{p}$. Alternatively, one could appeal to the classic invertibility results of Gale and Nikaido (1965) if the relevant functions are differentiable. Gale and Nikaido (1965) show that if all principal submatrices of the Jacobian of the function $F : [0, 1]^K \to \mathbb{R}^K$ have positive determinants everywhere on the support then $F$ is injective. A lower level condition for invertibility that does not require differentiability is that the function is inverse isotone, i.e., $F(t) \leq F(t')$ implies $t \leq t'$ where $\leq$ here denotes the component-wise inequality (Rheinboldt, 1970).\footnote{The inverse isotone condition has been used, for example, in the context of inverting demand systems Berry, Gandhi, and Haile (2013).}

If $\varepsilon_j$ is independent of $X, \theta$ and has a continuous and strictly increasing distribution function then, from equation (2.2), 2.6(ii) follows from the invertibility of $h_E(x, t) = (h_{E1}(x, t), \ldots, h_{EK}(x, t))$.

This is useful if the functions $h_{Ek}$ are derived from an explicit economic model, as in Example 2 discussed in Section 2. Invertibility of $\bar{p}$ can be derived from conditions on the individual conditional expectation functions, $p_j$, using either the Jacobian condition of Gale and
Nikaido (1965) or the inverse isotope condition, as demonstrated by the following result.

**Proposition 2.1**

(i) Suppose that for any $K$-tuplet of binary proxies, $M_{j1}, M_{j2}, \ldots, M_{jK}$ such that for each $k$ $M_{jk}$ belongs to the subset of variables that is averaged to produce $\bar{M}_k$, the corresponding conditional expectation functions, $(p_{j1}(x, \cdot), \ldots, p_{jK}(x, \cdot))$, together form a function that is differentiable with a Jacobian matrix with all principal submatrices having positive determinants. Then $\bar{p}$ is injective.

(ii) Suppose that for any $K$-tuplet of binary proxies, $M_{j1}, M_{j2}, \ldots, M_{jK}$ such that for each $k$ $M_{jk}$ belongs to the subset of variables that is averaged to produce $\bar{M}_k$, the corresponding conditional expectation functions, $(p_{j1}(x, \cdot), \ldots, p_{jK}(x, \cdot))$, together form a function that is inverse isotone. Then $\bar{p}$ is injective.

This result will not be useful however when different indicators are relevant only for different regions of the distribution of $\theta$. For example, some questions on a test may be easy enough that they are answered correctly with certainty for every individual above a certain level of ability. If there are enough questions on the test of various difficulty levels then $\bar{p}(x, \cdot)$ will be invertible but some functions $p_j(x, \cdot)$ may not be.

**Example 1, cont’d.** Suppose the test items and survey items together are divided into two groups. While condition (iii) of Assumption 2.5 requires only one item which does not depend on $\theta^N$ and one which does not depend on $\theta^C$, condition (i) of Assumption 2.6 will only be satisfied if there is sufficient information about both dimensions. For example, suppose that, as is often the case, the personality survey consists of a small number of items. Then since the number of indicators contributing to each of the two averages, $\bar{M}_1$ and $\bar{M}_2$, must be large (see Assumption 2.7(v) below), condition (iii) of Assumption 2.6 will only be satisfied if performance on some test questions also depends on $\theta^N$. Moreover, the two groups of items must be different functions of the two traits. One way to ensure that the Jacobian has nonzero determinant is if items in one group are more affected by $\theta^N$ than items in the other group.

Condition (ii) of Assumption 2.6 suggests that the exclusion restrictions of 2.2(iii) cannot be on any two items but instead these two items must provide information about separate
dimensions of the latent ability distribution. The most obvious way to satisfy this is to use a test item and an item from the personality survey. Assumption 2.6(iii) requires the items used for normalization 2.5(iii) to discriminate between individuals at all positions in the distribution of the corresponding ability.

Regularity conditions

Finally, I impose the following additional regularity conditions which play an important role since the identification analysis is a limit result as \( J \to \infty \).

Assumption 2.7

(i) The conditional distribution function of \( \theta \mid X \), denoted \( F_{\theta \mid X}(t \mid x) \), satisfies

\[
\max_k \sup_{x \in X, t_k \in [0,1], t'_k \in [0,1], t'_k \neq t_k} \frac{|F_{\theta_k \mid X}(t'_k \mid x) - F_{\theta_k \mid X}(t_k \mid x)|}{|t'_k - t_k|^{\alpha_F}} < M_F
\]

and

\[
\min_k \inf_{x \in X, t \in [0,1]} \frac{|F_{\theta \mid X}(t_k, t_k - k \mid x) - F_{\theta \mid X}(t_k, t_k - k \mid x)|}{|t'_k - t_k|^{\beta_F}} > \epsilon_F
\]

(ii) \( \bar{p} \) satisfies

\[
\sup_{x \in X, t \in [0,1], t' \in [0,1], t' \neq t} \frac{||\bar{p}(x, t') - \bar{p}(x, t)||}{||t' - t||^{\alpha_\beta}} < M_{\bar{p}}
\]

and

\[
\inf_{x \in X, t \in [0,1], t' \in [0,1], t' \neq t} \frac{||\bar{p}(x, t') - \bar{p}(x, t)||}{||t' - t||^{\beta_\beta}} > \epsilon_{\bar{p}}
\]

(iii) \( p_E \) satisfies

\[
\sup_{t \in [0,1], t' \in [0,1], t' \neq t} \frac{||p_E(t') - p_E(t)||}{||t' - t||^{\alpha_E}} < M_E
\]

and

\[
\inf_{t \in [0,1], t' \in [0,1], t' \neq t} \frac{||p_E(t') - p_E(t)||}{||t' - t||^{\beta_E}} > \epsilon_E
\]

(iv) \( p_N \) satisfies

\[
\sup_{x \in X, t \in [0,1], t' \in [0,1], t' \neq t} \frac{||p_N(x, t') - p_N(x, t)||}{||t' - t||^{\alpha_N}} < M_N
\]

and

\[
\min_{k=1, \ldots, K} \inf_{x \in X, t_k \in [0,1], t'_k \in [0,1], t'_k \neq t_k} \frac{|p_{N_k}(x, t'_k) - p_{N_k}(x, t_k)|}{|t'_k - t_k|^{\beta_N}} > \epsilon_N
\]
The sizes of the $K$ subgroups satisfy the limiting condition $J_k^{-1} = O(J^{-1})$ for each $k$. 

In addition, depending on which of the three structures discussed above is used for the outcome equation I need additional regularity conditions. Under Assumption 2.3 I can define an inverse function for each $x, t$, $h(x, t, y) := g^{-1}(y; x, t)$ by setting $h(x, t, y) = 1$ for all $y > g(x, t, 1)$ and $h(x, t, y) = 0$ for $y < g(x, t, 0)$. I assume that the functions $g$ and $h$ satisfy the following.

**Assumption 2.8** $g(x, t, u)$ satisfies

$$\sup_{x \in \mathcal{X}, t \in [0,1]^K, u, u' \in [0,1], u' \neq u} \frac{|g(x, t, u') - g(x, t, u)|}{|u' - u|^\alpha_g} < M_g$$

and $h(x, t, y) := g^{-1}(y; x, t)$ satisfies

$$\sup_{x \in \mathcal{X}, t \in [0,1]^K, t', y \in g(x, t', [0,1] \cap g(x, t, [0,1]) \cap g(x, t', [0,1]) \cap g(x, t, [0,1]) \cap g(x, t', [0,1]) \cap g(x, t, [0,1])} \frac{|h(x, t', y) - h(x, t, y)|}{||t' - t||^{\alpha_h}} < M_h$$

Under the binary outcome model of Assumption 2.4 the following regularity condition is used.

**Assumption 2.9** $h(x, t)$ satisfies the Holder condition

$$\sup_{x \in \mathcal{X}, t \in [0,1]^K, t', s \in [0,1]^K, t' \neq s} \frac{|h(x, t') - h(x, t)|}{||t' - t||^{\alpha_h}} < M_h$$

If neither Assumption 2.3 nor Assumption 2.4 holds then identification of $\bar{g}(x, t)$ uses the following regularity condition.

**Assumption 2.10** $\bar{g}(x, t)$ satisfies the Holder condition

$$\sup_{x \in \mathcal{X}, t \in [0,1]^K, t', s \in [0,1]^K, t' \neq s} \frac{|\bar{g}(x, t') - \bar{g}(x, t)|}{||t' - t||^{\alpha_h}} < M_h$$

These regularity conditions are used to control the size of the identified set as $J$ increases. There are two types of conditions in (i)-(iv). The first type is a Holder continuity condition. The second type of condition can be thought of as a Holder monotonicity condition. It could instead have been stated as Holder continuity on the inverse functions, which exist by
Assumption 2.6. The Holder continuity condition is stronger than assuming that the function is continuous and the Holder monotonicity condition similarly strengthens monotonicity (already imposed in Assumption M). These conditions ensure that the parameter space is compact.

Consider $f : [0, 1] \to [0, 1]$. If the function $f$ is differentiable on $[0, 1]$ then it is Holder continuous if the derivative is bounded and it is Holder monotonic if the derivative is bounded away from 0. The regularity conditions above could all be similarly stated in terms of bounded derivatives. As stated however these conditions allow derivatives to be 0 or unbounded at a point provided that they converge to 0 or $\infty$ at a slow enough rate.

In interpreting these regularity conditions it is important to keep in mind the normalizations, $\theta_k \sim Uniform(0, 1)$. Suppose, for example, that $\theta$ is scalar and for each $j \ M_j = 1(\mu_j + \beta_j' X + \alpha_j \tilde{\theta} \geq \varepsilon_j)$ where $\tilde{\theta} \sim F_{\tilde{\theta}}$ and $\varepsilon_j \sim F_{\varepsilon_j}$. Then $\theta := F_{\tilde{\theta}}(\tilde{\theta})$ is uniformly distributed and therefore, $p_j(x, t) = E(M_j | X = x, \theta = t) = F_{\varepsilon_j}(\mu_j + \beta_j' X + \alpha_j F_{\tilde{\theta}}^{-1}(t))$. Then

$$\frac{\partial p_j(x, t)}{\partial t} = \frac{f_{\varepsilon_j}(\mu_j + \beta_j' X + \alpha_j F_{\tilde{\theta}}^{-1}(t))}{f_{\tilde{\theta}}(F_{\tilde{\theta}}^{-1}(t))}$$

If the densities of $\tilde{\theta}$ and $\varepsilon_j$ are bounded and non-zero on $\mathbb{R}$ then $\frac{\partial p_j(x, t)}{\partial t}$ is defined for all $t \in (0, 1)$ but not necessarily for $t \in [0, 1]$. If $\tilde{\theta}$ has a bounded support, and $f_{\tilde{\theta}}$ is bounded away from 0 on this support, then this also holds on $[0, 1]$. If $\tilde{\theta}$ does not have bounded support, the regularity conditions are essentially controlling the relative tail size of $\tilde{\theta}$ and the relevant idiosyncratic error terms. See Remarks 2 and 3 following Theorem 3.2 for more discussion.

3. IDENTIFICATION

In this section I discuss identification of the model described in the previous section. The model is not nonparametrically identified for a fixed $J$ so instead I will consider an asymptotic analysis where $J \to \infty$ and show that the distance between any two observationally equivalent models converges to 0 in the limit. I focus on identification of the “structural” function — $g(x, t, u)$ under Assumption 2.3, $h(x, t)$ under Assumption 2.4, or $\tilde{g}(x, t)$ if neither 2.3 nor 2.4 holds.

To demonstrate how the binary indicators can serve as a proxy for $\theta$, enabling identification in the limit, I first study identification of a very simple version of the model in Section 3.1.
The main identification result is in Section 3.2.

### 3.1. Preliminary Result

In this section I look at a stripped down model to gain some intuition for the main results, which will be presented in Section 3.2. Suppose that

\[ Y = 1(h_Y(\theta) \geq U) \]
\[ M_j = 1(h_j(\theta) \geq \varepsilon_j), \]

for \( j = 1, \ldots, J \). This is essentially an item response model (Lord, 1980). Consider the following set of assumptions.

**Assumption 3.1**

(i) \( U \sim \text{Uniform}(0, 1) \)

(ii) \( U, \varepsilon_1, \ldots, \varepsilon_J \perp \perp \theta \)

(iii) \( \theta \sim \text{Uniform}(0, 1) \)

(iv) \( U, \varepsilon_1, \ldots, \varepsilon_J \) mutually conditionally independent given \( \theta \).

(v) Each \( \varepsilon_j \) has a continuous distribution function, \( F_{\varepsilon_j} \), and \( \bar{p}(t) := J^{-1} \sum_{j=1}^{J} p_j(t) \) is a strictly increasing function where \( p_j(t) := F_{\varepsilon_j}(h_j(t)) \).

