A non-parametric discrete instrumental variable model with application to female labour force participation*

Jeff Rowley[1,a]

[1]*University College London (Department of Economics),* [a]*jeffreyrowley@duck.com*

Abstract. This paper studies a non-parametric discrete instrumental variable model that embeds a conditional statistical independence condition—heterogeneity is statistically independent of instruments conditional upon covariates. The model is applicable when data comprises a dichotomous response, a dichotomous treatment, and a discrete instrumental variable; the model also allows for covariates to influence both the response decision and the treatment decision. This paper shows that the model partially identifies the conditional average treatment effect and its aggregation (i.e., linear combinations of this parameter for different conditioning levels), and that it is falsifiable. This paper also adopts a robust Bayesian (Giacomini & Kitagawa, 2021) posture to provide credible regions for the effect of additional children on female labour force participation using U.S. data from 1980 and 2008–2018 under weak restrictions on behaviour.

JEL subject classifications: C260, J130, J220
Keywords: Female labour force participation, Fertility, Instrumental variables, Partial (or set) identification, Robust Bayesian inference

Minimally restrictive economic modelling—the deliberate stripping back of the restrictions that a model embeds to the bare minimum as are needed to guarantee identification to some degree—potentially offers a middle ground to more conventional alternatives. A structural approach à la Heckman (1978) can (point) identify population average effects but relies upon parametric and strong assumptions about underlying behaviour in order to do so. A non-structural approach à la Imbens & Angrist (1994) does not rely upon strong assumptions about underlying behaviour but can only (point) identify local average effects that cannot inform *ex-ante* analysis. Advanced by Manski—and subsequently by others—what the broad class of minimally restrictive economic models *can* and *do* deliver in practice is, arguably, not well understood; we address this, proposing a discrete instrumental variable model that we apply to data. The model extends Balke & Pearl's pioneering contribution to the study of imperfect compliance (Balke & Pearl, 1997) in ways that are of practical use—by allowing for discrete instruments and covariates to influence choice. We show that the model is partially identifying and is falsifiable.

We frame the model in the context of female labour force participation—although we emphasise that the model is presented in a way that is not specific to that context—using U.S. data to provide some insight into the nature of the relationship between fertility and the extensive (employment) margin. Understanding this relationship is important for the design and evaluation of

*Version: January 26, 2025.

policy; although family size is most often not a stated objective and is not necessarily (directly) manipulable in any case, knowledge of its influence is key to predicting present and future contributions and costs to the public finances—a subject that is, arguably, growing in importance given the well-documented decline in fertility in and ageing of developed (and in some developing) economies. Children are an important feature of the tax code and tax policy is often designed with families with children in mind—reporting indicates bipartisan support for increasing child support and childcare provision with the intention of improving the incentive to work amongst parents (The New York Times, 2024). We find that the model is insufficiently restrictive to infer whether additional children upon maternal employment increase or decrease maternal employment, and yields wide and largely uninformative estimates. We contrast this result with Angrist & Evans (1998), Chesher & Rosen (2013)—whose empirical strategy we largely follow—which find that additional children have a negative effect upon maternal employment in the aggregate—in Angrist & Evans's case amongst some unidentifiable subset of mothers, in Chesher & Rosen's case under the assumption that additional children either increase or decrease maternal employment but not both (a restriction that we do not impose).

We adopt a Bayesian posture towards the problem of inferring the relationship between fertility and the extensive (employment) margin. We follow the robust approach of Giacomini & Kitagawa (2021). In doing so, we explore how this novel method of conducting statistical inference performs in practice, finding that the Jeffreys prior—and, to a lesser extent, a uniform prior—favours rejection of the model. This property is a consequence of the testable implications that we derive and which are typically violated by extreme conditional distributions (i.e., probability distributions that are located at the extremities of the simplex). We report estimates (posterior means) of and confidence regions (robust credible regions) for the effect of additional children on maternal employment. We find that the model—and our choice of instruments and covariates—is plausible.

The main contribution of this research is theoretical. We provide sharp characterisations of identified sets that can be used as the basis for estimation of policy-relevant parameters such as the average treatment effect, or that can be used to test whether an instrumental variable restriction is incompatible with data. A secondary contribution of this research is practical. We explore the implementation of the robust approach of Giacomini & Kitagawa (2021), and we provide estimates of the effect of additional children upon maternal employment—augmenting existing results by considering recent data and assuming very little about underlying behaviour.

We proceed as follows. We introduce related papers and works (Section 1); we introduce the model (Section 2); we introduce an alternative representation of the model (Section 3); we show that the model is partially identifying and falsifiable (Section 4); we introduce several criteria of interest and discuss the information content of the model (Section 5); we adopt a Bayesian posture (Section 6); we implement the model (Section 7); and we conclude (Section 8). We also include several appendices. We prove various theoretical results (Appendix A); we prove various auxiliary theoretical results (Appendix B); and we provide supplementary information relating to our application (Appendix C).

1. Literature

Our work is motivated by Chesher & Rosen (2013, 2020), which examine the question of what models *can* and *do* deliver in practice. The purpose of those papers is instructive, and so they deliberately—and rightly—study a model that delivers identification regions that are easier to

characterise.[1] Our starting point is to wonder what the monotonicity restriction that is present in the analysis of Chesher & Rosen (2013, 2020) purchases—specifically, the restriction that a change in the level of treatment induces either an increase or a decrease in the level of response but not both. That is, what moving from a structural equation linking response to treatment, covariates and heterogeneity such as

$$y = 1(g(t, x, u) > 0) \tag{1.1}$$

—akin to the model that we study—to one such as

$$y = 1(g(t, x) - u > 0) \tag{1.2}$$

—akin to the model that is studied in Chesher & Rosen (2013, 2020)—purchases, through the way in which it limits interaction inside the index equation. Monotonicity excludes certain behaviours that are generated by theoretical models of collective decision-making. For instance, the model that we propose is compatible with the sorts of behaviour that are described in Gronau (1977), which advances theory on the allocation of time in the household to leisure, home production and work. Such behaviours are excluded by monotonicity—or weak separability, as it translates to.

The model that we propose generalises Balke & Pearl's pioneering contribution to the study of imperfect compliance (Balke & Pearl, 1997) by allowing discrete instruments and covariates to influence choice. Balke & Pearl's analysis involves a graphical framework. Richardson & Robins (2014) advances a variant of the graphical framework—the Single-World Intervention Graph (SWIG)—in extending Balke & Pearl's model to incorporate discrete instruments. White & Chalak (2009) advances another flexible graphical framework—settable systems. Beresteanu, Molchanov & Molinari (2012) shows how Balke & Pearl's sharp characterisation can otherwise be derived, without using linear programming and a graphical framework.

Kitagawa (2021) proposes an extension to Balke & Pearl's model, considering various statistical independence conditions as we do; Kitagawa (2021) allows for continuous response. Chesher & Rosen (2017) builds upon results from random set theory to provide general results and tools that are applicable to and can be used for the analysis of a broad class of models, including the one that we propose. Gunsilius (2019) advances a path sampling approach that is able to handle the difficult case in which treatment is continuous.

The model that we propose does not embed weak separability. Mourifié (2015) studies a triangular model (see Strotz & Wold, 1960) that embeds weak separability. Mourifié's analysis is robust to the failure of the support condition that is present in the analysis of Shaikh & Vytlacil (2011).

Pearl (1995) provides a testable inequality of Balke & Pearl's model; Bonet (2001) shows that this inequality is insufficient to detect all possible violations of the model. Rather, Kédagni & Mourifié (2020) provides a comprehensive set of testable inequalities that are necessary and sufficient to detect all possible violations of the Balke & Pearl's model. More than this, Kédagni & Mourifié (2020) provides further results that are applicable when the statistical independence condition that Balke & Pearl's model embeds is weakened; it also provides a guide to implementing these testable inequalities in practice contingent upon the adoption of a frequentist posture.

[1] To be clear, Chesher & Rosen (2013) studies a *non-parametric threshold-crossing model*; Chesher & Rosen (2020) studies this model and many others.

We adopt a Bayesian posture and follow the robust approach of Giacomini & Kitagawa (2021). Kline & Tamer (2016), Norets & Tang (2014) provide related approaches. Moon & Schorfheide (2012) shows that an important property that holds in conventional settings—that Bayesian credible regions have an asymptotically valid frequentist interpretation—does not necessarily hold in partially identifying settings. Although the credible regions that Giacomini & Kitagawa (2021) proposes can have a valid frequentist interpretation, Kitagawa et al. (2020) shows that Giacomini & Kitagawa's crucial differentiability requirement does not hold in our setting.

To implement the robust approach of Giacomini & Kitagawa (2021), we sample from a Dirichlet distribution. Asymptotically, as the quantity of data becomes large (and the prior exerts a vanishingly small influence upon the posterior), the Dirichlet distribution that we obtain resembles the (frequentist) bootstrap distribution. Andrews & Han (2009) shows that the (frequentist) bootstrap is not valid here. Andrews & Soares (2010), Bugni (2010), Chernozhukov, Lee & Rosen (2013) propose alternative frequentist approaches that do attain asymptotically correct coverage. Kaido, Molinari & Stoye (2019) proposes a (frequentist) bootstrap procedure to correct for projection conservatism. Projection conservatism—that confidence regions do not have the correct size, and contain the truth in more than the pre-specified proportion of samples—is of concern whenever criteria of interest are not identified directly but as the projection of an identified set of several parameters.

We use U.S. data to provide some insight into the nature of the relationship between fertility and the extensive (employment) margin. We build upon the empirical strategy of Angrist & Evans (1998), which exploits covariation in family composition and size. Blau et al. (2020) explores the evolution of parental preferences over male and female children, which is one source of identifying variation that Angrist & Evans (1998) exploits. Angrist & Evans (1998) adopts an approach that shuns the strong parametric or non-parametric restrictions that are needed to (point) identify (global) average effects—whereas we conduct minimally restrictive structural economic modelling—and finds that additional children have a negative effect on maternal employment in the aggregate. We note that countless other studies have considered the nature of the relationship between fertility and the extensive (employment) margin—whether in the context of the U.S. economy or of others; we draw attention to but a few of these. Comparable studies of the British (Iacovou, 2001) and Egyptian (Al-Khaja, 2016) labour markets find the opposite effect. Benny (2021) adopts a difference-in-differences approach to study the Tanzanian labour market; Benny (2021) exploits a reduction in child mortality due to a policy intervention (the increased availability and coverage of treated bed nets) that reduces malarial transmission and finds that additional children have a negative effect on maternal employment in the aggregate.

One concern that we do not address is that instruments can be mismeasured; for instance, what we encode as the occurrence of a multiple second birth can include children born more than nine months apart. Chalak (2017), Jiang & Ding (2020) study how mismeasured instruments affect estimators.

2. Framework

In what follows, we suppose the existence of a standard probability space over $\mathbb{R}$ (the real numbers) that is equipped with the Borel σ-algebra. We further suppose that the Axiom of Choice holds (Zermelo, 1904).

We consider a non-parametric discrete instrumental variable model—discrete in that economic variates have finite support on some set that we label $\mathfrak{F} \subset \mathbb{N}$ (the natural numbers, in-

clusive of zero; we use script font to distinguish sets). The model embeds restrictions on a probability distribution over a collection of exogenous economic variates and on the functions that map these to a collection of endogenous economic variates.

The endogenous economic variates comprise two outcomes. We denote the first outcome by $y \in \mathfrak{Y}_y$ and refer to this as response. We denote the second outcome by $t \in \mathfrak{Y}_t$ and refer to this as treatment. We denote the functions that map the exogenous economic variates to response and treatment by $\mathbf{h}$ (we use bold font to distinguish concatenations) and refer to these as the structural equations, with h_y and h_t denoting the response equation and the treatment equation, respectively. We maintain the following assumption on the structural equations throughout.