This set of conditions essentially restates the relevant parts of Assumptions 2.1, 2.2, 2.4, 2.5, and 2.6 above for the model of equation (3.1) with scalar \( \theta \). By conditions (ii) and (v), \( \Pr(M_j = 1 \mid \theta) = p_j(\theta) \) and by conditions (i) and (ii) \( \Pr(Y = 1 \mid \theta) = h_Y(\theta) \). The joint distribution of \((Y, M_1, \ldots, M_J)\) can be written as

\[ \Pr(Y = y, M_1 = m_1, \ldots, M_J = m_J) = \int_{\mathbb{R}} h_Y(t)^y(1 - h_Y(t))^{1-y} \prod_{j: m_j = 1} p_j(t) \prod_{j: m_j = 0} (1 - p_j(t)) dt \]

This is a continuous mixture model. Since the unknown parameters, namely the vector of functions \((h_Y, p_1, \ldots, p_J)\), lie in an infinite-dimensional space the model is fundamentally unidentified.
3.1.1. The Identified Set

The object of interest in my analysis is the function $h_Y(\cdot)$. The distribution of the observed data, $(Y, M_1, \ldots, M_J)$, is characterized by the $2^{J+1} - 1$ moment conditions provided by equation (3.2) for $(y, m_1, \ldots, m_J) \in \{0, 1\}^{J+1}$. Therefore, the identified set consists of all functions $h_Y$ that satisfy each of these $2^{J+1} - 1$ moment conditions for some set of nuisance parameters $p_1, \ldots, p_J$. Without any additional structure a simpler way to construct the identified set is not apparent.

To visualize the identified set I can consider the identified set for $h_Y(x)$ for a fixed point $x \in [0, 1]$ or the identified set for $h_Y(x') - h_Y(x)$ for each pair of points $x, x' \in [0, 1]$. Several examples are provided in Williams (2013). It is apparent from the examples and accompanying discussion in Williams (2013) that the identified set is nontrivial but typically large when only a few items are available. It is also apparent that the identified set shrinks as $J$ increases and it appears to do so at a faster rate depending on certain properties of the model. In the next section I provide conditions under which the identified set converges to a point as $J \to \infty$ and show that the rate of this convergence can be characterized by certain properties of the model.

3.1.2. Identification as $J \to \infty$

Define the constant $B$ to be the smallest value $B^*$ such that if $h_Y$ and $h'_Y$ are any two functions in the identified set $\sup_{x \in [0, 1]} |h_Y(x) - h'_Y(x)| \leq B^*$. Since this is trivially true for $B^* = 1$ the constant $B$ must exist and, as demonstrated in Williams (2013) this constant $B$ is nontrivial (i.e., strictly less than 1) in many cases. In this section I will provide conditions under which $B$ converges to 0 as $J \to \infty$ and provide an upper bound on the rate of this convergence.

Let $\bar{M} = J^{-1} \sum_{j=1}^J M_j$ and define the conditional mean function $\bar{p}(t) := J^{-1} \sum_{j=1}^J p_j(t) = E(\bar{M} \mid \theta = t)$. The conditional independence in Assumption 3.1 implies that $\sup_{t \in [0, 1]} Var(\bar{M} \mid \theta = t) \to 0$ as $J \to \infty$. As a result, $\bar{M}$ converges in probability to its conditional mean $\bar{p}(\theta)$,\footnote{By Chebyshev’s inequality $Pr(|\bar{M} - E(\bar{M} \mid \theta)| > \delta \mid \theta) \leq Var(\bar{M} \mid \theta)/\delta^2$. Therefore, since $E(\bar{M} \mid \theta) = \bar{p}(\theta)$, it follows that $Pr(|\bar{M} - \bar{p}(\theta)| > \delta) = \int_0^1 Pr(|\bar{M} - \bar{p}(\theta)| > \delta \mid \theta = t) dt \leq \delta^{-2} \sup_{t \in [0, 1]} Var(\bar{M} \mid \theta = t) \to 0$.} Convergence in probability implies convergence in distribution, that is, $|F_{\bar{M}}(s) - F_{\bar{p}(\theta)}(s)| \to 0$ where $F_{\bar{M}}(s) := Pr(\bar{M} \leq s)$ and $F_{\bar{p}(\theta)}(s) := Pr(\bar{p}(\theta) \leq s)$.

This argument implies that the function $F_{\bar{p}(\theta)}$ is identified in the sense that if $F_{\bar{p}(\theta)}'$ is
observationally equivalent then \(|F_p(\theta) - F_p(\theta)'| \to 0\) But \(\theta \sim Uniform(0,1)\) and Assumption 3.1(iii) implies that the function \(\tilde{p}\) is strictly increasing, and hence invertible\(^{13}\) and therefore \(F_{\tilde{p}(\theta)}(s) = \tilde{p}^{-1}(s)\) which suggests that the function \(\tilde{p}\) is also identified in this same limiting sense. Another way to view this result is that \(\tilde{p}(t)\) is identified in the limit from the \(t^{th}\) quantile of \(\tilde{M}\). These arguments require that \(\tilde{p}\) and \(\tilde{p}^{-1}\) are continuous. In fact, if conditions (i) and (iii) in the statement of Theorem 3.1 below are satisfied then the convergence is uniform in \(s\) (and \(u\), respectively). Conditions (i) and (iii) in the statement of Theorem 3.1 below are slightly stronger than continuity of each \(\tilde{p}\) and \(\tilde{p}^{-1}\) because a limit of continuous functions can be discontinuous.

Next consider \(E(Y \mid \tilde{M} = s)\). Under conditions (i) and (ii) of Assumption 3.1,

\[
E(Y \mid \tilde{M} = s) = E(h_Y(\theta) \mid \tilde{M} = s) = \int_0^1 h_Y(t)f_{\theta|M}(t \mid s)dt
\]

The argument above that \(\tilde{M}\) converges in probability to \(\tilde{p}(\theta)\) can be used to show that, in a certain sense, \(\theta \mid \tilde{M} = s\) converges in distribution to \(\theta \mid \tilde{p}(\theta) = s\), which is a point mass at \(\tilde{p}^{-1}(s)\) since \(\tilde{p}\) is injective. If \(h_Y\) and \(\tilde{p}^{-1}\) are smooth enough then this implies that the integral above converges to \(h_Y(\tilde{p}^{-1}(s))\) as \(J \to \infty\). Since the function \(\tilde{p}\) is identified for large \(J\) from the distribution of \(\tilde{M}\) the function \(h_Y\) is identified in the limit as well.

The formal argument in the appendix uses the decomposition

\[
(3.3) \quad \int_0^1 h_Y(t)f_{\theta|M}(t \mid s)dt = \int_{t:|\tilde{p}(t) - m|<\epsilon} h_Y(t)f_{\theta|M}(t \mid s)dt + \int_{t:|\tilde{p}(t) - s|\geq\epsilon} h_Y(t)f_{\theta|M}(t \mid s)dt
\]

The first term is small for small \(\epsilon\) if \(h_Y \circ \tilde{p}^{-1}\) is continuous. The second term is small as \(J \to \infty\) according to the argument above that \(\tilde{M}\) converges to \(\tilde{p}(\theta)\) since it is bounded by \(Pr(|\tilde{p}(\theta) - s| \geq \epsilon \mid \tilde{M} = s)\). The rate of convergence as \(J \to \infty\) depends on the smoothness of the functions \(h_Y\) and \(\tilde{p}^{-1}\).

A similar argument was first formalized by Douglas (2001). The implication that \(h_Y(t)\) can be nonparametrically estimated by a kernel regression estimator of the conditional expectation function \(E(Y \mid F_M(\tilde{M}) = t)\) has been exploited by Ramsay (1991) and Douglas (1997). The following theorem extends a theorem of Douglas (2001) to allow for a broader class of

\(^{13}\)That is, there is an inverse function defined uniquely on the image of \(\tilde{p}\).
conditional expectation functions and by providing a rate that depends on the smoothness of these functions.

**Theorem 3.1** If (i) \( \sup_{t \neq t' \in [0,1]} \frac{|h_Y(t') - h_Y(t)|}{|t - t'|^{\alpha h}} < M_h \), (ii) \( \inf_{t \neq t' \in [0,1]} \frac{|\bar{p}(t') - \bar{p}(t)|}{|t - t'|^{\beta}} > \epsilon \) and (iii) \( \frac{|\bar{p}(t') - \bar{p}(t)|}{|t - t'|^{\alpha p}} < M_p \) then

\[
B = O \left( (\log J/J)^{\frac{1}{2} \alpha h \beta^{-1}} \right)
\]

The complete proof of this theorem is provided in the appendix. Note that the best possible rate is slightly slower than \( J^{-1/2} \). The rate is slower if \( \bar{p}(t) \) is too flat, which makes it difficult to distinguish between close values of \( \theta \), or if \( h_Y \) is too steep. The latter is not surprising because, even if \( \theta \) were observed, the convergence rate of an estimator of \( h_Y \) would depend on the smoothness of the function. Convergence of the identified set can be proven under alternative conditions as well. Note for example that the conditions of the theorem rule out a model with degenerate measurement error \( (p_j(t) = 1(t > c_j) \) for a constant \( c_j \)). See Williams (2013) for a discussion.

### 3.2. Main Identification Result

The main goal of this paper is identification of the outcome equation (2.1) using item response data generated according to equation (2.2). As the discussion in Section 3.1 suggests, though there is identifying information in the model for a finite \( J \), point identification is only possible in the limit as \( J \to \infty \). There are essentially three obstacles to identification: obtaining continuous variation in \( \theta \), being able to separately vary \( \theta \) and the observed covariates \( X \), and finding normalizations that will allow the scale of \( \theta \) to be determined and, in the case that \( \theta \) is multidimensional, finding further restrictions that enable separately varying the different dimensions. The large \( J \) asymptotics will overcome the first obstacle to identification. I show this by extending the argument, originally due to Douglas (2001), which was laid out above in Section 3.1. The idea is to use averages of the items to proxy \( \theta \). These sample averages will exhibit continuous variation in the limit as \( J \to \infty \), though this will reflect variation in \( X \) as well as \( \theta \). The second identification problem will be solved through the exclusion restrictions of Assumption 2.2(iii). I do not require any of the components of \( X \) to be excluded from the outcome equation. Lastly, the normalizations used to fix the scale of the multidimensional factor \( \theta \), in Assumption 2.5(iii), are reminiscent of conditions in factor models dating back to Anderson and Rubin (1956).
3.2.1. Outline of Identification

In this section I will lay out a sketch of the identification argument. I provide a formal statement of the main result in the following section and a proof in the appendix.

Recall that $\theta$ denotes the dimension of $\theta$, the $J$ responses (excluding those used to satisfy 2.2(iii) and 2.5(iii)) are divided into $K$ exhaustive and mutually exclusive groups, and $\bar{M}$ is the vector of averages of the responses in each group. By Assumption 2.6(i) the vector-valued conditional mean function $\bar{p}(x,t) = E(M \mid X = x, \theta = t)$ is invertible in $t$ for each $x$. By an extension of the argument above in Section 3.1.2, $\bar{M} \rightarrow p(X, \theta)$.

Under Assumption 2.3, $Pr(Y \leq y \mid \bar{M} = m, X = x) = E(g^{-1}(y; x, \bar{p}) \mid \bar{M} = m, X = x)$. Given the invertibility of $\bar{p}(x, \cdot)$, the argument used in equation (3.3) can be modified to show that

$$Pr(Y \leq y \mid \bar{M} = m, X = x) \approx g^{-1}(y; x, \bar{p}^{-1}(m; x))$$

Similarly, for each $k$

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

where $\bar{p}^{-1}(m; x) = (\bar{p}^{-1}_{(1)}(m; x), \ldots, \bar{p}^{-1}_{(K)}(m; x))$ and

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

To this point I am essentially modifying the argument of Section 3.1 — an argument adapted from Douglas (2001) — to allow for multivariate $\theta$ by using the $K$-dimensional proxy $\bar{M}$. The rest of the argument is entirely original to this paper.

His is as far as the existing results (Douglas, 2001) get us.\textsuperscript{14}

To provide a clear sketch of the main result I will now demonstrate how $g(x, t, u)$ is identified if the objects on the right-hand side of these three approximations are known. Since $g^{-1}(y; x, p^{-1}(m; x))$ and $p_{E}(\bar{p}^{-1}(m; x))$ are both known and since, by Assumption 2.6(ii), the vector-valued function $p_{E}$ is invertible with inverse $p_{E}^{-1}$, the function $g^{-1}(y; x, p_{E}^{-1}(m))$ is identified. Inverting this identifies the object of interest but defined on an unknown scale, $g(x, p_{E}^{-1}(m), u)$. That is, the endogeneity problem is solved because $g(x, p_{E}^{-1}(m), u) - g(x, p_{E}^{-1}(m), u)$ is a causal effect. This argument can be formalized to show, for example,

\textsuperscript{14}In fact this is already a slight modification of to allow multiple dimensions of $\theta$. 

that the average structural function, \( E(g(x, \theta, U)) \), is identified (Williams, 2013). However neither the distribution of such effects nor the effect at a particular value of \( \theta \) is identified.