ASSUMPTION 1 (Binary exclusion). The model restricts $\mathbf{h}$ to belong to the collection of structural equations satisfying

$$\begin{aligned} y &= h_y(t, x, u) \\ t &= h_t(z, x, u) \end{aligned} \tag{2.1}$$

such that $\mathfrak{Y}_y = \{1, 0\}$ and $\mathfrak{Y}_t = \{1, 0\}$.

The exogenous economic variates comprise a collection of attributes—observable components of the model—and heterogeneity—an unobservable component of the model. We interpret attributes and heterogeneity as causes of response and treatment. Given that the structural equations are non-parametric, that we regard attributes and heterogeneity as scalars is without loss of generality.

We denote attributes by $\mathbf{a} \in \mathfrak{A}_{\mathbf{a}}$ and decompose these into two sub-collections. We denote the first sub-collection by $x \in \mathfrak{X}_x$ and refer to attributes in this collection as covariates; we denote the second sub-collection by $z \in \mathfrak{Z}_z$ and refer to attributes in this collection as instruments. We adopt the convention that attributes are the concatenation of covariates and instruments, and we write $\mathbf{e}_x^{\intercal}\mathbf{a} = x$ and $\mathbf{e}_z^{\intercal}\mathbf{a} = z$ to access each sub-collection separately (i.e., covariates and instruments can be extracted via multiplication with a transposed elementary vector). We emphasise that $\mathfrak{X}_x \equiv \{\mathbf{e}_x^{\intercal}\mathbf{a} : \mathbf{a} \in \mathfrak{A}_{\mathbf{a}}\}$ and $\mathfrak{Z}_z \equiv \{\mathbf{e}_z^{\intercal}\mathbf{a} : \mathbf{a} \in \mathfrak{A}_{\mathbf{a}}\}$ with $\mathfrak{A}_{\mathbf{a}} \subseteq \mathfrak{X}_x \times \mathfrak{Z}_z$ (i.e., not every combination of covariates and instruments need occur, and we restrict attention only to those that do)—which translates to $\mathfrak{X}_{x|z} \subseteq \mathfrak{X}_x$ or $\mathfrak{Z}_{z|x} \subseteq \mathfrak{Z}_z$ as the corresponding requirement upon the conditional supports.

We denote heterogeneity by $u \in \mathfrak{U}_u$. We interpret heterogeneity as capturing those aspects of the economic environment that influence response and treatment but that are—for practical reasons or by assumption—latent, and that can introduce randomness to the process (i.e., that different levels of response or treatment can occur for the same level of attributes).

We denote a joint distribution over attributes and heterogeneity by P and refer to this as the population. We maintain the following assumption on the population throughout.

ASSUMPTION 2 (Statistical independence—Random assignment). We let

$$\mathfrak{P}_{\text{RA}} \equiv \{P : u \perp\!\!\!\perp z | x\} \tag{2.2}$$

for convenience. The model restricts $\mathfrak{U}_u$ to satisfy

$$\log_2(|\mathfrak{U}_u|) = 2 \cdot |\mathfrak{X}_x| + |\mathfrak{A}_{\mathbf{a}}| \tag{2.3}$$

and restricts P to belong to the family of probability distributions satisfying $P \in \mathfrak{P}_{\text{RA}}$.

The implications of restricting the support of heterogeneity to a finite set of the specified cardinality are discussed in Section 3.

An admissible structure is a joint distribution over the economic variates—attributes and heterogeneity, and response and treatment—that is induced by the combination of a population that is compatible with Assumption 2 and structural equations that are compatible with Assumption 1. The model constitutes the class of all admissible structures (Hurwicz, 1950).

The policymaker—or the econometrician acting on her behalf—does not directly and fully observe the data generating process (i.e., the admissible structure that generates realisations of the economic variates). Rather, only certain elements of the economic environment are revealed to her. We denote a joint distribution over attributes, response and treatment by Q and refer to this as the observable distribution. We maintain the following assumption on the observable distribution throughout.

ASSUMPTION 3 (Information). The policymaker has knowledge of Q, which is induced by an admissible structure and satisfies $0 \leq Q(z|x) < 1$ for all $z, x \in \mathfrak{F}_{\mathbf{a}}$.

We emphasise the statistical independence conditions that the model embeds. These are (i.) $u \perp\!\!\!\perp z|x$, (ii.) $y \perp\!\!\!\perp z|t, x, u$ and (iii.) $z \not\perp\!\!\!\perp t$. Together, these conditions constitute a classical instrumental variable restriction—there is no mechanism by which instruments can affect response except via their influence on attributes or treatment—and are implied by Assumptions 1 and 2. We exploit this instrumental variable restriction to conduct the types of *ceteris paribus* enquiries that economists are wont to do (Frisch, 1995), and remark that our naming of the economic variates reflects this.

The model captures processes of the sort described in Chesher & Rosen (2021)—augmented by support restrictions on observable elements, and for specific forms of conditional statistical independence—with the role of *classifier* in that paper played here by covariates.

3. Generality

In Section 2 we present the model in its structural equation form. One notable feature of the model is that we regard heterogeneity as being distributed on a finite set. We now consider the strength of this assumption and, thereby, the generality of the model.

We reiterate that the structural equations describe the levels of response and treatment that materialise for prescribed levels of their determinants—including for hypothetical levels. Fixing a particular level of heterogeneity, we obtain the so-called *potential outcomes* (Holland, 1986, Neyman, 1990), which constitute counterfactuals and of which only one pair is realised (the pair that is induced by the realised level of attributes). Knowledge of the potential outcomes is, therefore, sufficient to trace-out the levels of response and treatment that materialise in *every* counterfactual state. We note, however, that there are only a finite number of counterfactual states—and so potential outcomes—and only a finite number of levels of response or treatment that can occur in each of them.

We take an *equivalence class* to mean a collection that is invariant across its members; here, for instance, as the collection of levels of heterogeneity that induce the same levels of response or treatment across all counterfactual states. To illustrate this and the separate point that Assumption 1 has a threshold crossing representation (Vytlacil, 2002), we briefly suspend the requirement that heterogeneity is finite—and instead suppose that it is real-valued—and write

Assumption 1 as

$$\begin{aligned} y &= 1(g_y(t,x,u) > 0) \\ t &= 1(g_t(z,x,u) > 0) \end{aligned} \tag{3.1}$$

such that identification of $\mathbf{g}$ is akin to identification of $\mathbf{h}$. Then

$$\{v : 1(g_y(t,x,u) > 0) = 1(g_y(t,x,v) > 0) \,\forall\, t,x \in \mathfrak{F}_t \times \mathfrak{F}_x \,|\, u \in \mathbb{R}\} \tag{3.2}$$

describes the collection of heterogeneity that induce the same levels of response across all counterfactual states as a specific level of heterogeneity does, and

$$\{v : 1(g_t(z,x,u) > 0) = 1(g_t(z,x,v) > 0) \,\forall\, z,x \in \mathfrak{F}_{\mathbf{a}} \,|\, u \in \mathbb{R}\} \tag{3.3}$$

describes the collection of heterogeneity that induce the same levels of treatment across all counterfactual states as a specific level of heterogeneity does, with the intersection of eqs. (3.2) and (3.3) then well-defined and constituting an equivalence class. Although there are uncountably many levels of heterogeneity (since $\mathbb{R}$ is uncountably infinite) in this thought-exercise, the number of disjoint sets—or equivalence classes—that eqs. (3.2) and (3.3) traces out (by cycling over $u \in \mathbb{R}$) is finite and coincides with the cardinality of the support of heterogeneity that is specified in eq. (2.3). The implication is that Assumption 2 does not impose any constraint upon the behaviour of the economic agents that the model is intended to capture. This point is made even more concrete by noting that eq. (2.3) also describes the number of possible functions that map from attributes and heterogeneity to response and treatment when these two variates are binary (or dichotomous).

Remark 1. Saturation of the endogenous economic variates requires that heterogeneity—howsoever it is defined—can be collected into a finite number of equivalence classes (Balke & Pearl, 1997). Provided that the support of heterogeneity is sufficiently rich relative to the support of attributes then the representation of the model that is presented (i.e., with discrete heterogeneity) is general. An implication is that if attributes were too rich then heterogeneity could not be defined on a standard probability space without further restriction on the structural equations (Gunsilius, 2019).[2]

The point here is that heterogeneity can be supported on a finite set, even if the latent characteristics that underpin it are continuous (i.e., infinite). We emphasise the plural here; any high-dimensional set that is isomorphic to $\mathbb{R}$—which includes a general Euclidean space—can replace $\mathbb{R}$ in eqs. (3.2) and (3.3). The implication is that heterogeneity can be interpreted as capturing *any and all* of those aspects of the economic environment that influence response and treatment but that are—for practical reasons or by assumption—latent.

Separately, we categorise attributes and heterogeneity as exogenous economic variates but we do not preclude the possibility that there is some causal relationship between them. We require that there is no *direct* relationship between instruments and heterogeneity (Assumption 2), and we require that any relationship between instruments and covariates is not deterministic (Assumption 3). The model is, for instance, compatible with heterogeneity causing covariates or *vice versa*, or for these to have a common cause (spurious correlation).

[2] If attributes were infinite then—under the Axiom of Choice—the number of equivalence classes that heterogeneity could be collected into would also be infinite, and would exceed the number of Borel sets on the continuum.

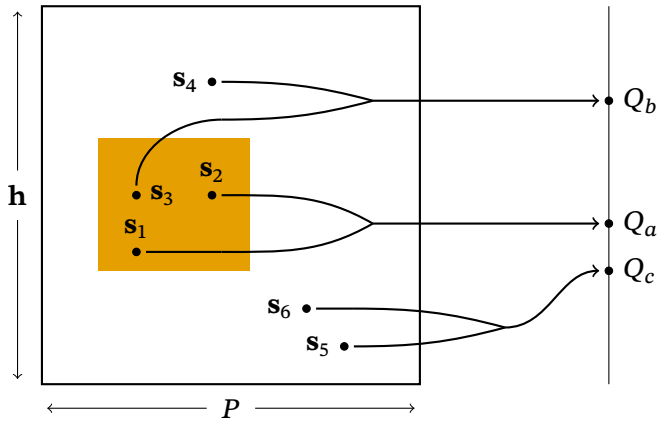


FIG 1. What it means for a model to be partially identifying and falsifiable—several compatible structures (combinations of P and $\mathbf{h}$ that are inside the shaded region) map to the same observable distribution (e.g., $\mathbf{s}_1$ and $\mathbf{s}_2$ map to Q_a; the model is partially identifying) and are said to be *observationally equivalent*, but some observable distributions are delivered by structures that are incompatible with the model (e.g., only $\mathbf{s}_5$ and $\mathbf{s}_6$ map to Q_c; the model is falsifiable). It is possible—and, indeed, likely—that structures that are and are not compatible with the model are observationally equivalent (e.g., $\mathbf{s}_3$ and $\mathbf{s}_4$ map to Q_b).

4. Identification

The identification problem is to infer knowledge of the data generating process from knowledge of the observable distribution. The model has complete identification power if every observable distribution is induced by a finite or countably infinite number of admissible structures and otherwise has incomplete identification power (Hurwicz, 1950). If there exist observable distributions that cannot be induced by any admissible structure then the model is falsifiable. We illustrate these concepts in fig. 1. To establish whether the model exhibits either property requires that we characterise those admissible structures that induce a given observable distribution.