To fully identify the function \( g(x, t, u) \) I must further show that \( p_E(t) \) is identified. First, the function \( T_k(m) := p_{Nk}(x, \varphi_{(k)}^{-1}(m; x)) \) is assumed to be known for each \( k = 1, \ldots, K \). Consider the distribution of \( T_k(M) \). Since \( M \approx \bar{p}(X, \theta) \), for large \( J \),

\[
Pr(T_k(M) \leq \tau_k | X = x) \approx Pr(T_k(\bar{p}(x, \theta)) \leq \tau_k | X = x) \\
= Pr(p_{Nk}(x, \theta) \leq \tau_k | X = x) \\
= F_{\theta|x}(p_{Nk}^{-1}(\tau_k; x) | x)
\]

Define the function \( \bar{p}_E(x, m) := p_E(\bar{p}^{-1}(m; x)) \) and let \( \bar{p}_E^{-1}(m; x) \) denote its inverse which is guaranteed to exist by condition (iii) of Assumption 2.6. Then \( Pr(T_k(\bar{M}) \leq T_k(\bar{p}_E^{-1}(x, m)) | X = x) \approx F_{\theta|x}(p_{E,(k)}^{-1}(m) | x) \) where \( p_{E,(k)}^{-1}(m) \) is the \( k \)th element of \( p_E^{-1}(m) \). Integrating with respect to the distribution of \( X \) and using the assumption that the unconditional distribution of \( \theta_k \) is \( Uniform(0, 1) \) (Assumption 2.5(ii)),

\[
\int_X F_{\theta|x}(p_{E,(k)}^{-1}(m) | x) dF_X(x) = F_{\theta|x}(p_{E,(k)}^{-1}(m)) \\
= p_{E,(k)}^{-1}(m)
\]

Applying this argument for each \( k \) and again ignoring the approximation error, \( p_E^{-1}(m) \) is identified and combining this with \( g(x, \bar{p}_E^{-1}(m, u) \) shows that the function \( g(x, t, u) \) is identified for each \( x, t, \) and \( u \). Identification of \( h(x, t) \) under Assumption 2.4 and identification of \( \bar{g}(x, t) \) follow the same argument because (i) under 2.4 \( Pr(Y = 1 | \bar{M} = m, X = x) = E(h(x, \theta) | \bar{M} = m, X = x) \) and (ii) \( E(Y | \bar{M} = m, X = x) = E(\bar{g}(x, \theta) | \bar{M} = m, X = x) \). And therefore

\[
Pr(Y = 1 | \bar{M} = m, X = x) \approx h(x, \bar{p}^{-1}(m, x))
\]

and

\[
E(Y | \bar{M} = m, X = x) \approx \bar{g}(x, \bar{p}^{-1}(m, x))
\]

The next section will provide the formal identification result for \( g(x, t, u) \), \( h(x, t) \), and \( \bar{g}(x, t) \) which accounts for the several approximations in the above argument.
3.2.2. Formal Result

As in Section 3.1, I wish to show that the identified set for the “structural function” collapses to a single point as $J \to \infty$. Since the focus of this paper is on this limiting result it is unnecessary to introduce additional cumbersome notation for the identified set and will instead define a concept of “large $J$ identification” directly. To do so, I focus for the moment on identification of the function $\bar{g}(x, t)$. The following definitions are easily altered in the other two cases where identification of $g(x, t, u)$ is studied under Assumption 2.3 and identification of $h(x, t)$ is studied under Assumption 2.4.

Let $\bar{g}^*(x, t)$ be a function satisfying Assumption 2.10 and define $Q_J(\bar{g}^*(x, t))$ as the class of possible probability distributions for $D = (Y, X, M_1, \ldots, M_J, M_E, M_N)$ if $D$ is generated according to the model of the previous section (without Assumptions 2.3, 2.4, 2.8, and 2.9) as the underlying primitives vary but the model parameter $\bar{g}(x, t)$ is fixed at the particular function $\bar{g}^*(x, t)$.

Now I can define the familiar concept of observational equivalence. Observational equivalence is defined relative to what data are assumed to be observed. The observed data depends on the number of measurements, $J$. If $J = 1$, $D = (Y, X, M_1, M_E, M_N)$; if $J = 3$, $D = (Y, X, M_1, M_2, M_3, M_E, M_N)$; etc. The $J$ subscript in the notation $Q_J(\bar{g}^*(x, t))$ is meant to emphasize this. The dimension of the space $Q_J(\bar{g}^*(x, t))$ is increasing in $J$. This results in the following definition.

**Definition 1** $\bar{g}^*(x, t)$ and $\bar{g}'(x, t)$ are $J$-observationally equivalent if

$$Q_J(\bar{g}^*(x, t)) \cap Q_J(\bar{g}'(x, t)) \neq \emptyset$$

The traditional notion of identification is that if $\bar{g}^*(x, t)$ and $\bar{g}'(x, t)$ are observationally equivalent then they must be equal (or at least equal with probability 1). Instead I will use a notion of identification that requires only that as $J \to \infty$, the distance between any observationally equivalent parameters collapses to 0.

**Definition 2** The model parameter $\bar{g}(x, t)$ is large $J$ identified at the rate $r_J$ if there exists a sequence $s_1, \ldots, s_J, \ldots$ such that
(i) for any $\bar{g}^*(x,t)$ and $\bar{g}'(x,t)$ that are $J$-observationally equivalent

$$\sup_{x,t} |\bar{g}^*(x,t) - \bar{g}'(x,t)| \leq s_J,$$

(ii) $\lim_{J \to \infty} s_J = 0$,

(iii) and $s_J = O(r_J)$

This is closely related to notions of identification used in panel data and in the weak IV literature though in those cases models are typically parametric and identification is not considered separately from consistency of estimators. This is the same notion of identification used in Douglas (2001). The following theorem is the main result of the paper.

**Theorem 3.2** $\bar{g}(x,t)$ and $h(x,t)$ are each large $J$ identified at the rate

$$r_J = \rho_{\alpha E} \frac{\alpha E}{\beta E} \rho_{\alpha N} \frac{\alpha N}{\beta N}$$

where $\rho = \left( \frac{\log(J)}{J} \right)^{1/2}$ and $g(x,t,u)$ is large $J$ identified at the rate $r_j^{\alpha_g}$.

**Remark 1** The rate is an upper bound on the distance between two observationally equivalent model. The rate derived varies depending on the degree of smoothness assumed in Assumptions 2.7–2.10 but is no faster than $\rho$, which is slightly slower than $J^{-1/2}$. The rate is independent of $K$ because each dimension of $\theta$ can be inferred directly using the restrictions in Assumption 2.5(iii). In Section 4 I show that alternative normalization strategies lead to a rate that is slower for higher dimensions of $\theta$.

**Remark 2** If some of the conditions described in Section 2 hold on $\mathcal{X} \times [t_l, t_u]^K$ but not on the entire support, $\mathcal{X} \times [0, 1]^K$ then it can also be shown that the structural functions are large $J$ identified on this restricted support in the sense that the difference between observationally equivalent versions of the structural function, restricted to this subspace of the support, converges to 0. This is worth noting because some models will be identified on the entire support only at a very slow rate because of behavior at 0 and 1 but will be identified on a restricted support at the rate $\rho$.

**Remark 3** If the regularity conditions are modified so that the functions, and their inverses, have a modulus of continuity that is exponential rather than polynomial then the convergence
4. ADDITIONAL RESULTS

The conditions imposed by Assumptions 2.2(iii), 2.5(ii), and 2.5(iii) are essentially normalizations so they can be replaced by alternative normalizations. Unlike with the standard notion of identification in a parametric model it is not apparent a priori that this is the case. These alternative normalizations will require new arguments and different regularity conditions, and, in some cases, will slow the rate of identification. I will explore this here with a few examples. In addition, condition 2.2(iii) is in fact stronger than necessary. I will show how identification can be obtained under a weaker condition, albeit possibly at a slower rate.

4.1. Avoiding the Exclusion Restrictions

The exclusion restriction in condition (iii) of Assumption 2.2 is in some cases difficult to justify. Consider Example 1 in Section 2. If the tests are administered before individuals complete their education then responses may be correlated with final education level conditional on innate ability (Hansen, Heckman, and Mullen, 2004). Thus education is endogenous. More educated individuals do better on the test both because of higher innate ability and because they are more educated. The test scores then reflect both ability and education. One strength of the approach described in this paper is that this problem can be solved if only $K$ questions on the test are not influenced by variation in education level. But this still may not be plausible in some contexts. An alternative approach is based on conditions suggested by Hansen, Heckman, and Mullen (2004). In particular they discuss identification of a linear factor model based on the assumptions that (i) performance on test questions is independent of final education level conditional on education level at the time of the test and ability and (ii) ability is independent of education level at the time of the test conditional on final education level. This motivates the following analysis.

Consider the model of Section 2 without condition 2.2(iii). Let $X = (X_1, X_2)$ and replace 2.2(iii) with the following two conditions.

**Assumption 4.1**

(i) $X_1 \perp \perp \theta \mid X_2$
(ii) \( M_E \perp \perp X_2 \mid X_1, \theta \)

These conditions are sufficient for identification. To see this, the first step is nearly identical to the proof of the Theorem 3.2. By condition (ii), the conditional expectation of \( M_E \) is

\[
p_E(x_1, t) := E(M_E \mid X = x, \theta = t) = E(M_E \mid X_1 = x_1, \theta = t).
\]

Conditioning on \( X \) and \( \tilde{M} \) provides the approximation

\[
E(M_E \mid \tilde{M} = m, X = x) \approx p_E(x_1, \tilde{p}^{-1}(m; x))
\]

If the right-hand side of this approximation is known, along with \( g^{-1}(y; x, \tilde{p}^{-1}(m; x)) \), then \( g(x, p_E^{-1}(m; x_1), u) \) is identified for each \( k \).

It also remains the case that

\[
Pr(T_k(\tilde{M}) \leq T_k(\tilde{p}_E^{-1}(x, m)) \mid X = x) \approx F_{\theta_k | X}(p_{E,1}^{-1}(m; x_1) \mid x)
\]

except that now \( p_{E,1}^{-1}(k) \) is a function of \( x_1 \) as well. By Assumption 4.1(i), the right-hand side object is equal to \( F_{\theta_k | X}(p_{E,1}^{-1}(m; x_1) \mid x_2) \). Integrating this over the distribution of \( X_2 \) produces \( p_{E,1}^{-1}(m; x_1) \) and hence the function \( g \) is fully identified. Thus the result of Theorem 3.2 will hold, with the same rates, for the model of Section 2 with 2.2(iii) replaced by Assumption 4.1. The proof is nearly identical to the proof of Theorem 3.2.

4.2. “Rotation” with Multivariate \( \theta \)

Since \( \theta \) is unobserved it can be “rescaled” by \( \bar{\theta} := G(\theta) \) for any invertible \( G : \mathbb{R}^K \to \mathbb{R}^K \). Then \( g(x, t, u) \) can be replaced by \( g(x, G^{-1}(t), u) \), and each \( p_j, p_{Nk}, \) and \( p_{Ek} \) can be similarly redefined. Assumption 2.5 is used to ensure that transforming the model this way does not produce an observationally equivalent model. Conditions (i) and (ii) of Assumption 2.5 set the scale of the marginal distribution of each \( \theta_k \). In the factor analysis literature this is known as a scale normalization. Assumption 2.5(iii) can be thought of as a rotation normalization. A rotation is a function \( G \) such that \( \bar{\theta} = G(\theta) \) is distributed differently from \( \theta \) but each \( \bar{\theta}_k \) has the same marginal distribution as \( \theta_k \).

The particular condition imposed by part (iii) of Assumption 2.5 to solve the rotation indeterminacy can be problematic if one wishes to give a clear interpretation to the latent variables \( \theta \). In some applications it is not plausible to assume that any of the binary proxies
only depends on $\theta_k$ (conditional on $X$). For example, if $\theta_1$ represents reading comprehension ability and $\theta_2$ represents quantitative ability then performance on every question on a math test may depend to some extent on $\theta_1$. In other words, imposing the normalization, or what is known in factor analysis as a rotation, 2.5(iii) disallows the interpretation of $\theta_2$ as pure quantitative ability. Consider instead the following assumption.

**Assumption 4.2**

(i) $\text{Supp}(X, \theta) = \mathcal{X} \times [0, 1]^K$

(ii) each component of $\theta$ satisfies $\theta_k \sim \text{Uniform}(0, 1)$

(iii) the components of $\theta$ are mutually independent

(iv) There are $K$ distinct values of $j - j_1, \ldots, j_K$ - such that for all $x, t, t'$

$$h_{jk}(x, t) = h_{jk}(x, t_1, \ldots, t_k, t'_{k+1}, \ldots, t'_K)$$

and $\varepsilon_{jk} \perp \perp \theta_{k+1}, \ldots, \theta_K \mid X, \theta_1, \ldots, \theta_k.$

This allows every question to depend on $\theta_1$ while requiring one question that does not depend on any other dimension of $\theta$, another that does not depend on $\theta_k$ for $k \geq 3$, etc. The tradeoff is that now I will require the different dimensions of $\theta$ to be independent. This upper triangular restriction\(^{15}\) is a common normalization in linear factor models (Anderson and Rubin, 1956). Also replace Assumptions 2.6 and 2.7 with the following two sets of conditions.

**Assumption 4.3**

(i) For each $x \in \mathcal{X}$, $\bar{p}(x, \cdot)$ is injective.

(ii) $p_E(x, \cdot)$ is injective.

(iii) For each $k$, each $x \in \mathcal{X}$ and each $t_1, \ldots, t_{k-1}$, $p_{Nk}(x, t_1, \ldots, t_{k-1}, \cdot)$ is a strictly increasing function.

Assumptions 4.2, 4.3, along with the modified regularity conditions stated below as Assumption 4.4 can be used in place of Assumptions 2.5, 2.6, and 2.7 to identify the model of Section 2. I will show this for $K = 2$.

\(^{15}\)Suppose $p_{Nk}(x, t) := E(M_{Nk} \mid X = x, \theta = t)$ is equal to $\mu_k(x) + \alpha_k't$. Let $\alpha$ denote the square matrix with $k^{th}$ row given by $\alpha_k'$. Then condition (iv) of Assumption 4.2 is satisfied if $\alpha$ is upper triangular.
First, note that the first half of the argument in Section 3.2.1 is still valid. Thus \( g(x, p_E^{-1}(m), u) \) is identified. Next, because \( M_{N_1} \) is independent of \( \theta_2, \ldots, \theta_K \) conditional on \( \theta_1 \) and \( X \), identification of \( p_{E,(1)}^{-1}(\tau_1) \) follows from the same argument used in Section 3.2.1. However because \( \theta_1 \) is not excluded from any equation the same argument can not be used to provide identification of \( p_{E,(2)}^{-1}(\tau_2) \).

Instead define the function \( T_2(m) := p_{N_2}(x, p^{-1}(m; x)) \) where \( p_{N_2}(x, t_1, t_2) \) is monotonic in \( t_2 \) for any value of \( (x, t_1) \) by Assumption 4.3(iii). This function is identified in the limit by \( E(M_{N_2} | X = x, M = m) \) so now consider the distribution of \( T_2(M) \).

\[
Pr(T_2(M) \leq \tau_2 | T_1(M) = \tau_1, X = x) \approx Pr(p_{N_2}(x, \theta) \leq \tau_2 | p_{N_1}(x, \theta_1) = \tau_1, X = x)
= F_{\theta_2|\theta_1, X}(p_{N_2}^{-1}(\tau_2; x, p_{N_1}^{-1}(\tau_1; x)) | x, p_{N_1}^{-1}(\tau_1; x))
\]

Again, combining this with previous results it can be shown that \( F_{\theta_2|\theta_1, X}(p_{E,2}^{-1}(\tau_2; p_{E,1}^{-1}(\tau_1)) | x, p_{E,1}^{-1}(\tau_1)) \) is identified. If \( \theta_2 | \theta_1 \sim Uniform(0, 1) \) then this implies, by integrating over the distribution of \( X \), that \( p_{E,2}^{-1}(\tau_2; p_{E,1}^{-1}(\tau_1)) \) is identified. And therefore \( p_E \) is identified and hence so is \( g(x, t, u) \).

The following regularity conditions are used to ensure that these approximations are valid.

**Assumption 4.4**

(i) The conditional distribution function of \( \theta | X \), denoted \( F_{\theta|X}(t | x) \), satisfies

\[
(a) \quad \min_k \inf_{x \in X, t \in [0,1]^K, t_k \in [0,1], t_k \neq t_k} \frac{|F_{\theta|X}(t_k, t, x) - F_{\theta|X}(t_k, t-k, x)|}{|t_k - t_k|^\beta_F} > \epsilon_F
\]

(b) the distribution of \( \theta_1 | X \) has a density function, \( f_{\theta_1|X} \), satisfying

\[
(b) \quad \sup_{x \in X, t \in [0,1]} \frac{|f_{\theta_1|X}(t_1, x) - f_{\theta_1|X}(t_1, x)|}{|t_1 - t_1|^\alpha_F} < M_F^{(1)}
\]

(c) \( \sup_{x \in X, t \in [0,1]^2, t_2 \in [0,1], t_2 \neq t_2} \frac{|F_{\theta_2|\theta_1, X}(t_2, t_1, x) - F_{\theta_2|\theta_1, X}(t_2, t_1, x)|}{|t_2 - t_2|^\alpha_F} < M_F^{(2)}
\]

(d) \( \sup_{x \in X, t \in [0,1]^2, t' \in [0,1]^2, t' \neq t} \frac{|F_{\theta_2|\theta_1, X}(t_2, t_1, x) f_{\theta_1|X}(t_1, x) - F_{\theta_2|\theta_1, X}(t_2, t_1, x) f_{\theta_1|X}(t_1, x)|}{|t' - t|^\alpha_F} < M_F^{(3)}
\]
(ii) \( \bar{p} \) satisfies

\[
\sup_{x \in X, t \in [0,1], t' \in [0,1], t' \neq t} \frac{||\bar{p}(x, t') - \bar{p}(x, t)||}{||t' - t||^{\alpha_{\bar{p}}}} < M_{\bar{p}}
\]

and

\[
\inf_{x \in X, t \in [0,1], t' \in [0,1], t' \neq t} \frac{||\bar{p}(x, t') - \bar{p}(x, t)||}{||t' - t||^{\beta_{\bar{p}}}} > \epsilon_{\bar{p}}
\]

(iii) \( p_E \) satisfies

\[
\sup_{t \in [0,1], t' \in [0,1], t' \neq t} \frac{||p_E(t') - p_E(t)||}{||t' - t||^{\alpha_E}} < M_E
\]

and

\[
\inf_{t \in [0,1], t' \in [0,1], t' \neq t} \frac{||p_E(t') - p_E(t)||}{||t' - t||^{\beta_E}} > \epsilon_E
\]

(iv) \( p_N \) satisfies

\[
\sup_{x,t,t'} \frac{||p_N(x, t') - p_N(x, t)||}{||t' - t||^{\alpha_N}} < M_N
\]

and

\[
\min_{k=1,\ldots,K} \inf_{x,t,t',t_k,t'_k} \frac{|p_{N_k}(x,t_{k-1},t'_k,t_{k+1};x,t_{k-1},t_k,t_{k+1};t)|}{|t'_k - t_k|^{\beta_N}} > \epsilon_N
\]

The proof of the following result is provided in the appendix.

**Theorem 4.1** Suppose \( K = 2 \). Under Assumptions 2.1, 2.2, 4.2, 4.3, and 4.4, \( \bar{g}(x,t) \) is large \( J \) identified at the rate

\[
r_J = \rho^{\frac{\alpha_E}{\beta_E}} + \rho^{\frac{\alpha_N}{\beta_N}}
\]

where \( \rho = \left( \frac{\log(J)}{J} \right)^{1/2} \) and

\[
\delta = \frac{(\alpha_N/\beta_N) \min\{\alpha_N/\beta_N, \alpha_E/\beta_E\} \min\{\alpha_F(1), \alpha_F(2)\}}{(\alpha_N/\beta_N) \min\{\alpha_F(1), \alpha_F(2)\} + 1}
\]

If, in addition, Assumptions 2.3 and 2.8 hold then \( g(x,t,u) \) is large \( J \) identified at the rate \( r_J^{\alpha_g} \). If instead Assumptions 2.4 and 2.9 hold then \( h(x,t) \) is large \( J \) identified at the rate \( r_J \).
4.3. Separate Exclusion Restrictions

When \( X \) consists of \( L > 1 \) covariates condition (iii) of Assumption 2.2 can be relaxed. Each component \( X_l \) of \( X \) must be excluded from \( K \) of the binary proxy variables. However the exclusions may be different for different \( l \). For example, ... The following is sufficient for identification.

**Assumption 4.5**

For each \( l = 1, \ldots, L \) there are \( K \) distinct values of \( j \) such that (a) \( X_l \) is excluded from \( h_j \), that is, \( h_j(x, t) \equiv h_j(x'_l, x - l, t) \) for all \( x, x'_l, t \) and (b) \( \varepsilon_j \perp X_1 | \theta \).

Under this assumption I can use \( M_{EI} = (M_{EI1}, \ldots, M_{EIK}) \) to denote the proxies from which \( X_l \) is excluded and define the associated conditional expectation functions, \( p_{EI}(x - l, t) \). Assumption 2.2(iii) can be used to identify the model in an iterative process.

First, for each \( k \), \( F_{\theta_k | X}(p^{-1}_{EI1,k}(\tau_k; x - 1) | x) \) is identified according to the usual argument. Integrating this with respect to the distribution of \( X_1 | X - 1 \) produces \( F_{\theta_k | X - 1}(p^{-1}_{EI1,k}(\tau_k; x - 1) | x - 1) \). The vector-valued function \( p_{EI1} \) can be linked to \( p_{E2} \) in a manner which should now be familiar to identify \( F_{\theta_k | X - 1}(p^{-1}_{E21,k}(\tau_k; x - 2) | x - 1) \) for each \( k \). Integrating this with respect to the distribution of \( X_2 | X - {1, 2} \) produces \( F_{\theta_k | X - {1, 2}}(p^{-1}_{E22,k}(\tau_k; x - 2) | x - {1, 2}) \). This can be continued until the \( L^{th} \) step produces \( F_{\theta_k}(p^{-1}_{EL,k}(\tau_k; x - L)) = p^{-1}_{EL,k}(\tau_k; x - L) \).

5. CONCLUSION

This paper introduces new results that demonstrate how a many binary indicators can be used obtain identification in a nonseparable model with endogeneity. This provides an approach that neither assumes conditional independence (exogeneity conditional on a vector of observed covariates) nor requires an instrument that is excluded from the outcome equation.

This paper is also a contribution to the increasingly important literature on high-dimensional data in econometrics. The results provide an alternative use of high-dimensional data in the context of an economic model with heterogeneity to current work, such as Belloni, Chernozhukov, Fernández-Val, and Hansen (2013). 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


Supplementary Material

The supplementary material includes additional examples, extensions, and theoretical applications of the results in the paper, as well as proofs of the main results. Proofs of any other results are available from the author upon request.

APPENDIX A: IDENTIFICATION PROOFS AND RELATED RESULTS

Proof of Theorem 3.1

The following lemma is used to prove Theorem 3.1. In order to state the lemma consider any two functions $h_Y, h'_Y$ in the identified set. There must be auxiliary functions $p_1, \ldots, p_J$ and $p'_1, \ldots, p'_J$ that complete the two models. Define $\bar{p} = J^{-1}\sum_j p_j$ and $\bar{p}' = J^{-1}\sum_j p'_j$. The lemma provides a bound on $\sup_{x\in[0,1]}|\bar{p}(x) - \bar{p}'(x)|$.

Lemma A.1 Under Assumption 3.1 and the conditions in Theorem 3.1, there exists $C > 0$ and $J_0$ such that for all $J \geq J_0$, $\sup_{x\in[0,1]}|\bar{p}(x) - \bar{p}'(x)| \leq C\rho$ where $\rho = (\log(J-1)/(J-1))^{1/2}$.

Proof of Lemma A.1: Suppose the conclusion of the lemma is not true. Then for any $C > 0$ there are infinitely many positive integers $J$ such that I can find functions $\bar{p}$ and $\bar{p}'$ in the identified set and a point $x \in [0,1]$ such that $\bar{p}(x) < kJ^{-1} - C\rho$ and $\bar{p}'(x) > kJ^{-1} + C\rho$ for some $k$.

Then the first inequality implies that

$$
P_r(\bar{M} \leq kJ^{-1}) \geq P_r(\bar{M} \leq kJ^{-1}, \bar{p}(\theta) \leq \bar{p}(x) + C\rho/2)$$

$$\geq \int_{x:\bar{p}(x^*) \leq \bar{p}(x) + C\rho/2} (1 - P_r(\bar{M} \geq kJ^{-1} | \theta = x^*)) dx^*$$

$$\geq \int_{x:\bar{p}(x^*) \leq \bar{p}(x) + C\rho/2} (1 - P_r(|\bar{M} - \bar{p}(x^*)| \geq C\rho/2 | \theta = x^*)) dx^*$$

By Hoeffding’s inequality $P_r(|\bar{M} - \bar{p}(x^*)| \geq C\rho/2 | \theta = x^*) \leq 2J^{-C^2/2}$ so

$$P_r(\bar{M} \leq kJ^{-1}) \geq (1 - 2J^{-C^2/2})P_r(\bar{p}(\theta) \leq \bar{p}(x) + C\rho/2)$$
On the other hand, if \( \bar{p}'(x) > kJ^{-1} + C \rho \) then

\[
Pr(\bar{M} \leq k/J^{-1}) \leq Pr(\bar{M} \leq \bar{p}'(x) - C \rho)
\leq \int_{\bar{p}'(x) < \rho(x) + C \rho/2} Pr(\bar{M} - \bar{p}'(x^*) \geq C \rho/2 | \theta = x^*) dx^*
+ \int_{\bar{p}'(x) - C \rho/2 > \rho(x^*)} dx^*
\leq 2J^{-C^2/2} + Pr(\bar{p}'(x) - C \rho/2 > \bar{p}'(\theta))
\]

By the conditions in the statement of Theorem 3.1, \( Pr(\bar{p}'(x) - C \rho/2 > \bar{p}'(\theta)) \leq Pr(x - \rho^* > \theta) = x - \rho^* \) and \( Pr(\bar{p}'(\theta) \leq \bar{p}(x) + C \rho/2) \geq Pr(\theta \leq x + \rho^*) = x + \rho^* \) where \( \rho^* = \left\{ \frac{c_3}{2M_3} \right\}^{1/\alpha_3} \).
Combining these results implies that \( \rho^* \leq 2J^{-C^2/2} \). This produces a contradiction because if \( C^2 > 1/\alpha_3 \) then this cannot hold for infinitely many positive integers \( J \).

\[Q.E.D.\]

I now prove Theorem 3.1.

**Proof of Theorem 3.1:** Consider any \( h_Y, h_Y' \) in the identified set. By definition there are functions \( p_1, \ldots, p_J \) and \( p_1', \ldots, p_J' \) that complete the two models, respectively. Define \( \bar{p} = J^{-1} \sum_{j=1}^J p_j \) and \( \bar{p}' = J^{-1} \sum_{j=1}^J p_j' \).

Take any \( x \in [0, 1] \). Define \( k = \lceil J\bar{p}(x) \rceil \) so that \( |\bar{p}(x) - kJ^{-1}| < J^{-1} \). By Lemma A.1 there exists \( C > 0 \) such that \( |\bar{p}(x) - \bar{p}'(x)| \leq C \rho_J \) where \( \rho_J = (\log(J)/J)^{1/2} \), and therefore by the triangle inequality \( |\bar{p}'(x) - kJ^{-1}| < 2C \rho_J \), at least for \( J \) sufficiently large. Thus both \( \bar{p}(x) \) and \( \bar{p}'(x) \) are within \( 2C \rho_J \) of \( kJ^{-1} \).