To facilitate this identification analysis, we define—compressing our notation here and in what follows by collecting the constituent elements of attributes together wherever it is possible and convenient to do so—

$$\begin{aligned} \text{Contour}(y,t,\mathbf{a}) &\equiv \{u : y = h_y(t,x,u)\,\&\,t = h_t(\mathbf{a},u)\} \\ \text{Capacity}(Q,\mathfrak{U},\mathbf{a}) &\equiv \textstyle\sum_{y,t\in\mathfrak{F}_y\times\mathfrak{F}_t} Q(y,t|\mathbf{a})\cdot 1(\text{Contour}(y,t,\mathbf{a})\cap\mathfrak{U}\neq\varnothing) \end{aligned} \tag{4.1}$$

which we emphasise are functionals that depend implicitly upon the population and structural equations,[3] and where $\mathfrak{U}\subseteq\mathfrak{F}_u$. These functionals map combinations of attributes and heterogeneity to combinations of response and treatment, and *vice versa*. We obtain the following theorem, which characterises the collection of admissible structures that are compatible with Q and that we refer to as the identification region.

THEOREM 1 (Chesher & Rosen, 2017). *We maintain the discrete instrumental variable model that comprises Assumptions 1 to 3. The discrete instrumental variable model identifies*

$$IR(\mathfrak{P},Q) \equiv \{P,\mathbf{h} : P\in\mathfrak{P}\,\&\,\mathbf{h}\ \textit{satisfies Assumption 1}\,|\,(P,\mathbf{h})\mapsto Q\ \textit{as per Assumption 3}\} \tag{4.2}$$

[3] We further emphasise that eq. (4.1) embeds Assumption 1.

as the collection of structures with which the model is compatible and that induce a given observable distribution, where $\mathfrak{P}$ is a placeholder for any family that is compatible with Assumption 2. We let

$$\begin{aligned} \mathfrak{S}(\mathbf{a}) &\equiv \{Contour(y,t,\mathbf{a}) : y,t \in \mathfrak{F}_y \times \mathfrak{F}_t\} \\ Core(\mathfrak{P},Q) &\equiv \{P,\mathbf{h} : P(\mathfrak{U}|\mathbf{a}) \leq Capacity(Q,\mathfrak{U},\mathbf{a}) \, \forall \, \mathfrak{U},\mathbf{a} \in \mathfrak{S}(\mathbf{a}), \mathfrak{F}_{\mathbf{a}} \,|\, P \in \mathfrak{P}\} \end{aligned} \tag{4.3}$$

for convenience, and refer to these objects as the core-determining sets and the core, respectively. Then

$$IR(\mathfrak{P},Q) = Core(\mathfrak{P},Q) \tag{4.4}$$

which is a sharp characterisation.

If the right-hand side of eq. (4.4) is finite for every observable distribution then the model has complete identification power (and is has unique identification power if the right-hand side of eq. (4.4) consists of just one structure for every observable distribution); and if the right-hand side of eq. (4.4) is empty for at least one observable distribution then the model is falsifiable. We obtain the following theorem, which enumerates the boundary of the identification region and derives conditions under which the identification region is empty.

THEOREM 2. *We maintain the discrete instrumental variable model that comprises Assumptions 1 to 3. We let*

$$\begin{aligned} \overline{q}_{yt}(x) &\equiv \max_{z \in \mathfrak{F}_{z|x}} (Q(y,t|z,x)) \\ \underline{q}_{yt}(x) &\equiv \min_{z \in \mathfrak{F}_{z|x}} (Q(y,t|z,x)) \end{aligned} \tag{4.5}$$

and

$$\begin{aligned} \overline{q}_{y_1+y_0}(x) &\equiv \max_{z \in \mathfrak{F}_{z|x}} (Q(y_1,1|z,x) + Q(y_0,0|z,x)) \\ \underline{q}_{y_1+y_0}(x) &\equiv \min_{z \in \mathfrak{F}_{z|x}} (Q(y_1,1|z,x) + Q(y_0,0|z,x)) \end{aligned} \tag{4.6}$$

for convenience. We define

$$\Theta_{Manski}(Q) \equiv \{P,\mathbf{h} : P(h_y(t,x,u) = 1-y|x) \leq 1 - \overline{q}_{yt}(x) \, \forall \, y,t,x \in \mathfrak{F}_y \times \mathfrak{F}_t \times \mathfrak{F}_x\} \tag{4.7}$$

and

$$\Theta_{Pearl}(Q) \equiv \{P,\mathbf{h} : P(h_y(1,x,u) = y_1, h_y(0,x,u) = y_0|x) \leq \underline{q}_{y_1+y_0}(x) \, \forall \, y_1,y_0,x \in \mathfrak{F}_y^2 \times \mathfrak{F}_x\} \tag{4.8}$$

as those structures that are compatible with the Manski bounds (Manski, 1990) and the Pearl bounds (Balke & Pearl, 1997), respectively. If a given observable distribution is such that $\Theta_{Manski}(Q) \cap \Theta_{Pearl}(Q) = \varnothing$ then $IR(\mathfrak{P}_{RA},Q) = \varnothing$. The discrete instrumental variable model yields

$$\max_{x \in \mathfrak{F}_x} \circ \max_{t \in \mathfrak{F}_t} (\overline{q}_{1t}(x) + \overline{q}_{0t}(x)) \leq 1 \tag{4.9}$$

as a testable implication—for violations that lead to $\Theta_{Manski}(Q) = \varnothing$*; and, additionally, yields*

$$
\begin{aligned}
&\min_{x\in\mathfrak{F}_x}(\underline{q}_{1+1}(x) + \underline{q}_{1+0}(x) - \overline{q}_{11}(x)) \geq 0\\
&\min_{x\in\mathfrak{F}_x}(\underline{q}_{1+1}(x) + \underline{q}_{0+1}(x) - \overline{q}_{10}(x)) \geq 0\\
&\min_{x\in\mathfrak{F}_x}(\underline{q}_{0+0}(x) + \underline{q}_{0+1}(x) - \overline{q}_{01}(x)) \geq 0\\
&\min_{x\in\mathfrak{F}_x}(\underline{q}_{0+0}(x) + \underline{q}_{1+0}(x) - \overline{q}_{00}(x)) \geq 0
\end{aligned}
\tag{4.10}
$$

and

$$
\min_{x\in\mathfrak{F}_x}(\overline{q}_{1+1}(x) + \overline{q}_{1+0}(x) + \overline{q}_{0+1}(x) + \overline{q}_{0+0}(x)) \geq 1 \tag{4.11}
$$

as testable implications—for violations that lead to $\Theta_{Manski}(Q) \cap \Theta_{Pearl}(Q) = \varnothing$*. These testable implications are necessary and sufficient to detect any observable distributions that are incompatible with the discrete instrumental variable model.*

The model has incomplete identification power—the model is partially identifying, to use more modern parlance—and so stands in contrast to other models in which continuous variation (Lewbel, 2000) or parametric restrictions (Heckman, 1978) grant not just complete but unique (or point) identification power. Continuous variation of covariates or instruments cannot, however, be introduced in this framework without technical modification or without further restriction on the structural equations,[4] due to the effect that continuous variation of attributes has on the number of counterfactuals and measurability; parametric restrictions can be introduced to the structural equations, by replacing the general expressions of probability in eqs. (4.7) and (4.8) with those implied by the parametric restrictions.

One further property of that model is that it is complete—every combination of the exogenous economic variates induces a single level of the endogenous economic variates. This property is due to the presence of a treatment equation, which is often omitted elsewhere. A consequence of this property is that the inequality relations in eqs. (4.7) and (4.8) hold with equality (Chesher & Rosen, 2017, §Corollary 2).

We defer proof of Theorems 1 and 2 to Appendix A.

5. Information content

Although random assignment is a very natural assumption to make given its straightforward economic interpretation, an important question to ask is whether it is necessary. To answer this question, we consider two other common statistical independence conditions. For convenience, we define these alternatives in terms of the so-called potential outcome notation (Holland, 1986, Neyman, 1990) that associates heterogeneity with the levels of response and treatment that are effected for hypothetical levels of attributes. We denote the potential outcomes by $y_u(t, x) \in \mathfrak{F}_y$

[4]Continuous differentiability of the structural equations, for instance, would be more than sufficient to restrict the number of counterfactuals to $\aleph_1$ (the cardinality of the continuum), which is necessary for heterogeneity to be measurable. The maximum and minimum operators in Theorem 2 would be replaced by the more general supremum and infimum operators in this setting.

and $t_u(\mathbf{a}) \in \mathfrak{F}_t$ and refer to these as the response counterfactuals and treatment counterfactuals, respectively, and as the counterfactuals, collectively. That is, we define

$$\begin{aligned} y_u(1,x) &\equiv h_y(1,x,u) \\ y_u(0,x) &\equiv h_y(0,x,u) \end{aligned} \tag{5.1}$$

and

$$t_u(\mathbf{a}) \equiv h_t(\mathbf{a},u) \tag{5.2}$$

for which me make explicit the dependence of the counterfactuals upon heterogeneity. We present the following definitions.

DEFINITION 1 (Joint statistical independence). We let

$$\mathfrak{P}_{\text{JSI}} \equiv \{P : y_u(1,x), y_u(0,x) \perp\!\!\!\perp z|x\} \tag{5.3}$$

for convenience. We say that P satisfies joint statistical independence if $P \in \mathfrak{P}_{\text{JSI}}$.

DEFINITION 2 (Marginal statistical independence).

$$\mathfrak{P}_{\text{MSI}} \equiv \{P : y_u(1,x) \perp\!\!\!\perp z|x \,\&\, y_u(0,x) \perp\!\!\!\perp z|x\} \tag{5.4}$$

We say that P satisfies marginal statistical independence if $P \in \mathfrak{P}_{\text{MSI}}$.

Every probability distribution that satisfies random assignment also satisfies joint statistical independence; and every probability distribution that satisfies joint statistical independence also satisfies marginal statistical independence.[5] An immediate and obvious consequence is that

$$\text{IR}(\mathfrak{P}_{\text{RA}}, Q) \subseteq \text{IR}(\mathfrak{P}_{\text{JSI}}, Q) \subseteq \text{IR}(\mathfrak{P}_{\text{MSI}}, Q) \tag{5.5}$$

but what is less clear is whether this ordering is strict (i.e., the subset relation is without equality). We obtain the following corollary, which clarifies the nature of this ordering.

COROLLARY 1 (to Theorem 2). *If a given observable distribution is such that* $\Theta_{Manski}(Q) = \varnothing$ *then* $IR(\mathfrak{P}_{MSI}, Q) = \varnothing$*; if a given observable distribution is such that* $\Theta_{Manski}(Q) \cap \Theta_{Pearl}(Q) = \varnothing$ *then* $IR(\mathfrak{P}_{JSI}, Q) = \varnothing$ *and* $IR(\mathfrak{P}_{RA}, Q) = \varnothing$.

An interesting property of the model is that random assignment does not transmit any additional information content beyond joint statistical independence—illustrated by the coincidence of the identified region under these two statistical independence conditions as per Corollary 1—and is also an insight of Kédagni & Mourifié (2020).[6] This similarity in conclusion is not unexpected since the model is a straightforward extension of the one that is studied in Kédagni

[5] We include joint statistical independence and marginal statistical independence for completeness since they are arguably of limited use in our setting for our chosen application. Joint statistical independence is of practical relevance in models that do not feature a treatment equation and for which potential treatments are not well-defined; marginal statistical independence is of theoretical relevance only and is difficult to imagine—let alone rationalise—in practice.

[6] And also of Richardson & Robins (2014), albeit in the context of a slightly different framework that does not include covariates.