Now consider \( E(Y \mid |\bar{M} - kJ^{-1}| < \kappa \rho_J) \) for some \( \kappa > 0 \). I will show that this converges to both \( h_Y(x) \) and \( h_Y'(x) \) as \( J \to \infty \). First, since \( E(Y \mid \theta) = h_Y(\theta) \),

\[
E(Y \mid |\bar{M} - kJ^{-1}| < \kappa \rho_J) = E(h_Y(\theta) \mid |\bar{M} - kJ^{-1}| < \kappa \rho_J)
\]
Therefore,
\[
|h_Y(x) - E(Y \mid |\bar{M} - kJ^{-1}| < \kappa \rho J)| \leq \int |h_Y(x) - h_Y(x^*)| dF_{\bar{M} - kJ^{-1} < \kappa \rho J}(x^*)
\]
\[
= \int_{|x-x^*| \leq \eta} |h_Y(x) - h_Y(x^*)| dF_{\bar{M} - kJ^{-1} < \kappa \rho J}(x^*)
\]
\[
+ \int_{|x-x^*| > \eta} |h_Y(x) - h_Y(x^*)| dF_{\bar{M} - kJ^{-1} < \kappa \rho J}(x^*)
\]

Take \( \eta = \gamma M_2 \rho^{\alpha_2} \). The first term is bounded by \( M_1 \eta^{\alpha_1} = O(\rho^{\alpha_1 \alpha_2}) \) since \( h_Y \) is Holder continuous as described in the statement of the theorem. And because \( h_Y \) is bounded between 0 and 1 the second term is no larger than \( Pr(|\theta - x| > \eta \mid |\bar{M} - kJ^{-1}| < \kappa \rho J) \).

To further bound this probability note that by the conditions of the theorem, \(|\theta - x| > \eta \) implies that \(|\bar{p}(\theta) - \bar{p}(x)| > (\eta/M_2)^{1/\alpha_2} \). This in turn implies that \(|\bar{p}(\theta) - kJ^{-1}| > \frac{2}{3}(\eta/M_2)^{1/\alpha_2} \) since \( \gamma \) can be chosen large enough in defining \( \eta \) that \( 2C \rho < \frac{1}{3}(\eta/M_2)^{1/\alpha_2} \). Therefore,

\[
Pr(|\theta - x| > \eta \mid |\bar{M} - kJ^{-1}| < \kappa \rho J) \leq Pr(|\bar{p}(\theta) - kJ^{-1}| > \frac{2}{3}(\eta/M_2)^{1/\alpha_2} \mid |\bar{M} - kJ^{-1}| < \kappa \rho J)
\]
\[
\leq Pr(|\bar{M} - \bar{p}(\theta)| > \frac{1}{3}(\eta/M_2)^{1/\alpha_2} \mid |\bar{M} - kJ^{-1}| < \kappa \rho J)
\]

where the second inequality follows because \(|\bar{p}(\theta) - kJ^{-1}| \leq |\bar{M} - \bar{p}(\theta)| + |\bar{M} - kJ^{-1}| \) and \( \gamma \) can be chosen large enough in the definition of \( \eta \) so that \( \kappa \rho < \frac{1}{3}(\eta/M_2)^{1/\alpha_2} \).

Then

\[
Pr(|\bar{M} - \bar{p}(\theta)| > \frac{1}{3}(\eta/M_2)^{1/\alpha_2} \mid |\bar{M} - kJ^{-1}| < \kappa \rho J) \leq \frac{Pr(|\bar{M} - \bar{p}(\theta)| > \frac{1}{3}(\eta/M_2)^{1/\alpha_2})}{Pr(|\bar{M} - kJ^{-1}| < \kappa \rho J)}
\]

By Hoeffding’s inequality, and since \( \frac{1}{3}(\eta/M_2)^{1/\alpha_2} = \frac{1}{3} \gamma^{1/\alpha_2} \rho \),

\[
Pr(|\bar{M} - \bar{p}(\theta)| > \frac{1}{3}(\eta/M_2)^{1/\alpha_2}) = \int Pr(|\bar{M} - \bar{p}(\theta)| > \frac{1}{3}(\eta/M_2)^{1/\alpha_2} \mid \theta = x) dx \leq 2 \exp(-2J(\frac{1}{3} \gamma^{1/\alpha_2} \rho)^2) = 2J^{-\frac{2}{3} \gamma^{1/\alpha_2}}
\]
Also,

\[
Pr(|\bar{M} - kJ^{-1}| < \kappa \rho J) \geq Pr(|\bar{M} - kJ^{-1}| < \kappa \rho J, |\bar{p}(\theta) - kJ^{-1}| < \frac{1}{2} \kappa \rho J)
\]

\[
= \int_{x^* : |\bar{p}(x^*) - kJ^{-1}| < \frac{1}{2} \kappa \rho J} (1 - Pr(|\bar{M} - kJ^{-1}| \geq \kappa \rho J | \theta = x^*)) dx^*
\]

\[
\geq \int_{x^* : |\bar{p}(x^*) - kJ^{-1}| < \frac{1}{2} \kappa \rho J} (1 - Pr(|\bar{M} - \bar{p}(x)| \geq \frac{1}{2} \kappa \rho J | \theta = x^*)) dx^*
\]

\[
\geq (1 - 2J^{-\frac{1}{2}\kappa^2}) Pr(|\bar{p}(\theta) - kJ^{-1}| < \frac{1}{2} \kappa \rho J)
\]

If \( \kappa \) is chosen large enough that \( \kappa > 8C \) then I can conclude that

\[
Pr(|\bar{M} - kJ^{-1}| < \kappa \rho J) \geq (1 - 2J^{-\frac{1}{2}\kappa^2}) Pr(|\bar{p}(\theta) - \bar{p}(x)| < \frac{1}{4} \kappa \rho J)
\]

\[
\geq (1 - 2J^{-\frac{1}{2}\kappa^2}) Pr(|\theta - x| < \left(\frac{\kappa \rho}{4M_3}\right)^{1/\alpha^3})
\]

\[
\geq (1 - 2J^{-\frac{1}{2}\kappa^2}) \left(\frac{\kappa \rho}{4M_3}\right)^{1/\alpha^3}
\]

Combining these two results, it is clear that if \( \gamma \) is chosen large enough then for sufficiently large \( J \),

\[
Pr(|\bar{M} - \bar{p}(\theta)| > \frac{1}{3} (\eta/M_2)^{1/\alpha^2} | |\bar{M} - kJ^{-1}| < \kappa \rho J)
\]

\[
\leq \frac{2J^{-\frac{2}{9}1/\alpha^2}}{(1 - 2J^{-\frac{1}{2}\kappa^2}) \left(\frac{\kappa \rho}{4M_3}\right)^{1/\alpha^3}} = o(\rho^{\alpha_1\alpha_2})
\]

and therefore, \( |h_Y(x) - E(Y | |\bar{M} - kJ^{-1}| < \kappa \rho J)| = O(\rho^{\alpha_1\alpha_2}) \).

By the same argument, \( |h'_Y(x) - E(Y | |\bar{M} - kJ^{-1}| < \kappa \rho J)| = O(\rho^{\alpha_1\alpha_2}) \) and hence \( |h'_Y(x) - h'_Y(x)| = O(\rho^{\alpha_1\alpha_2}) \).

\( Q.E.D. \)
Proof of Theorem 3.2

First I state and prove three lemma that will be used.

**Lemma A.2** For any \( \delta > 0 \), \( \sup_{t \in [0,1]^K, x \in \mathcal{X}} Pr(||\bar{M} - \bar{p}(x, \theta)|| > \delta \mid X = x, \theta = t) \leq 2K \exp(-2\delta^2 J) \).

**Proof:** Since

\[
Pr(||\bar{M} - \bar{p}(x, \theta)|| > \delta \mid X = x, \theta = t) \leq \sum_{k=1}^{K} Pr(||\bar{M}_k - \bar{p}_k(x, \theta)|| > \delta \mid X = x, \theta = t)
\]

the result follows by Hoeffding’s inequality. \( Q.E.D. \)

**Lemma A.3** There exists a constant \( \kappa \) such that for any \( (x, t) \in \mathcal{X} \times [0,1]^K \), \( Pr(||\bar{M} - \bar{p}(x, t)|| < \rho \mid X = x) \geq \kappa \rho \beta \alpha^{-1} (1 - 2 \exp(-\frac{1}{2} J \rho^2)) \).

**Proof:** Noting that

\[
Pr(||\bar{M} - \bar{p}(x, t)|| < \rho \mid X = x) \geq \int_{||\bar{p}(x, t^*) - \bar{p}(x, t)|| < \rho / 2} Pr(||\bar{M} - \bar{p}(x, t)|| < \rho \mid X = x, \theta = t^*) dF_{\theta|X=x}(t^*)
\]

the result follows from Assumptions ... and Hoeffding’s inequality. \( Q.E.D. \)

Next is an additional result that follows directly from Lemmas 1 and 2.

**Lemma A.4** There exists \( \kappa > 0 \) such that for any \( x, t \) and there is a \( t' \in [0,1]^K \) such that

\[ ||\bar{p}(x, t) - \bar{p}'(x, t')|| \leq \kappa (\log(J)/J)^{1/2} \]

Next, the proof of Theorem 3.2.

**Proof:** If \( g(x, t, u) \) and \( g'(x, t, u) \) are \( J \)-observationally equivalent then I can define \( \bar{p}(x, t) \) and \( \bar{p}'(x, t) \) associated with the two respective probability distributions.
Fix any $x \in X, t \in [0,1]^K$ and define $m = \mathbf{p}(x,t)$. According to Lemma 3 there must be $t' \in [0,1]^K$ such that $||\mathbf{p}'(x,t') - m|| < \rho_0$ where for any $\kappa > 0$ I define $\rho(\kappa) := \kappa(\log(J)/J)^{1/2}$ and $\rho_0 := \rho(\kappa_0)$ for some $\kappa_0 > 0$.

By Assumptions 2.1, 2.2, and 2.3, for any $y \in \mathcal{Y}$

$$Pr(Y \leq y \mid ||\mathbf{M} - m|| < \rho(\kappa), X = x) = E(g^{-1}(y;x,\theta) \mid ||\mathbf{M} - m|| < \rho(\kappa), X = x)$$

So, for any $\kappa > 0$, defining $\eta > 0$ such that $\rho(\kappa) \leq \frac{1}{3} \epsilon_\beta \eta^3$,

$$||Pr(Y \leq y \mid ||\mathbf{M} - m|| < \rho(\kappa), X = x) - g^{-1}(y;x,t)||$$

$$\leq M_\kappa \eta^3 + Pr(||\theta - t|| > \eta \mid ||\mathbf{M} - m|| < \rho(\kappa), X = x)$$

$$\leq M_\kappa \eta^3 + Pr(||\mathbf{p}(x,\theta) - \mathbf{p}(x,t)|| > \epsilon_\beta \eta^3 \mid ||\mathbf{M} - m|| < \rho(\kappa), X = x)$$

$$\leq M_\kappa \eta^3 + Pr(||\mathbf{M} - \mathbf{p}(x,\theta)|| > \frac{1}{3} \epsilon_\beta \eta^3 \mid ||\mathbf{M} - m|| < \rho(\kappa), X = x)$$

where the first inequality uses Assumption 2.8, the second uses Assumption 2.7(ii), and the third holds because $\rho(\kappa) \leq \frac{1}{3} \epsilon_\beta \eta^3$. The second term in the last expression can be bounded as follows.

$$Pr(||\mathbf{M} - \mathbf{p}(x,\theta)|| > \frac{1}{3} \epsilon_\beta \eta^3 \mid ||\mathbf{M} - m|| < \rho, X = x)$$

$$\leq \frac{Pr(||\mathbf{M} - \mathbf{p}(x,\theta)|| > \frac{1}{3} \epsilon_\beta \eta^3 \mid X = x)}{Pr(||\mathbf{M} - m|| < \rho \mid X = x)}$$

$$\leq b_0 \exp(-b_1 \eta^2 \rho^2) \rho^{-\alpha_3^-1}$$

where the last inequality holds for sufficiently large $J$ for some $b_0, b_1 > 0$ according to Lemmas 1 and 2 provided that $J \rho^2 \to \infty$. Taking $\eta$ such that $\frac{1}{3} \epsilon_\beta \eta^3 = \rho$ and $\kappa$ sufficiently large, for some $\kappa_0 > 0$ and all sufficiently large $J$

$$||Pr(Y \leq y \mid ||\mathbf{M} - m|| < \rho, X = x) - g^{-1}(y;x,t)|| \leq \eta_1 := \kappa_0(\log(J)/J)^{\alpha_3 \beta_3^{-1/2}}$$

Similarly, since $||\mathbf{p}'(x,t') - m|| \leq \rho_0$, the same decomposition can be applied for $\kappa > \kappa_0$ to show that

$$||Pr(Y \leq y \mid ||\mathbf{M} - m|| < \rho(\kappa), X = x) - g^{-1}(y;x,t')|| \leq \eta_1^2$$
Therefore, if \( ||t - t'|| < \psi \) then

\[
|g^{-1}(y; x, t) - g^{-1}(y; x, t')| \leq |g^{-1}(y; x, t) - g^{-1}(y; x, t')| + |g^{-1}(y; x, t') - g^{-1}(y; x, t)| \\
\leq 2r_1^* + M\psi^\alpha_h
\]

Since this holds for any \( y \in \mathcal{Y} \) it implies that for any \( u \in [0, 1] \),

\[
|g(x, t, u) - g'(x, t, u)| \leq (r_1^*)^\alpha_g + \psi^\alpha_h
\]

Repeating the same argument shows that if \( ||t - t'|| < \psi \) then, because of Assumption 2.4 and 2.9,

\[
|h(x, t) - h'(x, t)| \leq r_1^* + \psi^\alpha_h
\]

and, using Assumption 2.10,

\[
|\bar{g}(x, t) - \bar{g}'(x, t)| \leq r_1^* + \psi^\alpha_h
\]

So it remains to find \( \psi \) such that \( ||t - t'|| < \psi \).