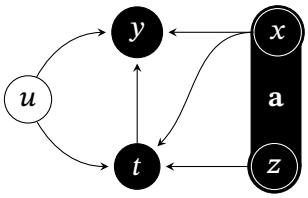


FIG 2. A graphical representation of the model under strong random assignment.

& Mourifié (2020).[7] Something that is also of interest though is whether this apparent lack of additional information content survives the imposition of a further statistical independence condition. We introduce the following assumption on the population to discuss this point.

ASSUMPTION 4 (Strong random assignment). We let

$$\mathfrak{P}_{\text{S–RA}} \equiv \{P : u \perp\!\!\!\perp \mathbf{a}\} \tag{5.6}$$

for convenience. The model restricts P to belong to the family of probability distributions satisfying $\mathfrak{P}_{\text{S–RA}}$.

Chesher & Rosen (2021) points out that strong random assignment can be obtained by augmenting random assignment with the statistical independence condition that is $u \perp\!\!\!\perp x$. The implication is that strong random assignment transmits at least as much information content as random assignment when imposed in conjunction with Assumptions 1 and 3. We obtain the following theorem, which enumerates the boundary of the identification region and derives conditions under which the identification region is empty.

THEOREM 3. *We maintain the discrete instrumental variable model that comprises Assumptions 1, 3 and 4, and we maintain the definitions of Theorem 2. We define*

$$\Theta_M(Q) \equiv \{P, \mathbf{h} : P(h_y(t,x,u) = 1-y) \leq 1 - \overline{q}_{yt}(x) \, \forall \, y,t,x \in \mathfrak{F}_y \times \mathfrak{F}_t \times \mathfrak{F}_x\} \tag{5.7}$$

and

$$\Theta_P(Q) \equiv \{P, \mathbf{h} : P(h_y(1,x,u) = y_1, h_y(0,x,u) = y_0) \leq \underline{q}_{y_1+y_0}(x) \, \forall \, y_1, y_0, x \in \mathfrak{F}_y^2 \times \mathfrak{F}_x\} \tag{5.8}$$

as the unconditional analogues of the Manski bounds and Pearl bounds of Theorem 2. If a given observable distribution is such that $\Theta_M(Q) \cap \Theta_P(Q) = \varnothing$ then $IR(\mathfrak{P}_{S-RA}, Q) = \varnothing$. The model yields eqs. (4.9) to (4.11) as testable implications—for violations that lead to $\Theta_M(Q) \cap \Theta_P(Q) = \varnothing$. These testable implications are necessary and sufficient to detect any observable distributions that are incompatible with the discrete instrumental variable model.

[7] In fact, Kédagni & Mourifié (2020, §Supplementary material 5.) considers a similar extension; the model that is studied therein is written in potential outcome form but does not make explicit the relationship between covariates and other observable variates as we do. Despite this omission, consideration of the maintained statistical independence conditions à la Pearl (2009) yields that covariates *can* cause response and treatment (as we suppose) but not the reverse.

Although the identification regions under random assignment and strong random assignment are similar, we emphasise that they are technically distinct—the former identifying the conditional distribution, the latter identifying the unconditional distribution. To illustrate this difference we extract

$$P(y_u(t,x) = 1 - y | x_n) \leq 1 - \overline{q}_{yt}(x) \tag{5.9}$$

and

$$P(y_u(1,x) = y_1, y_u(0,x) = y_0 | x_n) \leq \underline{q}_{-y_1+y_0}(x) \tag{5.10}$$

from the statements of the Manski bounds and the Pearl bounds, respectively; whereas random assignment requires that eqs. (5.9) and (5.10) hold for $x_n = x$, strong random assignment requires that eqs. (5.9) and (5.10) hold for all $x_n \in \mathfrak{F}_x$. Despite this difference, the testable implications that are derived under random assignment and strong random assignment are identical. This coincidence is an artefact of how eqs. (5.9) and (5.10) constrain the population. In particular, eqs. (5.9) and (5.10) combine over all possible levels of covariates to form a convex region; provided that this region is non-empty, then it contains the product distribution (i.e., the joint distribution that is obtained by multiplying eqs. (5.9) and (5.10) over all levels of covariates).[8] Accordingly, we conclude that any population that is admitted under random assignment but that is not admitted under strong random assignment differs on test sets that have no bearing on what is observed—every structure that is admitted under random assignment is observationally equivalent to a structure that is admitted under strong random assignment—and it is for this reason that the testable implications are the same.

From a practical perspective, it is likely of no consequence to the policymaker whether the model identifies the conditional distribution or the unconditional distribution. For one thing, if covariates are not manipulable (either directly, or indirectly due to manipulation of another aspect of the economic environment) then the policymaker cannot influence this margin; the policymaker must take as given whatever conditional distribution presents, whether for the purpose of implementing some policy or simply to learn about the effect of one economic variate upon another. We focus upon the conditional criteria that are

$$\begin{aligned} \overline{y}_1(x) &\equiv P(y(1,x) = 1 | x) \\ \overline{y}_0(x) &\equiv P(y(0,x) = 1 | x) \end{aligned} \tag{5.11}$$

and

$$\delta_{\overline{y}}(x) \equiv \overline{y}_1(x) - \overline{y}_0(x) \tag{5.12}$$

which we term the conditional average structural functions and the conditional average treatment effect, respectively, in recognition of their relevance and similarity—or dissimilarity—to

[8] The joint distribution that we propose in proving Theorem 2 is a product distribution. The inclusion of the product distribution in the identification region is an artefact of how each lower bound is obtained from the respective upper bound on the complementary counterfactual; provided that the population satisfies the appropriate bounds for each level of covariates then there exists a population that is proper (i.e., sums to one) irrespectively of what values each conditional probability actually takes (i.e., their values are unrelated).

the average structural functions (Blundell & Powell, 2003) and in line with Abrevaya, Hsu & Lieli (2015); and upon the aggregated criteria that are

$$\begin{aligned}\beta_1(\mu) &\equiv \textstyle\sum_{x\in\mathfrak{X}_x} \mu(x)\cdot \overline{y}_1(x)\\ \beta_0(\mu) &\equiv \textstyle\sum_{x\in\mathfrak{X}_x} \mu(x)\cdot \overline{y}_0(x)\end{aligned} \tag{5.13}$$

and

$$\delta_\beta(\mu) \equiv \beta_1(\mu) - \beta_0(\mu) \tag{5.14}$$

which we term the aggregated average structural functions and the aggregated average treatment effect, respectively. Here, μ is a probability measure that either coincides with the population or, else, is hypothetical.

We assert that the conditional criteria—eqs. (5.11) and (5.12)—and the aggregated criteria—eqs. (5.13) and (5.14)—are useful to the policymaker. The conditional average structural functions report the average response that would materialise if the policymaker were able to implement a particular level of treatment in the sub-population (real or hypothetical) distinguished by a particular level of covariates. The conditional average structural functions are, therefore, useful in instances where the effect of treatment varies across different sub-populations, and where discriminatory policy can be of benefit—which is true whenever the conditional average treatment effect is non-zero. The aggregated average structural functions report the average response that would materialise if the policymaker were able to uniformly implement a particular level of treatment in the population (real or hypothetical). The aggregated average structural functions are, therefore, useful in instances where the effect of treatment does not vary across different sub-populations, or where discriminatory policy is not possible (due to financial, legal, moral, or political constraints)—with the aggregated average treatment effect suggesting the best policy to adopt.

The model has incomplete identification power and this extends to the conditional average structural functions, which are partially identified. Specifically, the model identifies the conditional average structural functions as

$$\begin{aligned}\max(\overline{q}_{11}(x), \overline{q}_{1+0}(x) - \underline{q}_{0+0}(x)) \leq \overline{y}_1(x) \leq \min(1-\overline{q}_{01}(x), \underline{q}_{1+1}(x) + \underline{q}_{1+0}(x))\\ \max(\overline{q}_{10}(x), \overline{q}_{0+1}(x) - \underline{q}_{0+0}(x)) \leq \overline{y}_0(x) \leq \min(1-\overline{q}_{00}(x), \underline{q}_{1+1}(x) + \underline{q}_{0+1}(x))\end{aligned} \tag{5.15}$$

which are then sufficient to identify the other criteria via eqs. (5.12) to (5.14) and Minkowski arithmetic.[9] This *cost* is a trade-off for the additional credibility that the model has versus more conventional approaches to instrumental variable problems, and for the opportunity to identify more *general* effects in the population than, say, the local average treatment effect (Imbens & Angrist, 1994). Comparison with the local average treatment effect is useful here for another reason—for elucidating the motivation for including covariates, something that is not obvious if interest is in the aggregated criteria and strong random assignment is maintained. Like the

[9]Named for Hermann Minkowski (1864–1909), Minkowski arithmetic involves convex operations on sets—say, A and B. It includes addition (and subtraction)—if $a \in A$ and $b \in B$ then $A + B = \{a + b : a \in A, b \in B\}$—and multiplication by a scalar—if $a \in A$ and $\lambda > 0$ then $\lambda \cdot A = \{\lambda \cdot a : a \in A\}$. Such arithmetic is necessary here since eq. (5.13) involves the addition of and multiplication by scalars of sets since the conditional average structural functions constitute sets, as per eq. (5.15).

policymaker can choose between the Wald estimator and the two-stage least squares estimator to recover the local average treatment effect, so the policymaker can conceive of instead using an auxiliary model here that omits covariates. It is well-known that the Wald estimator and the two-stage least squares estimator diverge when covariates influence response and covariates and instruments are correlated; a similar thing happens if an auxiliary model is used here—this divergence manifests in eq. (5.13) as variation in the conditional average structural functions across covariates and replacement of μ by a probability measure that is conditional on instruments. If the policymaker believes that covariates influence response and covariates and instruments are correlated then an auxiliary model should not be used.

We note that neither Theorem 2 nor Theorem 3, in defining the identified set of structures, depend upon the form of the treatment equation. There is, in a certain sense, limited information transmission between the structural equations given the non-parametric form of the response equation and what this implies for heterogeneity. As such, the treatment equation only conveys how the policymaker thinks about the economic environment—that attributes cause treatment, for instance, but response does not.

We defer proof of Theorem 3 to Appendix A.

6. Inference

In what follows, we suppose that the model constitutes Assumptions 1 to 3—we disregard Assumption 4. Moreover, we fix the structural equations—and so impose a particular labelling of heterogeneity—as a normalisation, which is valid in view of eqs. (3.2) and (3.3) and the wider discussion of Section 3. The identification and inferential problems are then solely focussed upon identification of and inference about the population.

We adopt a Bayesian posture towards the inferential problem. We follow the robust approach of Giacomini & Kitagawa (2021), specifying a prior over the reduced-form parameters rather than over the structural parameters (terms that we define precisely in due course). As is explained in Giacomini & Kitagawa (2021), the principal advantage of this approach—and what makes it robust—is that it is not sensitive to the choice of prior in the same way that a conventional Bayesian approach is when there is only partial updating (see Kitagawa, 2012, for a detailed illustration of this point in the context of a missing data example that is not dissimilar to the problem at hand).