- Let \( x \) vary through \( \mathcal{X} \) but keep \( t \) fixed; by Assumption 2.5(i) \( t \) remains in the support of \( \theta \mid X = x \) as \( x \) varies so I can define \( m(x) := \bar{p}(x, t) \) and define \( t'(x) \) to be a vector in \( [0, 1]^K \) provided by Lemma 3 such that \( ||\bar{p}'(x, t'(x)) - m(x)|| \leq \rho_0 := \kappa_0(\log(J)/J)^{1/2} \) for some \( \kappa_0 > 0 \). I will now show that for any \( x_1 \in \mathcal{X} \), \( \sup_{x_2 \in \mathcal{X}} |t'(x_2) - t'(x_1)| \leq (r_2^*)^{1/2} \).

- By Assumption 2.2, \( E(M_E \mid \bar{M} = m(x), X = x) = E(p_E(\theta) \mid \bar{M} = m(x), X = x) \) and therefore, using Assumptions 2.7(ii) and 2.7(iii) and the same argument employed above,

\[
|E(M_E \mid \bar{M} = m(x), X = x) - p_E(t)| \leq M_E\eta^\alpha_E + Pr( ||\theta - t|| > \eta \mid \bar{M} = m(x), X = x) \\
\leq r_2^* := \kappa_2^*(\log(J)/J)^{\alpha_E\beta_E^{1/2}}
\]

for some \( \kappa_2^* > 0 \), and likewise \( |E(M_E \mid \bar{M} = m(x), X = x) - p'_E(t'(x))| \leq r_2^* \), and

\[\text{If } \bar{p} \text{ and } \bar{p}' \text{ are bijective, as functions of } t \text{ then I am essentially arguing here that } \bar{p}'^{-1}(\bar{p}(x, t); x) \text{ does not vary with } x \text{ in the limit. This is the first step in showing that } \bar{p} \approx \bar{p}'.\]
Therefore,

\[ |p'_E(t'(x_2)) - p'_E(t'(x_1))| \leq |p'_E(t'(x_2)) - p_E(t)| + |p'_E(t'(x_1)) - p_E(t)| \]
\[ \leq |E(M_E | M = m(x_2), X = x_2) - p'_E(t'(x_2))| + |E(M_E | M = m(x_2), X = x_2) - p_E(t)| \]
\[ + |E(M_E | M = m(x_1), X = x_1) - p'_E(t'(x_1))| + |E(M_E | M = m(x_1), X = x_1) - p_E(t)| \]
\[ \leq 4r_2^* \]

By Assumption 2.7(iii), \( |p'_E(t'(x_2)) - p'_E(t'(x_1))| < 4r_2^* \) implies \( |t'(x_2) - t'(x_1)| < (4r_2^*/\epsilon_E)\beta_E^{-1} \).

- By assumption 2.5(ii), for each \( k, \theta_k \sim \text{Uniform}(0, 1) \) and, as a result, for any \( s \in \mathbb{R}^K \),
\[ s_k = F_{\theta_k}(s_k) = \int_X F_{\theta_k|X}(s_k | x^*)dF_X(x^*) \]. The same thing is of course true if \( F_{\theta_k|X} \) is replaced by the observationally equivalent \( F'_{\theta_k|X} \). Therefore,

\[ |t_k - t'_k(x)| = \left| \int_X F_{\theta_k|X}(t_k | x^*)dF_X(x^*) - \int_X F'_{\theta_k|X}(t'_k(x) | x^*)dF_X(x^*) \right| \]
\[ \leq \left| \int_X F_{\theta_k|X}(t_k | x^*)dF_X(x^*) - \int_X F'_{\theta_k|X}(t'_k(x^*) | x^*)dF_X(x^*) \right| \]
\[ + \left| \int_X F'_{\theta_k|X}(t'_k(x^*) | x^*)dF_X(x^*) - \int_X F'_{\theta_k|X}(t'_k(x) | x^*)dF_X(x^*) \right| \]

Therefore, using Assumption 2.7(i) and applying the previous result for \( x_2 = x^* \) and \( x_1 = x \),

\[ |t_k - t'_k(x)| \leq \sup_{x \in \mathcal{X}} |F_{\theta_k|X}(t_k | x^*) - F'_{\theta_k|X}(t'_k(x^*) | x^*)| + M_F(4r_2^*/\epsilon_E)\alpha_E\beta_E^{-1} \]

So the last step is to derive an upper bound on \( \sup_{x^* \in \mathcal{X}} |F_{\theta_k|X}(t_k | x^*) - F'_{\theta_k|X}(t'_k(x^*) | x^*)| \).

- For each \( k \) and any \( x^* \in \mathcal{X} \), define \( \tau_k = p_{N_k}(x^*, t) \). Recall from the text that

\[ T_k(m) := E(M_{N_k} | ||\tilde{M} - m|| < \rho, X = x^*) \]

The dependence on \( \rho \) is implicit in the notation; take \( \rho = \kappa(\log(J)/J)^{1/2} \) where \( \kappa > 0 \).
is chosen to be sufficiently large.

First consider

\begin{equation}
Pr(T_k(\bar{M}) \leq \tau_k \mid X = x^*) \geq \int_{p_{NK}(x^*,t^*) \leq \tau_k - r^*} (1 - Pr(|T_k(\bar{M}) - p_{NK}(x^*,t^*)| > r \mid X = x^*, \theta = t^*))dF_{\theta \mid X}(t^* \mid x^*)
\end{equation}

Note that by the argument used twice above, and Assumption 2.7(ii) and 2.7(iv), if $||\bar{p}(\bar{x},\bar{t}) - \bar{m}|| < \kappa(\log(J)/J)^{1/2}$ then there is $\kappa_3 > 0$ such that $|T_k(\bar{m}) - p_{NK}(\bar{x},\bar{t})| < r^* := \kappa_3^*(\log(J)/J)^{\alpha N \beta^{-1}}$. Therefore,

\[Pr(|T_k(\bar{M}) - p_{NK}(x^*,t^*)| > r_3^* \mid X = x^*, \theta = t^*) \leq Pr(||\bar{M} - \bar{p}(x^*,\theta)|| > \kappa(\log(J)/J)^{1/2} \mid X = x^*, \theta = t^*) \leq 2K \exp(-2\kappa^2 \log(J))\]

and therefore, using $r = r_3^*$ in equation (A.1) and applying Assumption 2.5(iii) and 2.7(iv),

\[Pr(T_k(\bar{M}) \leq \tau_k \mid X = x^*) \geq (1 - 2K \exp(-2\kappa^2 \log(J))) \int_{p_{NK}(x^*,t^*) \leq \tau_k - r_3^*} dF_{\theta \mid X}(t^* \mid x^*) \geq (1 - 2K \exp(-2\kappa^2 \log(J))) F_{\theta \mid X}(t_k - (r_3^*/\epsilon_N)\beta^{-1} \mid x^*)\]

On the other hand,

\[Pr(T_k(\bar{M}) \leq \tau_k \mid X = x^*) \leq Pr(p_{NK}(x^*,\theta) \leq \tau_k + r_3^* \mid X = x^*) + \int_{p_{NK}(x^*,t^*) > \tau_k + r_3^*} Pr(|T_k(\bar{M}) - p_{NK}(x^*,t^*)| > r_3^* \mid X = x^*, \theta = t^*)dF_{\theta \mid X}(t^* \mid x^*) \leq F_{\theta \mid X}(t_k + (r_3^*/\epsilon_N)\beta^{-1} \mid x^*) + 2K \exp(-2\kappa^2 \log(J))\]

Using Assumption 2.5(iii) and 2.7(iv) I can conclude that

\[|F_{\theta \mid X}(t_k \mid x^*) - Pr(T_k(\bar{M}) \leq \tau_k \mid X = x^*)| \leq 2K \exp(-2\kappa^2 \log(J)) + (r_3^*/\epsilon_N)^{\beta^{-1}}\]
Next,

\[ |p_{N_k}(x^*, t) - p'_{N_k}(x^*, t'(x^*))| \leq 2r_3^* \]

since \( p_{N_k}(x^*, t) \) and \( p'_{N_k}(x^*, t'(x^*)) \) are each within \( r_3^* \) of \( T_k(m) \).

Therefore,

\[
\begin{align*}
(A.2) \quad &Pr(T_k(\bar{M}) \leq \tau_k | X = x^*) \\
&\geq \int_{p_{N_k}(x^*, t^*) \leq p'_{N_k}(x^*, t^*) - 3r_3^*} (1 - Pr(|T_k(\bar{M}) - p'_{N_k}(x^*, t^*)| > r_3^* | X = x^*, \theta = t^*)) dF'_{\theta|X}(\theta)
\end{align*}
\]

\[
(A.3) \quad \geq (1 - 2K \exp(-2\kappa^2 \log(J))) F'_{\theta|X}(t_k - (3r_3^*/\epsilon N)^{\beta_N^{-1}} | x^*)
\]

and

\[
Pr(T_k(\bar{M}) \leq \tau_k | X = x^*) \leq F'_{\theta|X}(t_k + (3r_3^*/\epsilon N)^{\beta_N^{-1}} | x^*) + 2K \exp(-2\kappa^2 \log(J))
\]

Combining these results,

\[
|F_{\theta|X}(t_k | x^*) - F'_{\theta|X}(t'_k(x^*))| \leq 4K \exp(-2\kappa^2 \log(J)) + 2M_F(3r_3^*/\epsilon N)^{\alpha_F \beta_N^{-1}}
\]

\[
\leq (r_3^*)^{\alpha_F \beta_N^{-1}}
\]

where the asymptotic bound is valid if \( \kappa \) is large enough.

\textit{Q.E.D.}

The proof of Theorem 4.1 follows.

\textbf{PROOF:} Consider two \( J \)-observationally equivalent models. Fix any \( x \in \mathcal{X}, t \in [0,1]^K \) and define \( m = \bar{p}(x, t) \). According to Lemma 3 there must be \( t'(x) \in [0,1]^K \) such that

\[ ||p'(x, t'(x)) - m|| \leq \rho_0 := \kappa_0 (\log(J)/J)^{1/2} \]

for some \( \kappa_0 > 0 \).

As shown in the proof of Theorem 3.2, if \( ||t - t'(x)|| \leq \psi \) then

\[ |g(x, t, u) - g'(x, t, u)| \leq (r_1^*)^\alpha_y + \psi^\alpha_y, |h(x, t) - h'(x, t)| \leq r_1^* + \psi^\alpha_h, \] and \( |\bar{g}(x, t) - \bar{g}'(x, t)| \leq r_1^* + \psi^\alpha_h \). So then remains to find \( \psi \) such that \( ||t - t'(x)|| < \psi \). I will use the result from the proof of Theorem 3.2 that for any \( x_1 \in \mathcal{X}, \sup_{x_2 \in \mathcal{X}} |t'(x_2) - t'(x_1)| \leq (r_2^*)^\beta_{k^{-1}} \), which is still valid under the assumptions of Theorem 4.1.
By assumption \( \theta \sim Uniform(0,1) \) and, as a result, for any \( s_1 \in \mathbb{R} \), \( s_1 = F_{\theta_1}(s_1) = \int_{\mathcal{X}} F_{\theta_1|X}(s_1 \mid x^*)dF_X(x^*) \). The same thing is of course true if \( F_{\theta_1|X} \) is replaced by the observationally equivalent \( F'_{\theta_1|X} \). Therefore,

\[
|t_1 - t'_1(x)| = \left| \int_{\mathcal{X}} F_{\theta_1|X}(t_1 \mid x^*)dF_X(x^*) - \int_{\mathcal{X}} F'_{\theta_1|X}(t'_1(x) \mid x^*)dF_X(x^*) \right|
\]

\[
\leq \left| \int_{\mathcal{X}} F_{\theta_1|X}(t_1 \mid x^*)dF_X(x^*) - \int_{\mathcal{X}} F'_{\theta_1|X}(t'_1(x^*) \mid x^*)dF_X(x^*) \right|
\]

\[
+ \left| \int_{\mathcal{X}} F'_{\theta_1|X}(t'_1(x^*) \mid x^*)dF_X(x^*) - \int_{\mathcal{X}} F''_{\theta_1|X}(t''_1(x) \mid x^*)dF_X(x^*) \right|
\]

which implies that

\[
|t_1 - t'_1(x)| \leq \sup_{x^* \in \mathcal{X}} |F_{\theta_1|X}(t_1 \mid x^*) - F'_{\theta_1|X}(t'_1(x^*) \mid x^*)| + (4r^*_\theta / \epsilon_E)^{\beta_E^{-1}}
\]

The argument used in the proof of Theorem 3.2 can be applied to show that \( \sup_{x^* \in \mathcal{X}} |F_{\theta_1|X}(t_1 \mid x^*) - F'_{\theta_1|X}(t'_1(x^*) \mid x^*)| \leq (r^*_\theta)^{\beta_N^{-1}} \). (n.b.: The exponent \( \alpha_F \) does not appear in the rate here as it did in Theorem 3.2 because under Assumption 4.4(i), \( \alpha_F = 1 \).) In fact, this holds regardless of the value of \( t \) initially chosen. That is, or any \( s \in [0, 1]^K \), if \( s' \) is defined such that

\[
||p(x, s) - \bar{p}'(x, s')|| \leq \kappa(\log(J)/J)^{1/2} \text{ then } |s_1 - s_1'| \leq (r^*_\theta)^{\beta_E^{-1}} + (r^*_\theta)^{\beta_N^{-1}} \] and therefore

\[
|p_{N1}(x, s) - p'_{N1}(x, s)| \leq |p_{N1}(x, s) - p'_{N1}(x, s')| + |p'_{N1}(x, s') - p'_{N1}(x, s)|
\]

\[
\leq r^*_s := (r^*_\theta)^{\beta_E^{-1}} \alpha_N + (r^*_\theta)^{\beta_N^{-1}} \alpha_N
\]

where the second line follows because \( p'_{N1}(x, s') = p'_{N1}(x, s_1', s_2) \).