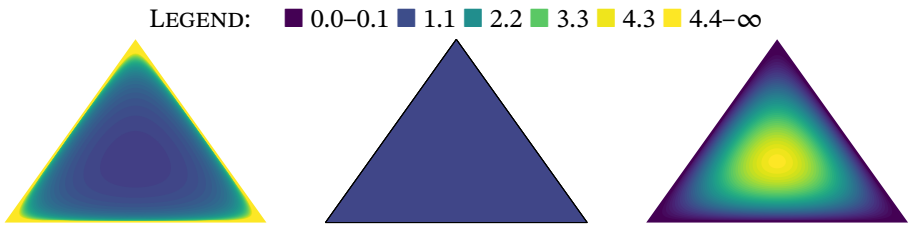
We denote the reduced-form parameters by η (we do not distinguish between the collection of parameters and an individual parameter for convenience) and define these in terms of a probability distribution over response, treatment and attributes as

$$\eta \equiv \Pr(y, t|\mathbf{a}) \tag{6.1}$$

We denote the structural parameters by θ (we do not distinguish between the collection of parameters and an individual parameter for convenience) and define these in terms of a probability distribution over heterogeneity and covariates as

$$\theta \equiv \Pr(u|x) \tag{6.2}$$

Both eqs. (6.1) and (6.2) are expressed in terms of an as yet unspecified probability measure that is the manifestation of some held belief. It is immediate given our understanding of the relationship between the observable distribution—that eq. (6.1) resembles—and the population—that



(a) $\alpha = 1/2$ (b) $\alpha = 1$ (c) $\alpha = 2$

Density contours are plotted on the unit simplex for symmetric Dirichlet distributions over simplexes in two dimensions (i.e., feasible combinations of three probabilities).

FIG 3. Some symmetric Dirichlet distributions and their densities.

eq. (6.2) resembles—that any held belief about the reduced-form parameters is fully updated by data and that any held belief about the structural parameters is partially updated by data. Moreover, any held belief about the reduced-form parameters defines a class of beliefs about the structural parameters.

Leveraging Bayes' theorem, we write the relationship between the likelihood (of data), the posterior and the prior as

$$[\text{Posterior belief} : P = \Pr\,|\text{Data}] \propto \textstyle\sum_{\Pr}[\text{Likelihood} : \text{Data}\,|\,\Pr] \cdot [\text{Prior belief} : P = \Pr] \tag{6.3}$$

where the summation on the right-hand side is over all probability measures which the prior ascribes positive probability to. Fixing one such probability measure, the associated likelihood is the product of a specified marginal distribution over attributes and each reduced-form parameter, both raised by the number of times that the corresponding combination of response, treatment and attributes occurs (i.e., eq. (6.1); the product yields a joint distribution over response, treatment and attributes). For this prior to reflect the model, it must ascribe probability one to observable distributions that satisfy the testable implications of the model (i.e., observable distributions that satisfy Theorem 2), although we do not impose this condition at this stage.

A natural choice of prior in view of the fact that each economic variate has finite support (and so is distributed according to a multinomial distribution) is the Dirichlet distribution.[10] The Dirichlet distribution is characterised by a finite number of (positive) concentration parameters; each concentration parameter is associated with a particular combination of response, treatment and attributes, and captures the relative chance of that combination occurring (a larger relative value corresponding to a higher relative chance). An important special case is when the concentration parameters are symmetric and all equal to one half, in which case the Dirichlet distribution coincides with the Jeffreys prior. As is evident from fig. 3, the Jeffreys prior (the left-hand panel) favours more extreme probability distributions than either the uniform prior (the centre-panel) or a more concentrated prior (the right-hand panel). The Dirichlet distribution is a conjugate prior and, in the special case of the Jeffreys prior (and more generally for other symmetric Dirichlet distributions), yields a posterior with an especially simple form; if data reveals a particular combination of response, treatment and attributes a total of n times then the concentration

[10] A probability distribution over one or several economic variates with finite support can be represented by a point on the unit simplex (of one dimension less than the number of variates). The Dirichlet distribution—as a probability distribution over such a probability distribution—is, therefore, a probability distribution over points on the unit simplex.

parameter corresponding to that combination is equal to $n + 1/2$. We assume the Jeffreys prior over response, treatment and attributes in what follows.

We reiterate that our chosen prior ascribes positive probability to every point on the simplex. This includes points that lead to violations of the testable implications of the model. An attractive feature of the robust approach of Giacomini & Kitagawa (2021) is that it can deal with such an improper prior. To conduct statistical inference, we implement the following algorithm—in which we term what was an unspecified probability measure in eqs. (6.1) and (6.2) a pseudo-distribution.

ALGORITHM 1. INPUT—$\mathfrak{Y}_y$ (the support of response), $\mathfrak{Y}_t$ (the support of treatment), $\mathfrak{Y}_{\mathbf{a}}$ (the support of attributes), and $\mathbf{n}$ (the number of times each combination of outcomes occurs in data).
SPECIFY—$\bar{r} \in \mathbb{N}_+$.
SET—$R = 0$ and $r = 0$.

1. Increment R.
2. Draw a pseudo-distribution from the Dirichlet distribution with parameters equal to $\mathbf{n} + 1/2$ (the posterior).
3. Compute the reduced-form parameters from the pseudo-distribution.
4. Compute the identification region from the reduced-form parameters.
5. Do the reduced-form parameters satisfy the testable implications of the model (i.e., is the identification region empty)?
   - ✓ Retain the pseudo-distribution and increment r.
   - ✗ Discard the pseudo-distribution.
6. Is $r < \bar{r}$?
   - ✓ REPEAT.
   - ✗ END.

OUTPUT—R (the number of pseudo-distributions generated), the reduced-form parameters corresponding to each retained pseudo-distribution, and the identification region corresponding to each retained pseudo-distribution.

We report the α-proportion robust credible region for a criterion of interest—amongst those defined in Section 5—by projecting the identification region for each retained pseudo-distribution that we output from Algorithm 1; we find the smallest possible set that contains α-proportion of these projections. Like Giacomini & Kitagawa (2021) we take *smallest* to mean the set with the lowest Lebesgue volume out of all possible sets that satisfy the coverage constraint.[11]

[11]The criteria that we consider are convex projections and are identified up to closed intervals on the reals. If the intersection of the projections demarcates a non-empty area then we can compute the smallest possible set that contains α-proportion of these projections by performing a series of sort-subset-eliminate operations. We order the projections according to their lower bound; we set the lower boundary of the credible region to the least lower bound; we order the projections according to their upper bound; we set the upper boundary of the credible region to the α-proportion upper bound (i.e., α-proportion of projections have smaller upper bounds); we eliminate the projection that has the smallest lower bound and repeat with the remaining projections. Each time we eliminate a projection, we obtain a credible region with at least α-proportion coverage; by comparing these credible regions, we obtain the smallest credible region with exactly α-proportion coverage. This procedure involves relatively few operations and is guaranteed to return the set with the lowest Lebesgue volume out of all possible sets that satisfy the coverage constraint.

The interpretation of the α-proportion robust credible region is that the true interval for the criterion under consideration is contained within the robust credible region with probability equal to α. By exploiting that any held belief about the reduced-form parameters defines a class of beliefs about the structural parameters, we are able to extend this interpretation to conventional posteriors; the probability that the true interval for the criterion under consideration is contained within the robust credible region is equal to α according to any conventional posterior in the aforementioned class.

REMARK 2. Giacomini & Kitagawa (2021) establishes a set of conditions under which the robust credible region asymptotically attains correct frequentist coverage. One of these conditions is that the endpoints of the interval that a criterion is identified up to are continuously differentiable about the truth and have non-zero derivatives. Since every criterion that we consider can be obtained by arithmetic on eqs. (5.9) and (5.10), these endpoints are not differentiable (Kitagawa et al., 2020) and so the robust credible regions that we report do not have a frequentist interpretation. Incidentally, non-differentiability is also the cause of the failure of the frequentist bootstrap (Andrews & Han, 2009); we assert that our inability to interpret the robust credible region in a frequentist way can also be framed as being due to this, as the bootstrap distribution (of the relative frequency of combinations of response, treatment and attributes) resembles the Dirichlet distribution as the quantity of data becomes large (and the prior exerts a vanishingly small influence upon the posterior).

We also report the posterior plausibility (Giacomini & Kitagawa, 2021) of the model. To do so, we compute $\bar{r}/R$ using our output from Algorithm 1 to derive a simple measure of how credible Assumptions 1 to 3 are (with a higher value suggesting that they are more credible, or—loosely speaking—less likely to be proven false by data).[12]

In what follows, we use estimate to mean the posterior mean, which we obtain by projecting the identification region for each retained pseudo-distribution that we output from Algorithm 1 and averaging.

7. Application

We consider the question of whether and by how much additional children affect maternal employment at the extensive margin. We use data from the Integrated Public Use Microdata Series (IPUMS: Minneapolis, MN, 2024); the data consist of individual-level observations of U.S. households, and were collected as part of the decennial census and the American Community Survey (ACS). We augment this data with demographic statistics and projections (U.S. Census Bureau, 2020) that we use to inform our overall design. To apply our model to this context, we build upon the empirical strategy of Angrist & Evans (1998). We associate response with a binary indicator for whether a mother worked in the previous year; we associate treatment with with a binary indicator for whether a mother has more than two children; we associate instruments with discrete indicators for family composition (i.e., the sex of a mother's first two children) and the occurrence of a multiple second birth (i.e., twins), and combinations thereof; and we associate covariates with discrete indicators for age, and race and ethnicity—race, for brevity—and combinations thereof.

[12] We set r equal to 10,000 in practice.

We separate our consideration of the question of whether and by how much additional children affect maternal employment at the extensive margin into several parts. The model embeds statistical restrictions that translate to arguably weak constraints on economic behaviour; we discuss the meaning of these restrictions in the context of the question. We use this discussion to motivate the remainder of our analysis. This motivation centres on two things in particular. First, is there broad evidence to support an investigation of how much additional children affect maternal employment at the extensive margin? To address this, we provide summary statistics that show evidence of correlation between the number of children that a mother has and her state of employment. Second, is there broad evidence to support our approach? To address this, we provide summary statistics that show evidence of correlation between family composition and the number of children that a mother has. Upon establishing these two things, we proceed to implement the model and its associated methods and results, reporting the posterior plausibility of the model and robust credible regions of criteria of interest.

7.1. The meaning and credibility of the model

The model embeds statistical restrictions that translate to arguably weak constraints on economic behaviour. The model allows for a vast array of unspecified economic variates to influence the employment decision and the fertility decision; these generate dependence between the employment decision and the fertility decision; and these generate rich variety in the effect of fertility upon employment. We discuss these positive aspects of the model in greater detail now.

First and principally, the model allows for a vast array of unspecified economic variates to influence the employment decision and the fertility decision. These economic variates can capture differences in the endowments that households possess, the preferences that they hold, and the prices that they face; for example, if mothers face low childcare costs due to the presence of nearby family. The ability of the model to incorporate latent elements of the economic environment is discussed in Section 3 and is a consequence of its flexibility—the model imposes little constraint on what these unspecified economic variates are, how they are distributed, and how they influence the employment decision and the fertility decision. This flexibility is in contrast to parametric models that typically impose non-trivial constraints upon the distributions of endowments, of preferences, and of prices, and usually involve aggregation of these elements.

Second, these unspecified economic variates generate dependence between the employment decision and the fertility decision; for example, the model allows for mothers that are presented with poor employment opportunities (and a low opportunity cost of pregnancy) to choose to try for an additional child. This joint determination is common also to conventional parametric models and is generally viewed as a sensible—if not necessary—property to capture (Benny, 2021, Roy, 1951).

Third, these unspecified economic variates generate rich variety in the effect of fertility upon employment; for example, the model allows for mothers (in the aggregate) to move in to or to move out of employment following the birth of an additional child. This behavioural diversity is in contrast to conventional parametric models that typically impose monotonicity upon the effect. Indeed, the widely used single index model is known to exclude one of the above behaviours—(in the aggregate) mothers can either move in to or out of employment following the birth of an additional child, but not both—by virtue of it imposing weak separability (Vytlacil, 2002). Of course, this is not to say that parametric models—and the single index model in particular—cannot generate variety in the effect of fertility upon employment—they can—but this is limited to variety across observable groups and not within observable groups, which is

what we intend by *rich*. Rich variety is an important property of the model that is studied in Gronau (1977); there, differences in the endowments that families (since Gronau's is a model of a collective decision-making process) possess and the prices that they face generate different employment responses at the extensive margin. Formally, this manifests via the influence that children have on the cost of employment—say, the need to provide childcare—and the value of employment—children increase the marginal utility of income—both of which depend upon endowments and prices; depending upon how the initial and subsequent change in the cost and value of employment compare, the birth of an additional child can precipitate differential employment reactions.

The model does, however, also encode several statistical independence conditions. Together, these conditions constitute a classical instrumental variable restriction. At least some of these statistical independence conditions are unverifiable, and must be defended. Of course, any study that is to provide meaningful insight into the mechanics of an economic process must impose at least some framework on the problem at hand; credibility then hinges on how constricting and, ultimately, believable this framework is. Relative to alternatives, the statistical independence conditions that the model encodes are weak. We discuss these statistical independence conditions in greater detail now.

First, the statistical independence condition that is implied by Assumption 2 translates to the requirement that employment opportunities are not influenced by family composition once age and race are controlled for. Violation is unlikely since sex is determined at random according to a probability distribution that skews towards male children overall but that is more favourable towards female children as age increases. In the case of the occurrence of a multiple second birth, similar reasoning applies. The argument in both scenarios is that any correlation between instruments and employment opportunities is spurious. We emphasise that this is an unverifiable component.

Second, the exclusion restriction that is implied by Assumption 1 translates to the requirement that family composition does not directly influence the employment decision. Violation occurs if children of different sex incur additional costs *vis à vis* children of the same sex; for example, if female and male children participate in different activities at distinct venues, or if they attend different schools. In the case of the occurrence of a multiple second birth, similar reasoning applies; for example, if having children of the same age requires duplication, versus handing-down or re-using (of clothes and such like). The argument in both scenarios is that instruments influence the employment decision via their effect upon the budget, and not simply through their influence on the decision of whether to have additional children.

Third, the relevance condition that is implied by Assumption 1 translates to the requirement that family composition influences the fertility decision. Violation occurs if parents either do not express a preference for variety or if parents are equally balanced over all possible combinations of sex (i.e., as many parents prefer mixtures of female and male children as prefer just female or just male). In the case of the occurrence of a multiple second birth, violation is impossible since a mother necessarily has more than two children.

In what follows, we introduce evidence in support of the second positive aspect of the model—we show that there is broad evidence to support an investigation of how much additional children affect maternal employment at the extensive margin—and evidence in support of the relevance condition—we show that there is broad evidence to support our approach. This evidence is presented for a specific sample of women.

TABLE 1
Descriptions of sub-samples used for data analysis.

Sub-sample	Description	*n*
S1	Women aged 20–34 years of age.	4,302,947
S2	Women aged 35–49 years of age.	4,088,279
S3	Mothers aged 20–34 years of age with at least one child.[1.]	1,893,763
S4	Married mothers aged 20–34 years of age with at least one child, and whose husband is present.[1.,2.]	1,053,890
S5	Mothers aged 20–34 years of age with at least two children.[1.,3.]	994,040
S6	Married mothers aged 20–34 years of age with at least two children, and whose husband is present.[2.,3.,4.]	580,091

Notes: 2.6% of households are excluded because the age or sex of at least one member of the household is imputed.
[1.] Eldest child aged 17 years or younger and born after mother's 16th birthday.
[2.] Mother and husband both married and married only once.
[3.] Second-oldest child aged one year or older.
[4.] Eldest child aged 17 years or younger and born after mother's 16th birthday and that of her husband.

7.2. Sample and sub-sample design

We use data from 1980—collected for the decennial census—and from 2008 through 2018—collected for the ACS.[13] We augment this data with demographic statistics and projections (U.S. Census Bureau, 2020) that we use to inform our overall design. We largely follow Angrist & Evans (1998); the rationale for focusing on particular age-groups of mothers in conjunction with particular age-groups of children is as relevant there as it is here. We do deviate from Angrist & Evans (1998) in three ways though; in how we handle imputation, as a deliberate choice; in how we handle marriage, as a deliberate choice; and in the age-groups that we focus upon, a reflection of what we intend to do. We construct several sub-samples from the remaining data, which we summarise in Table 1, and utilise in our description of the environment and in our statistical analysis.

Angrist & Evans's sample design excludes women whose own age or sex or those of her first two children (where this is applicable) have been flagged; we, instead, exclude women whose own age or sex or those of *any* member of her household have been flagged. Our more conservative design excludes 3.5% of all women aged 20–49 years; Angrist & Evans's more liberal design excludes 2.3% of all women aged 20–49 years. The bulk of the additional exclusions come from the 1980 survey year, which contains around four times as many observations and has around double the imputation rate as compared to other survey years. Since accurate definition of our two instruments—family composition and the occurrence of a multiple second birth—is reliant upon the correct recording of age and sex—and not just of a mother and her first two children, as incorrect classification can lead to the wrong child being recognised as such—we err on the side of caution so-to-speak, and choose to impose a more conservative requirement.

[13] We omit data from 1990 and 2000, despite it being available, due to it not containing information about quarter of birth, which—although is imperfect (for reasons discussed in due course)—is crucial to defining a multiple second birth. We omit data from 2001 through 2007, despite it being available, due to it not containing information about marriage, which is crucial to our sample construction. We omit data from 2019 through 2022, despite it being available, due to the irregular economic environment (directly and indirectly) caused by the COVID-19 pandemic.

Angrist & Evans's sample design excludes married mothers whose eldest child was born before marriage; we, instead, do not impose this requirement. In disregarding the legitimacy of the eldest child at the time of their birth, our sample design admits 13.1% or 18.2% more women than Angrist & Evans's sample design depending upon the sub-sample that is considered. The bulk of the additional inclusions come from the 2008 through 2018 survey years, which have at least triple the illegitimacy rate as compared to the remaining survey year.[14] Since neither the decennial census nor the ACS provide a complete employment history, since the absence of marriage does not preclude the presence of a stable relationship, and since a more conservative requirement would impact more upon certain racial groups,[15] we choose to impose a more liberal requirement.

Angrist & Evans's sample design includes only women aged, primarily, 21–35 years and, secondarily, aged 36–50 years; we, instead, include only women aged, primarily, 20–34 years and, secondarily, aged 35–49 years. The demographic statistics and projections (U.S. Census Bureau, 2020) that we use to inform our overall design—our choice of sample and what we intend to do—collects women into age groups of five year intervals (women aged 20–24 years, women aged 25–29 years, etc.). Since the rationale for focusing on particular age-groups of mothers in conjunction with particular age-groups of children is arguably unchanged by a shift of one year, and since comparable statistics are not available for different age-groups, we choose to impose a different requirement.

We note that we do not distinguish between adoptive—whether formal or informal—or biological child–parent relationships due to this information being unavailable.[16]

We also note that the indicator that we use for the occurrence of a multiple second birth cannot distinguish between a genuine multiple second birth and children that were born more than nine months but less than one year apart in data collected for the ACS. Whereas data collected for the decennial census provides a snapshot on a single—and fixed—day, and age and quarter of birth can be combined to construct an accurate entry for year of birth (something that is not directly asked for), the ACS occurs throughout the year. Year of birth is incorrect for all individuals whose birthday falls after the survey date; to illustrate, if the date of survey is February, 2024, then an individual that was born in January, 2024, and an individual that was born in March, 2023, would both report their age as zero and their quarter of birth as one—they would be categorised as twins, even though they were born 10 months apart. By using data from the 1980 survey year (which is immune to this issue), we provide an insight into the misclassication rate; a total of 48,593 multiple second births are correctly recognised as such, alongside 65 false negatives and 64 false positives. Since the misclassification rate is negligible and balances across the sample, we do not view this issue as a major concern.

7.3. Employment and fertility statistics

We introduce evidence in support of the second positive aspect of the model—evidence of dependence between the employment decision and the fertility decision. We present summary statistics relating to employment and fertility in Tables 2 to 4. Each table reports estimates of the mean and

[14] We add that this rate is—perhaps unsurprisingly—increasing over time.

[15] Our sample design includes 40.3% or 56.1% more black mothers, 36.3% or 49.7% more American Indian or Alaska Native mothers, and 10.8% or 14.7% more white mothers than than Angrist & Evans's sample design depending upon the sub-sample that is considered.

[16] Adoptive relationships are typically under-reported in any case, which makes IPUMS: Minneapolis, MN's current redefinition of the step-parent variables less of a concern than it otherwise would be.

TABLE 2
Fertility and employment statistics amongst women in sub-samples S1 and S2.

	Year	n	Children ever born		Number of children		Three or more children		Worked last year	
Sub-sample S1	1980	1,393,616	1.125	(1.251)	1.053	(1.200)	0.121	(0.327)	0.739	(0.439)
	2008	249,056			0.894	(1.172)	0.107	(0.309)	0.814	(0.389)
	2009	254,878			0.880	(1.169)	0.105	(0.307)	0.798	(0.402)
	2010	260,191			0.869	(1.172)	0.104	(0.306)	0.778	(0.416)
	2011	263,569			0.837	(1.167)	0.101	(0.302)	0.763	(0.425)
	2012	263,414			0.818	(1.156)	0.099	(0.298)	0.767	(0.422)
	2013	268,352			0.798	(1.149)	0.095	(0.293)	0.774	(0.418)
	2014	267,616			0.778	(1.139)	0.093	(0.290)	0.780	(0.414)
	2015	268,863			0.755	(1.129)	0.089	(0.285)	0.787	(0.409)
	2016	267,512			0.741	(1.124)	0.088	(0.284)	0.798	(0.401)
	2017	271,822			0.723	(1.114)	0.085	(0.280)	0.804	(0.397)
	2018	274,058			0.696	(1.101)	0.082	(0.275)	0.813	(0.390)
Sub-sample S2	1980	887,785	2.751	(1.874)	1.797	(1.448)	0.284	(0.451)	0.669	(0.470)
	2008	313,612			1.399	(1.241)	0.168	(0.374)	0.799	(0.401)
	2009	309,614			1.421	(1.245)	0.173	(0.378)	0.790	(0.408)
	2010	304,753			1.430	(1.250)	0.175	(0.380)	0.774	(0.419)
	2011	292,657			1.420	(1.261)	0.176	(0.380)	0.761	(0.426)
	2012	289,470			1.437	(1.267)	0.179	(0.383)	0.763	(0.425)
	2013	288,439			1.448	(1.263)	0.181	(0.385)	0.766	(0.423)
	2014	282,062			1.452	(1.268)	0.182	(0.386)	0.768	(0.422)
	2015	278,983			1.456	(1.266)	0.182	(0.386)	0.774	(0.418)
	2016	277,621			1.463	(1.273)	0.185	(0.388)	0.780	(0.414)
	2017	281,580			1.464	(1.274)	0.185	(0.388)	0.785	(0.411)
	2018	281,703			1.465	(1.276)	0.184	(0.388)	0.791	(0.407)

standard deviation of answers to several survey questions for two sub-samples that are interesting to compare.[17] The answers to these survey questions detail the number of children that a woman reports ever having—including those that are, for whatever reason, no longer present in the household—(asked only in the 1980 survey year),[18] the number of children that a woman reports having that are present in the household, whether a woman reports having three or more children that are present in the household, and whether a woman reports having worked at all for profit, pay, or as an unpaid family worker during the previous 12 months.

There is a stark contrast in the answers that respondents give between the 1980 survey year and the 2008 through 2018 survey years. This contrast is apparent in any one of Tables 2 to 4. The number of children that a woman reports having that are present in the household is typically higher in the 1980 survey year; whether a woman reports having three or more children that are present in the household is typically more common in the 1980 survey year, and whether a woman reports having worked is typically less common in the 1980 survey year. All this speaks to a decrease in fertility alongside an increase in employment.