On the other hand, since \( \theta_2 \mid \theta_1 \sim Uniform(0,1) \), for any \( s \in \mathbb{R}^2 \),

\[
s_2 = F_{\theta_2|\theta_1}(s_2 \mid s_1)
\]

\[
= \int_{\mathcal{X}} F_{\theta_2|\theta_1,X}(s_2 \mid s_1, x^*)dF_{X|\theta_1}(x^* \mid s_1)
\]

\[
= \int_{\mathcal{X}} F_{\theta_2|\theta_1,X}(s_2 \mid s_1, x^*)f_{\theta_1|X}(s_1 \mid x^*)dF_X(x^*)
\]
and therefore,
\begin{align*}
|t_2 - t'_2(x)| &\leq \sup_{x^* \in \mathcal{X}} |F_{\theta_2|\theta_1,X}(t_2 \mid t_1, x^*) f_{\theta_1|X}(t_1 \mid x^*) - F'_{\theta_2|\theta_1,X}(t_2'(x^*) \mid t_1'(x^*), x^*) f'_{\theta_1|X}(t_1'(x^*) \mid x^*)| \\
&+ \sup_{x^* \in \mathcal{X}} |F'_{\theta_2|\theta_1,X}(t_2'(x^*) \mid t_1'(x^*), x^*) f'_{\theta_1|X}(t_1'(x^*) \mid x^*) - F''_{\theta_2|\theta_1,X}(t_2'(x) \mid t_1'(x), x) f''_{\theta_1|X}(t_1'(x) \mid x)| \\
&\leq \sup_{x^* \in \mathcal{X}} |F_{\theta_2|\theta_1,X}(t_2 \mid t_1, x^*) f_{\theta_1|X}(t_1 \mid x^*) - F'_{\theta_2|\theta_1,X}(t_2'(x^*) \mid t_1'(x^*), x^*) f'_{\theta_1|X}(t_1'(x^*) \mid x^*)| \\
&+ (4r_2^*/\epsilon_E)^{\alpha_F^3}\beta_E^{-1}
\end{align*}

So the last step is to derive an upper bound on
\[
\sup_{x \in \mathcal{X}} |F_{\theta_2|\theta_1,X}(t_2 \mid t_1, x) f_{\theta_1|X}(t_1 \mid x) - F'_{\theta_2|\theta_1,X}(t_2'(x) \mid t_1'(x), x) f'_{\theta_1|X}(t_1'(x) \mid x)|
\]

Define $\tau_2 = p_{N2}(x,t)$. Recall that $T_2(m) := E(M_{N2} \mid ||\bm{M} - m|| < \rho_2, X = x)$ where $\rho_2 := \kappa_2(log(J)/J)^{1/2}$ is distinct from the $\rho_1 = \kappa_1(log(J)/J)^{1/2}$ in the definition of $T_1(m)$. Henceforth I will drop the last argument, $t_2$, from the function $p_{N1}$ to simplify notation since, by Assumption 4.2(iv) it does not vary in this argument. Also, let $p_{N1}^{-1}(u; x)$ denote the unique inverse function that takes the value 1 for $u > p_{N1}(x, 1)$ and the value 0 for $u < p_{N1}(x, 0)$.

- I will first show that for $\rho > 0$ and $\zeta > 0$ such that $\min\{\rho \zeta, M_N(\epsilon_N^{-1}(1-\zeta))^{2 / \alpha_N}\} > r_3^*$

\[
(A.5) \quad |Pr(T_2(\bm{M}) \leq \tau_2, |T_1(\bm{M}) - \tau_1| < \rho \mid X = x) \\
- F_{\theta_2|\theta_1,X}(t_2 \mid t_1, x) f_{\theta_1|X}(t_1 \mid x)|p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{-1}(p_{N1}(x, t_1) - \rho; x)|| \\
\leq r_4^*|p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{-1}(p_{N1}(x, t_1) - \rho; x)| \\
+ |Pr([p_{N1}(x, \theta_1) - \tau_1] < \rho(1 + \zeta) \mid X = x) - Pr([p_{N1}(x, \theta_1) - \tau_1] < \rho(1 - \zeta) \mid X = x)| \\
+ M_f|p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{-1}(p_{N1}(x, t_1) - \rho; x)|^{1 + \alpha_F^{(1)}}
\]

where $r_4^* = (\rho(1 + \zeta))^{2 / \alpha_N \alpha_F^{(2)}}$ and likewise,

\[
(A.6) \quad |Pr(T_2(\bm{M}) \leq \tau_2, |T_1(\bm{M}) - \tau_1| < \rho \mid X = x) \\
- F'_{\theta_2|\theta_1,X}(t_2'(x) \mid t_1'(x), x)|p_{N1}^{-1}(p_{N1}'(x, t_1') + \rho; x) - p_{N1}^{-1}(p_{N1}'(x, t_1') - \rho; x)|| \\
\leq r_4^*|p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{-1}(p_{N1}(x, t_1) - \rho; x)| \\
+ |Pr'([p_{N1}'(x, \theta_1) - \tau_1] < \rho(1 + \zeta) \mid X = x) - Pr'([p_{N1}'(x, \theta_1) - \tau_1] < \rho(1 - \zeta) \mid X = x)| \\
+ M_f|p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{-1}(p_{N1}(x, t_1) - \rho; x)|^{1 + \alpha_F^{(1)}}
\]
If these two inequalities hold then it follows that

\[ |F_{\theta_2|t_2, X}(t_2 \mid t_1, x) - F_{\theta_2|t_1, X}(t_1 \mid x)| \]
\[ \leq r^*_4 + |p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{-1}(p_{N1}(x, t_1) - \rho; x)|^{\alpha_1^{(1)}(t)} \]
\[ + |Pr(|p_{N1}(x, \theta_1) - \tau_1| < \rho(1 + \zeta) \mid X = x) - Pr(|p_{N1}(x, \theta_1) - \tau_1| < \rho(1 - \zeta) \mid X = x)| \]
\[ + \frac{|p_{N1}^{-1}(p_{N1}(x, t_1') + \rho; x) - p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x)|}{p_{N1}^{-1}(p_{N1}(x, t_1') - \rho; x) - p_{N1}^{-1}(p_{N1}(x, t_1) - \rho; x)} \]
\[ + \frac{|p_{N1}^{-1}(p_{N1}(x, t_1') - \rho; x) - p_{N1}^{-1}(p_{N1}(x, t_1') - \rho; x)|}{p_{N1}^{-1}(p_{N1}(x, t_1') + \rho; x) - p_{N1}^{-1}(p_{N1}(x, t_1') - \rho; x)} \]

Before deriving (A.5) and (A.6) I will work with each of the terms following \( r^*_4 \) in the bound above. First, by Assumption 2.7(iv),

\[ |p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{-1}(p_{N1}(x, t_1) - \rho; x)|^{\alpha_1^{(1)}} \leq \rho^{a_1^{(1)}(t)} \]

Next, since this same factor is in the denominator of the other terms a lower bound is needed.

\[ |p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{-1}(p_{N1}(x, t_1) - \rho; x)| \geq \rho^{a_1^{(1)}} \]

Next, by an argument developed below,

\[ Pr(|p_{N1}(x, \theta_1) - \tau_1| < \rho(1 + \zeta) \mid X = x) \leq |p_{N1}^{-1}(\tau_1 + \rho(1 + \zeta)) - p_{N1}^{-1}(\tau_1 - \rho(1 + \zeta))| \]

and

\[ Pr(|p_{N1}(x, \theta_1) - \tau_1| < \rho(1 - \zeta) \mid X = x) \leq |p_{N1}^{-1}(\tau_1 + \rho(1 - \zeta)) - p_{N1}^{-1}(\tau_1 - \rho(1 - \zeta))| \]

Since \( \rho(1 + \zeta) - \rho(1 - \zeta) = 2\rho\zeta \) it follows that

\[ |Pr(|p_{N1}(x, \theta_1) - \tau_1| < \rho(1 + \zeta) \mid X = x) - Pr(|p_{N1}(x, \theta_1) - \tau_1| < \rho(1 - \zeta) \mid X = x)| \]
\[ \leq (\rho\zeta)^{a_1^{(1)}} \]
Now consider the final two terms. First,

\[ |p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{t_1}(p_{N1}(x, t_1) + \rho; x)| \]
\[ \leq |p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{t_1}(p_{N1}(x, t_1) + \rho; x)| \]
\[ + |p_{N1}^{t_1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{t_1}(p_{N1}(x, t_1') + \rho; x)| \]
\[ \leq |p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{t_1}(p_{N1}(x, t_1) + \rho; x)| + |p_{N1}(x, t_1) - p_{N1}^{t_1}(x, t_1')|^{\beta_N^{-1}} \]
\[ \leq |p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{t_1}(p_{N1}(x, t_1) + \rho; x)| + (r_5^*)^{\beta_N^{-1}} \]

so I will now consider the first term on the final line separately.

Suppose \( p_{N1}(x, t_1) + \rho = p_{N1}(x, t_1') \). This means that

\[ |p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{t_1}(p_{N1}(x, t_1) + \rho; x)| \]
\[ \leq |p_{N1}^{t_1}(p_{N1}(x, t_1') + \rho; x) - p_{N1}^{t_1}(p_{N1}(x, t_1') + \rho; x)| \]
\[ \leq |p_{N1}^{t_1}(x, t_1') - p_{N1}(x, t_1')|^{\beta_N^{-1}} \]
\[ \leq (r_5^*)^{\beta_N^{-1}} \]

where the final line follows from (A.4) above. If there is no \( t_1' \in [0, 1] \) such that \( p_{N1}(x, t_1) + \rho = p_{N1}(x, t_1') \) then \( p_{N1}(x, t_1) + \rho > 1 \). By definition, \( p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) = 1 = p_{N1}^{t_1}(p_{N1}(x, t_1) + \rho; x) \) so

\[ |p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{t_1}(p_{N1}(x, t_1) + \rho; x)| \]
\[ \leq |p_{N1}^{t_1}(p_{N1}(x, 1) + \rho; x) - p_{N1}(x, 1) + \rho; x)| \]

Then, \( p_{N1}(x, t_1) + \rho > p_{N1}(x, 1) > p_{N1}^{t_1}(x, 1) - \kappa r_5^* \) for some constant \( \kappa > 0 \). Thus, either \( p_{N1}(x, t_1) + \rho > p_{N1}(x, 1) \), in which case \( p_{N1}^{t_1}(p_{N1}(x, t_1) + \rho; x) = 1 \) by definition, or \( |p_{N1}(x, t_1) + \rho - p_{N1}(x, 1)| \leq r_5^* \). Therefore, it is still the case that

\[ |p_{N1}^{-1}(p_{N1}(x, t_1) + \rho; x) - p_{N1}^{t_1}(p_{N1}(x, t_1) + \rho; x)| \]
\[ \leq (r_5^*)^{\beta_N^{-1}} \]
Combining the above results,

\[ |F_{\theta_2|\theta_1,X}(t_2 \mid t_1, x)f_{\theta_1|X}(t_1 \mid x) - F'_{\theta_2|\theta_1,X}(t_2' \mid t_1', x)f'_{\theta_1|X}(t_1' \mid x)| \]

\( \leq r_4^* + \rho^{\alpha(N - 1)} + (\rho\zeta)^{\beta N - 1} + (r_5^*)^{\alpha(N - 1)} \)

(A.7)

It remains prove that (A.5) and (A.6) hold.