[17]We note that the sample size in each survey year is sufficiently large that no standard error exceeds 0.05 or 4.8% of the estimated mean of answers to any survey question in Tables 2 to 4.

[18]The text of the survey question is the following. *How many babies has she ever had, not counting stillbirths? Do not count her stepchildren or children she has adopted. Count all children born alive, including any who have died (even shortly after birth) or who no longer live with her.*

TABLE 3
Fertility and employment statistics amongst women in sub-samples S3 and S4.

	Year	n	Children ever born		Number of children		Three or more children		Worked last year	
Sub-sample S3	1980	731,333	1.95	(1.025)	1.909	(0.940)	0.213	(0.410)	0.616	(0.486)
	2008	111,246			1.889	(0.961)	0.218	(0.413)	0.744	(0.436)
	2009	112,307			1.886	(0.962)	0.217	(0.412)	0.728	(0.445)
	2010	112,687			1.889	(0.970)	0.219	(0.413)	0.709	(0.454)
	2011	109,455			1.896	(0.982)	0.221	(0.415)	0.699	(0.459)
	2012	107,842			1.887	(0.981)	0.219	(0.414)	0.703	(0.457)
	2013	107,244			1.885	(0.980)	0.217	(0.412)	0.709	(0.454)
	2014	104,486			1.881	(0.980)	0.216	(0.412)	0.713	(0.453)
	2015	102,115			1.879	(0.982)	0.215	(0.411)	0.721	(0.449)
	2016	99,648			1.882	(0.980)	0.217	(0.412)	0.731	(0.443)
	2017	99,003			1.881	(0.976)	0.215	(0.411)	0.739	(0.439)
	2018	96,397			1.879	(0.981)	0.215	(0.411)	0.746	(0.435)
Sub-sample S4	1980	467,752	1.963	(0.970)	1.939	(0.910)	0.216	(0.411)	0.581	(0.493)
	2008	57,721			1.915	(0.930)	0.216	(0.411)	0.703	(0.457)
	2009	57,259			1.923	(0.938)	0.220	(0.415)	0.693	(0.461)
	2010	56,544			1.927	(0.941)	0.222	(0.416)	0.683	(0.465)
	2011	53,176			1.934	(0.954)	0.223	(0.416)	0.673	(0.469)
	2012	52,940			1.924	(0.955)	0.221	(0.415)	0.678	(0.467)
	2013	53,358			1.919	(0.955)	0.215	(0.411)	0.680	(0.467)
	2014	52,057			1.915	(0.950)	0.217	(0.412)	0.680	(0.467)
	2015	51,133			1.920	(0.965)	0.218	(0.413)	0.683	(0.465)
	2016	50,754			1.918	(0.957)	0.219	(0.414)	0.692	(0.462)
	2017	50,962			1.916	(0.954)	0.216	(0.412)	0.699	(0.459)
	2018	50,234			1.906	(0.955)	0.214	(0.410)	0.705	(0.456)

Outside of this headline though, the story is perhaps a little more nuanced. From Tables 2 to 4 and some background calculations, there appears to be a general trend towards women having fewer children overall and having them later when they do (we refer the reader to Appendix C and fig. 8). The proportion of women that report having at least one child that is present in the household (see Table 2) decreases between the 1980 survey year and the 2008 through 2018 survey years, and decreases year-on-year during the 2008 through 2018 survey years. Amongst those women with at least one child, the number of children that a woman reports having that are present in the household (see Table 3) decreases between the 1980 survey year and the 2008 through 2018 survey years, and decreases year-on-year during the 2008 through 2018 survey years. Amongst those women with at least two children, the number of children that a woman reports having that are present in the household (see Table 4) increases between the 1980 survey year and the 2008 through 2018 survey years, and increases year-on-year during the 2008 through 2018 survey years. Despite seeming contradictory, all three results are compatible with a general trend towards women having fewer children overall, and average behaviour in the tail being increasingly driven by those with more children. Separately, the age at first birth amongst women with at least one child increases by approximately two years over all of the survey years, rising from roughly 21 or 22 years to roughly 23 or 24 years depending upon the sub-sample that is considered.

Whilst the proportion of women that report having worked increases between the 1980 survey year and the 2008 through 2018 survey years, this is not true during the 2008 through 2018 survey

TABLE 4
Fertility and employment statistics amongst women in sub-samples S5 and S6.

	Year	n	Children ever born		Number of children		Three or more children		Worked last year	
Sub-sample S5	1980	401,788	2.557	(0.919)	2.531	(0.800)	0.387	(0.487)	0.559	(0.496)
	2008	57,181			2.581	(0.825)	0.421	(0.494)	0.703	(0.457)
	2009	57,757			2.581	(0.825)	0.420	(0.494)	0.684	(0.465)
	2010	58,023			2.589	(0.834)	0.422	(0.494)	0.665	(0.472)
	2011	56,656			2.598	(0.851)	0.425	(0.494)	0.654	(0.476)
	2012	54,928			2.603	(0.849)	0.428	(0.495)	0.658	(0.474)
	2013	54,679			2.596	(0.852)	0.422	(0.494)	0.664	(0.472)
	2014	52,878			2.599	(0.852)	0.424	(0.494)	0.663	(0.473)
	2015	51,408			2.602	(0.858)	0.424	(0.494)	0.671	(0.470)
	2016	50,130			2.604	(0.852)	0.429	(0.495)	0.684	(0.465)
	2017	50,100			2.596	(0.848)	0.423	(0.494)	0.696	(0.460)
	2018	48,512			2.601	(0.855)	0.426	(0.494)	0.699	(0.459)
Sub-sample S6	1980	268,713	2.523	(0.851)	2.500	(0.766)	0.374	(0.484)	0.528	(0.499)
	2008	30,750			2.542	(0.797)	0.402	(0.490)	0.654	(0.476)
	2009	30,758			2.553	(0.800)	0.408	(0.491)	0.641	(0.480)
	2010	30,599			2.554	(0.802)	0.407	(0.491)	0.633	(0.482)
	2011	28,854			2.561	(0.825)	0.408	(0.491)	0.618	(0.486)
	2012	28,225			2.569	(0.824)	0.412	(0.492)	0.622	(0.485)
	2013	28,340			2.561	(0.835)	0.403	(0.490)	0.630	(0.483)
	2014	27,456			2.562	(0.823)	0.408	(0.491)	0.622	(0.485)
	2015	26,907			2.577	(0.844)	0.412	(0.492)	0.623	(0.485)
	2016	26,526			2.576	(0.834)	0.416	(0.493)	0.635	(0.481)
	2017	26,899			2.564	(0.830)	0.407	(0.491)	0.646	(0.478)
	2018	26,064			2.568	(0.834)	0.409	(0.492)	0.646	(0.478)

years; instead, the proportion fluctuates over this period. We assert that this apparent lack of trend need not be due to a deliberate reduction in labour supply over this period, and can instead be attributed to the prevailing macroeconomic conditions (we refer the reader to Appendix C and figs. 9 to 11).

7.4. Fertility and family composition statistics

We introduce evidence in support of the relevance condition—evidence that family composition influences the fertility decision. We present summary statistics relating to fertility and family composition in figs. 4 and 5. Each figure reports estimates of the probability and associated 95% frequentist confidence interval of whether a woman that reports having at least one child (see fig. 4) or at least two children (see fig. 5) has another child. We decompose this probability according to family composition—thereby comparing the probability of a given choice across different permutations of the two sexes—and by year (the 1980 survey year left-most, the 2018 survey year right-most).

There is no difference between the probability that a woman with one child has another if her first child is female versus if her first child is male. In other words, there appears to be no explicit preference over the two sexes. This similarity in probability is apparent in fig. 4. There is a difference between the probability that a woman with two children has another if her first two children have the same sex versus if her first two children have different sexes. In other words, there appears to be preference for variety in the sex of children.

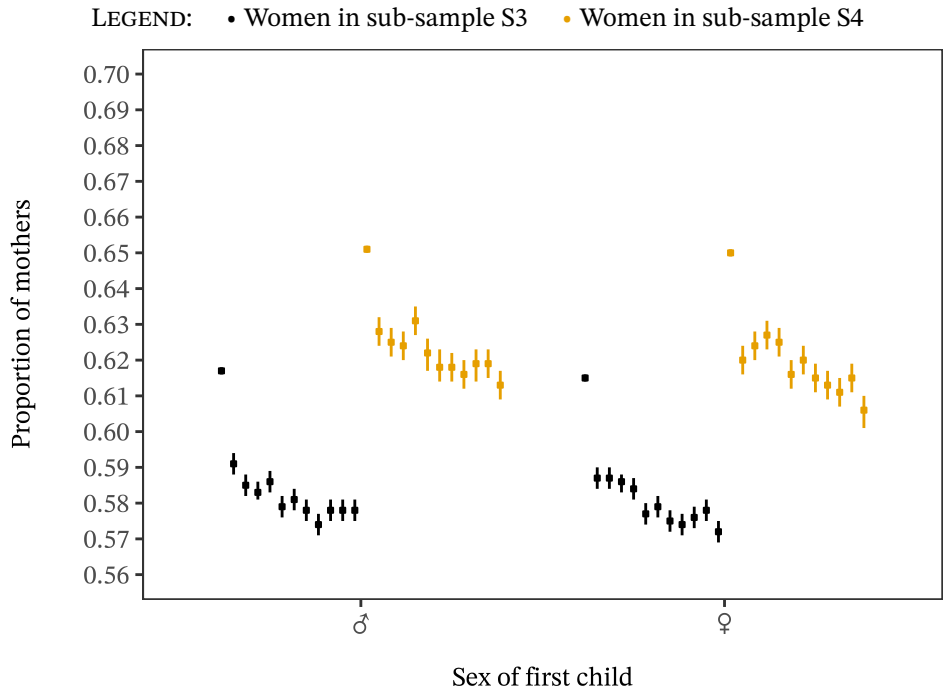


FIG 4. Fertility choice amongst women in sub-samples S3 and S4.

FIG 5. Fertility choice amongst women in sub-samples S5 and S6.

Sex is determined at random—and so too the occurrence of a multiple second birth—and is a pre-determined and immutable characteristic. It is, however, a misconception that male and

TABLE 5
Demographic membership amongst women in the U.S.A. (estimated for 2023).

Ethnicity	Race	Age group		
		20–24 years	25–29 years	30–35 years
Hispanic	White	2,226,971	2,121,581	2,087,278
	Black	130,242	121,416	130,470
	American Indian or Alaska Native	76,798	73,286	71,248
	Asian	25,641	24,702	26,086
	Native Hawaiian and other Pacific Islander	8,814	8,401	8,897
	Multiracial	93,080	82,022	74,335
Non-Hispanic	White	5,518,156	5,670,972	6,205,777
	Black	1,487,082	1,555,513	1,738,159
	American Indian or Alaska Native	86,927	88,237	95,217
	Asian	638,423	747,223	893,693
	Native Hawaiian and other Pacific Islander	21,655	22,684	26,130
	Multiracial	370,572	326,425	283,577

Source: U.S. Census Bureau (2020).

female children are equally likely; male children are slightly more likely than female children (we refer the reader to Appendix C and fig. 12).