- First,

\[ Pr(T_2(\tilde{M}) \leq \tau_2, |T_1(\tilde{M}) - \tau_1| < \rho \mid X = x) \]

\[ = \int_{\mathcal{R}_1} \xi(t^*)dF_{\theta|X=x}(t^*) + \int_{\mathcal{R}_2} \xi(t^*)dF_{\theta|X=x}(t^*) + \int_{\mathcal{R}_3} \xi(t^*)dF_{\theta|X=x}(t^*) \]

where

\( \mathcal{R}_1 := \{ t^* : |p_{N1}(x, t_1^*) - p_{N1}(x, t_1)| > \rho(1 + \zeta) \} \)

\( \mathcal{R}_2 := \{ t^* : |p_{N1}(x, t_1^*) - p_{N1}(x, t_1)| \leq \rho(1 + \zeta), p_{N2}(x, t_1, t_2^*) \leq p_{N2}(x, t_1, t_2) + \epsilon \} \)

\( \mathcal{R}_3 := \{ t^* : |p_{N1}(x, t_1^*) - p_{N1}(x, t_1)| \leq \rho(1 + \zeta), p_{N2}(x, t_1, t_2^*) > p_{N2}(x, t_1, t_2) + \epsilon \} \)

and

\[ \xi(t^*) := Pr(T_2(\tilde{M}) \leq \tau_2, |T_1(\tilde{M}) - \tau_1| < \rho \mid X = x, \theta = t^*) \]

I will bound each term in this decomposition separately.

\[ \int_{\mathcal{R}_1} \xi(t^*)dF_{\theta|X=x}(t^*) \]

\[ \leq \int_{\mathcal{R}_1} Pr(|T_1(\tilde{M}) - \tau_1| < \rho \mid X = x, \theta = t^*)dF_{\theta|X=x}(t^*) \]

\[ \leq \int_{\mathcal{R}_1} Pr(|T_1(\tilde{M}) - p_{N1}(x, \theta_1)| > \rho\zeta \mid X = x, \theta = t^*)dF_{\theta|X=x}(t^*) \]

\[ \leq \int_{\mathcal{R}_1} Pr(|\tilde{M} - p(x, \theta)| > \kappa_1(\log(J)/J)^{1/2} \mid X = x, \theta = t^*)dF_{\theta|X=x}(t^*) \]

(A.9) \( \leq 2K \exp(-2\kappa_1^2 \log(J)) \)

where the third inequality follows if \( \rho\zeta > r_3^* \) and the fourth follows from Lemma 2.
For the third term in (A.8)

\[
\int_{\mathcal{R}_3} \xi(t^*) dF_{\theta|X=x}(t^*) \\
\leq \int_{\mathcal{R}_3} Pr(T_2(\bar{M}) \leq \tau_2 \mid X=x, \theta = t^*) dF_{\theta|X=x}(t^*) \\
\leq \int_{\mathcal{R}_3} Pr(T_2(\bar{\tilde{M}}) \leq p_{N^2}(x, t_1, t_2^*) - \epsilon \mid X=x, \theta = t^*) dF_{\theta|X=x}(t^*) \\
\leq \int_{\mathcal{R}_3} Pr(T_2(\bar{\tilde{M}}) \leq p_{N^2}(x, t_1^*, t_2^*) - \tilde{\epsilon} \mid X=x, \theta = t^*) dF_{\theta|X=x}(t^*)
\]

where \( \tilde{\epsilon} = \epsilon - M_N(\epsilon_N^{-1}\rho(1+\zeta))^{\beta_N^{-1}\alpha_N} \) and the last inequality follows because \(|t_1 - t_1^*| < (\epsilon_N^{-1}|p_{N1}(x, t_1^*) - p_{N1}(x, t_1)|)^{\beta_N^{-1}} < (\epsilon_N^{-1}\rho(1+\zeta))^{\beta_N^{-1}} \), which implies that \(p_{N^2}(x, t_1, t_2^*) < p_{N^2}(x, t_1^*, t_2^*) + M_N(\epsilon_N^{-1}\rho(1+\zeta))^{\beta_N^{-1}\alpha_N} \). Take \( \epsilon = 2M_N(\epsilon_N^{-1}\rho(1+\zeta))^{\beta_N^{-1}\alpha_N} \), and assuming that \(M_N(\epsilon_N^{-1}\rho(1+\zeta))^{\beta_N^{-1}\alpha_N} > r_3^2 \), it follows that

\[
\int_{\mathcal{R}_3} \xi(t^*) dF_{\theta|X=x}(t^*) \\
\leq \int_{\mathcal{R}_3} Pr(|T_2(\bar{\tilde{M}}) - p_{N^2}(x, \theta)| > M_N(\epsilon_N^{-1}\rho(1+\zeta))^{\beta_N^{-1}\alpha_N} \mid X=x, \theta = t^*) dF_{\theta|X=x}(t^*) \\
\leq \int_{\mathcal{R}_3} Pr(|\bar{\tilde{M}} - \bar{p}(x, \theta)| > \kappa_2(\log(J)/J)^{1/2} \mid X=x, \theta = t^*) dF_{\theta|X=x}(t^*)
\]

(A.10) \( \leq 2K \exp(-2\kappa_2^2 \log(J)) \)

And for the second term in (A.8),

\[
\int_{\mathcal{R}_2} \xi(t^*) dF_{\theta|X=x}(t^*) \\
\leq \int_{\mathcal{R}_2} Pr(p_{N^2}(x, t_1, \theta_2) \leq p_{N^2}(x, t_1, t_2) + \epsilon \mid X=x, \theta_1 = t_1^*) dF_{\theta|X=x}(t^*) \\
\leq F_{\theta_2|\theta_1}(t_2 + (\epsilon/\epsilon_N)^{\beta_N^{-1}} \mid t_1, x) Pr(|p_{N1}(x, \theta_1) - \tau_1| < \rho(1+\zeta) \mid X=x) \\
\leq F_{\theta_2|\theta_1}(t_2 \mid t_1, x) Pr(|p_{N1}(x, \theta_1) - \tau_1| < \rho \mid X=x) \\
\leq M_F(\epsilon_N/\epsilon_N)^{\beta_N^{-1}\alpha_F^{(2)}} Pr(|p_{N1}(x, \theta_1) - \tau_1| < \rho \mid X=x) \\
+ \{Pr(|p_{N1}(x, \theta_1) - \tau_1| < \rho(1+\zeta) \mid X=x) - Pr(|p_{N1}(x, \theta_1) - \tau_1| < \rho \mid X=x)\}
\]
On the other hand, define

\[ \mathcal{R}_2^* = \{ t^* : |p_{N1}(x, t_1^*) - p_{N1}(x, t_1)| \leq \rho (1 - \zeta), p_{N2}(x, t_1, t_2^*) \leq p_{N2}(x, t_1, t_2) - \epsilon \} \]

and then

\[
Pr(T_2(\bar{M}) \leq \tau_2, |T_1(\bar{M}) - \tau_1| < \rho \mid X = x) \\
\quad \geq \int_{\mathcal{R}_2^*} \xi(t^*)dF_{\theta|X=x}(t^*) \\
\quad \geq \int_{\mathcal{R}_2^*} (1 - Pr(T_2(\bar{M}) > \tau_2 \mid X = x, \theta = t^*))dF_{\theta|X=x}(t^*)
\]

(A.12) \quad -Pr(|T_1(\bar{M}) - \tau_1| \geq \rho \mid X = x, \theta = t^*))dF_{\theta|X=x}(t^*)

and each of these three terms can be treated using arguments similar to those just employed to provide an upper bound on \( Pr(T_2(\bar{M}) \leq \tau_2, |T_1(\bar{M}) - \tau_1| < \rho \mid X = x) \). Briefly,

\[
\int_{\mathcal{R}_2^*} Pr(T_2(\bar{M}) > \tau_2 \mid X = x, \theta = t^*)dF_{\theta|X=x}(t^*) \\
\leq \int_{\mathcal{R}_2^*} Pr(|T_2(\bar{M}) - p_{N2}(x, \theta_1, \theta_1)| \geq (\epsilon - M_N(\epsilon N^{-1}(1 - \zeta))^{\beta_{\alpha N}^{-1}}) \mid X = x, \theta = t^*)dF_{\theta|X=x}(t^*)
\]

Take \( \epsilon = 2M_N(\epsilon N^{-1}(1 - \zeta))^{\beta_{\alpha N}^{-1}} \), and assuming that \( M_N(\epsilon N^{-1}(1 - \zeta))^{\beta_{\alpha N}^{-1}} > r_{3*}^* \), it follows that

(A.13) \quad \int_{\mathcal{R}_2^*} Pr(T_2(\bar{M}) > \tau_2 \mid X = x, \theta = t^*)dF_{\theta|X=x}(t^*) \leq 2K \exp(-2\kappa_2^2 \log(J))

Next, if \( \rho \zeta > r_{3*}^* \),

\[
\int_{\mathcal{R}_2^*} Pr(|T_1(\bar{M}) - \tau_1| \geq \rho \mid X = x, \theta = t^*)dF_{\theta|X=x}(t^*) \\
\leq \int_{\mathcal{R}_2^*} Pr(|T_1(\bar{M}) - p_{N1}(x, \theta_1)| \geq \rho \zeta \mid X = x, \theta = t^*)dF_{\theta|X=x}(t^*)
\]

(A.14) \quad \leq 2K \exp(-2\kappa_1^2 \log(J))
And lastly

\[
\int_{\mathbb{R}_+^2} dF_{\theta|X=x}(t^*) \\
\geq \sum_{p_{N_1}(x,t^*) < p_{N_1}(x,t_1)} F_{\theta_2|\theta_1,X}(t_2 - (\epsilon/\epsilon_N)^{\beta_N^1} | t_1, x)dF_{\theta_1|X=x}(t_1^*) \\
\geq F_{\theta_2|\theta_1,X}(t_2 | t_1, x) Pr([p_{N_1}(x, \theta_1) - \tau_1 < \rho | X = x) \\
\geq M_f(\epsilon/\epsilon_N)^{\beta_N^1 a_F^1(2)} Pr([p_{N_1}(x, \theta_1) - \tau_1 < \rho | X = x) \\
\geq \{ Pr([p_{N_1}(x, \theta_1) - \tau_1 < \rho | X = x) - Pr([p_{N_1}(x, \theta_1) - \tau_1 < \rho(1 - \zeta) | X = x) \}
\]

Together (A.8)-(A.15) imply that

\[
|Pr(T_2(\bar{M}) \leq \tau_2, [T_1(\bar{M}) - \tau_1 < \rho | X = x) \\
-F_{\theta_2|\theta_1,X}(t_2 | t_1, x) Pr([p_{N_1}(x, \theta_1) - \tau_1 < \rho | X = x) \\
\leq r_3^* Pr([p_{N_1}(x, \theta_1) - \tau_1 < \rho | X = x) \\
+ |Pr([p_{N_1}(x, \theta_1) - \tau_1 < \rho(1 + \zeta) | X = x) - Pr([p_{N_1}(x, \theta_1) - \tau_1 < \rho(1 - \zeta) | X = x) | \\
+ \exp(-2\kappa_1^2 \log(J)) + \exp(-2\kappa_2^2 \log(J))
\]

Next, by the monotonicity of \(p_{N_1}\) and the definition of \(p_{N_1}^{-1}\), \(Pr([p_{N_1}(x, \theta_1) - \tau_1 < \rho | X = x) = F_{\theta_1|X}(p_{N_1}^{-1}(\tau_1 + \rho; x) | x) - F_{\theta_1|X}(p_{N_1}^{-1}(\tau_1 - \rho; x) | x)\). Using 2.7(i) this can be rewritten using a first order Taylor approximation as \(f_{\theta_1|X}(p_{N_1}^{-1}(\tau_1; x))(p_{N_1}^{-1}(\tau_1 + \rho; x) - p_{N_1}^{-1}(\tau_1 - \rho; x)) + R^*\) where \(R^* \leq M_f(p_{N_1}^{-1}(\tau_1 + \rho; x) - p_{N_1}^{-1}(\tau_1 - \rho; x))^{1+a_F^1(3)}\). Since \(\tau_1 = p_{N_1}(x, t_1)\), the bound in (A.5) follows if \(\kappa_1\) and \(\kappa_2\) are chosen to be sufficiently large.

Then (A.6) follows a similar argument. Since \(|p_{N_1}(x, t_1) - p_{N_1}'(x, t')| \leq r_3^* \) and \(|p_{N_2}(x, t) - p_{N_2}'(x, t')| \leq r_3^*\), it can be shown that

\[
|Pr(T_2(\bar{M}) \leq \tau_2, [T_1(\bar{M}) - \tau_1 < \rho | X = x) \\
-F_{\theta_2|\theta_1,X}(t_2 | t_1, x) Pr([p_{N_1}'(x, \theta_1) - p_{N_1}'(x, t_1) < \rho | X = x) \\
\leq r_3^* Pr([p_{N_1}'(x, \theta_1) - p_{N_1}'(x, t_1) < \rho | X = x) \\
+ |Pr([p_{N_1}'(x, \theta_1) - p_{N_1}'(x, t_1) < \rho(1 + \zeta) | X = x) - Pr([p_{N_1}'(x, \theta_1) - p_{N_1}'(x, t_1) < \rho(1 - \zeta) | X = x) | \\
+ \exp(-2\kappa_1^2 \log(J)) + \exp(-2\kappa_2^2 \log(J))
\]
The asymptotic bound in (A.7) is minimized by choosing $\rho \zeta = r^*_3$ and $\rho = \kappa (\log(J)/J)^{1/2} \gamma_1/(\gamma_2 + \alpha N)$ where

$$\gamma_1 = \min\{\alpha / \beta, \alpha_E / \beta_E\} \frac{\alpha N}{\beta N} \beta^{-1}$$

and

$$\gamma_2 = \min\{\alpha_F(1), \alpha_F(2) / \beta N\} \beta^{-1}$$

This can be seen since

$$r^*_4 + \rho^{\alpha F(1) \beta N^{-1}} + (\rho \zeta)^{\beta N^{-1}} \rho^{-\alpha N^{-1}} + (r^*_5)^{\beta N^{-1}} \rho^{-\alpha N^{-1}} \leq \rho^{\gamma_2} + (\log(J)/J)^{1/2} \gamma_1 \rho^{-\alpha N^{-1}}$$

Q.E.D.