7.5. Implementation

We assume the Jeffreys prior over response, treatment and attributes. That is, we associate every combination—comprising a level of response, a level of treatment and a level of attributes—with a parameter of the Dirichlet distribution, and set these parameters equal to one half. We do this regardless of how we define attributes, with one exception; the occurrence of a multiple second birth is incompatible with a mother having fewer than three children, and we set the parameter associated with this combination equal to zero in all cases in which we utilise information about this event. We emphasise that the Jeffreys prior—and the Dirichlet prior that is obtained if the aforementioned exception is made—is improper in this setting, in the sense that is ascribes positive probability to pseudo-distributions that falsify the model. We update the Jeffreys prior using data contained in either sub-sample S5 or sub-sample S6, by adding the number of times that a combination of response, treatment and attributes occurs to its associated parameter.

REMARK 3. The Jeffreys prior favours rejection of the model—a property that is driven by how much density it ascribes to distributions at the extremities of the simplex.[19] The posterior inherits this property if data does not or insufficiently updates some parameters of the Jeffreys prior; for example, if some combinations of race rarely arise. We emphasise that this is a practical issue—being driven by our choice of prior—and is arguably not something to concern the pure

[19]We conduct a simple simulation exercise in which we repeatedly draw from a Jeffreys prior with eight parameters. This exercise is designed to mimic an empirical setting in which there are two levels of instruments and a single level of covariates. We find that 69.9% of pseudo-distributions falsify the model. This probability compounds, and is 100.0% if there are ten levels of covariates. In contrast, the corresponding probabilities are 46.4% and 99.8% if we repeatedly draw from a uniform prior, with this probability falling further as the level of the (symmetric) concentration parameter is increased (i.e., as draws increasingly concentrate around the mid-point of the simplex).

TABLE 6
The occurrence of a multiple second birth and family composition amongst women in S5 and S6.

	Multiple second birth	Sex of first two children			
		♂♀	♀♂	♂♀	♂♂
Sub-sample S5	12,966	269,702	273,890	276,530	295,150
Sub-sample S6	7,245	156,956	174,703	163,871	160,582

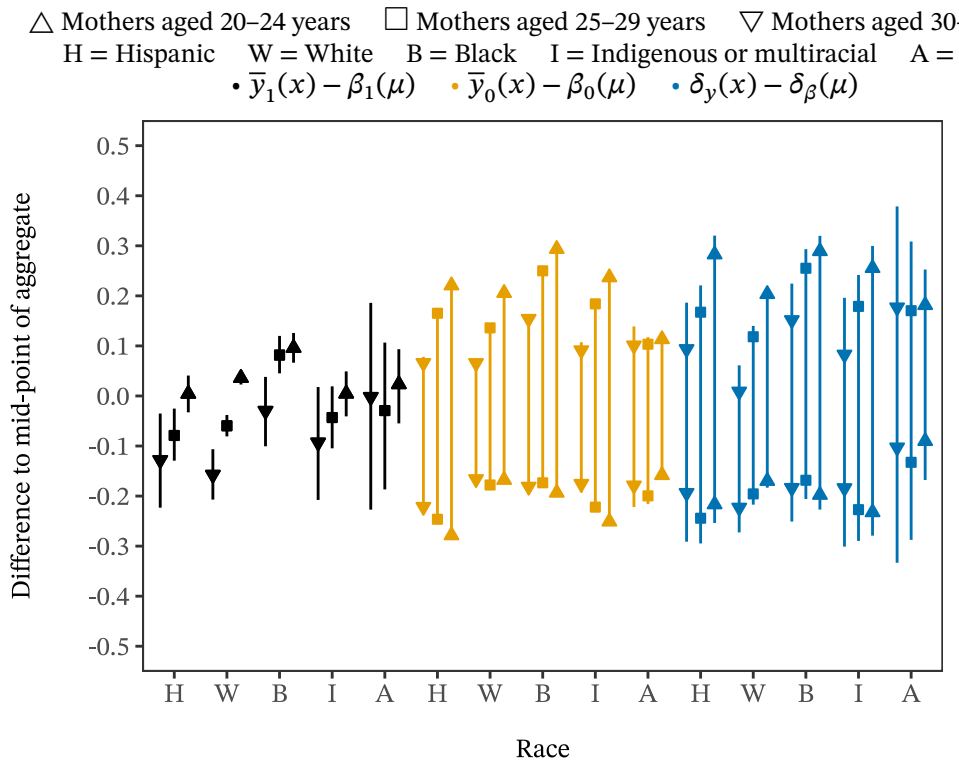
Bayesian who has full faith in their choice. We, however, select the Jeffreys prior because it is uninformative with the consequence that the resulting posterior is largely data-driven; that the Jeffreys prior holds that the model is highly implausible is an unintended—and, from our perspective, unfortunate—side-effect that imposes a high burden of proof. We recommend, therefore, adopting a categorisation of age and of race and ethnicity, that is sufficiently saturated by data, in the same way that one might take care in specifying regressors so as to avoid degeneracy of a design matrix in the context of linear regression.[20]

In view of Remark 3, we choose to limit the number of covariates that we include. More covariates increase the probability that the model is rejected, both mechanically—because each included covariate increases the dimension of the simplex from which the posterior samples and so too the probability of rejection—and due to more diluted updating of the prior—because the same number of observations update more parameters. Including too many covariates (relative to the quantity and concentration of data) increases the probability of rejection to the point at which the question of whether the model is a suitable representation of reality moot. To give the model a fighting chance so-to-speak, we allocate mothers to one of 15 categories based upon their age and race. Our choice of categories is informed by demographic statistics and projections (U.S. Census Bureau, 2020) and what these suggest about the prevalence of certain demographic groups in the wider US population (see Table 5). We divide mothers into one of three age groups (mothers aged 20–24 years, mothers aged 25–29 years, and mothers aged 30–34 years); and we divide mothers into one of five racial groups (white and Hispanic—Hispanic, for brevity; white and non-Hispanic—white, for brevity; black; American Indian or Alaska Native, or multiracial—indigenous or multiracial, for brevity; and Asian, or Native Hawaiian or other Pacific Islander—Asian, for brevity). Our choice of categories is not an ideal one but is a pragmatic one, intended to ensure that at least 1.0% of the wider US population falls into each group.[21]

We calculate estimates and α-proportion robust credible regions for three separate specifications. We set α to the conventional 95.0% level throughout. Our first specification ignores a mother's age and race (i.e., all mothers are allocated to a single category, to mimic a model absent covariates), and is included as an interesting baseline. Our second specification divides mothers

[20]We speculate that this is why Angrist & Evans's regression design translates to six separate racial categories, rather than the (more than) 45 separate categories that the decennial census and the ACS contain information about.

[21]We achieve this objective in the wider US population; the least-populous category—indigenous or multiracial mothers aged 30–34 years—accounts for 1.6% of the wider US population. We do not achieve this objective in sub-samples S5 or S6; the least-populous category—Asian mothers aged 20–24 years—accounts for 0.1% or 0.1% of mothers depending upon the sub-sample that is considered—equating to 881 or 544 mothers. NB: Asian mothers constitute only 1.2% or 1.4% of mothers aged 20–24 depending upon the sub-sample that is considered; in contrast, Asian women constitute 4.4% of all women aged 20–24 in sub-sample S1. Part of this under-representation is due to the growth in the share of this racial group combined with an increasing age at first birth; part of this under-representation is due to a preference for fewer children—Asian women are around half as likely as their white counterparts to have a second child.

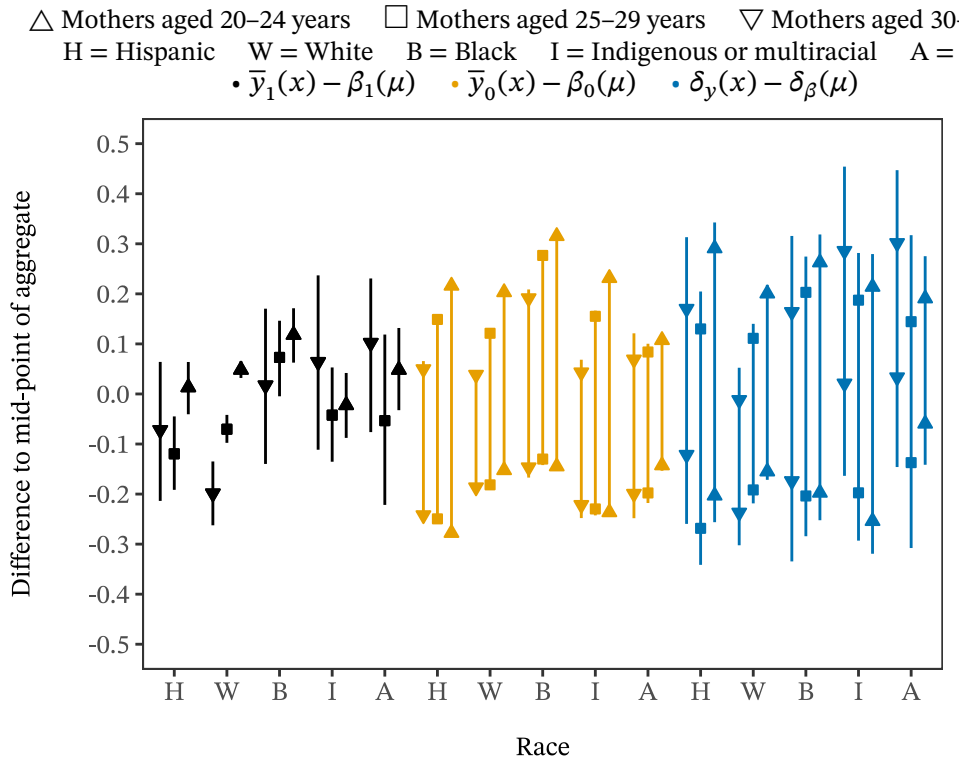


The midpoints of estimates of aggregate criteria (using stochastic weights) are subtracted from conditional criteria; α-proportion robust credible regions are depicted as lines; estimates are depicted as the intervals between points.

FIG 6. Differences between conditional criteria relative to aggregate criteria amongst women in sub-sample S5.

into the aforementioned 15 categories, and aggregates conditional criteria according to the shares of their corresponding categories in each pseudo-distribution. That is, we use the marginal distributions that are obtained from each pseudo-distribution to obtain the aggregate criteria, with the result that the weights that we use are stochastic. Our third specification divides mothers into the aforementioned 15 categories, and aggregates conditional criteria according to the shares of their corresponding categories in the demographic statistics and projections (U.S. Census Bureau, 2020). That is, we use the marginal distribution that is implied by Table 5, with the result that the weights that we use are fixed. In all three specifications, we use family composition—specifically, whether a mother's first two children have the same sex—in conjunction with the occurrence of a multiple second birth.

Although the occurrence of a multiple second birth is an especially rare event (see Table 6), it conveys considerable information. Indeed, the mechanical effect of the occurrence of a multiple second birth upon family size—it necessarily means that a mother has more than two children—is sufficient to uniquely identify one of the two conditional average structural functions, as an artefact of how the conditional probability of treatment contingent upon instruments is zero. Generally, such *sharp* instruments (i.e., instruments that set the probability of treatment equal to zero or one) are useful in partially identifying settings for delivering sets that have low Lebesgue volume. By using family composition in conjunction with the occurrence of a multiple second birth, we balance the high incidence–low information yield of the former with the low incidence–high information yield of the latter. This construction is made possible by the flexibility of the treatment equation; the lack of functional form facilitates using any combination of the occurrence of a multiple second birth with family composition. Since family composition is irrelevant



The midpoints of estimates of aggregate criteria (using stochastic weights) are subtracted from conditional criteria; α-proportion robust credible regions are depicted as lines; estimates are depicted as the intervals between points.

FIG 7. Differences between conditional criteria relative to aggregate criteria amongst women in sub-sample S6.

to the fertility decision if a multiple second birth occurs, we propose a hierarchical definition that adheres to the primacy of the occurrence of a multiple second birth. Moreover, since there appears to be preference for variety in the sex of children (see fig. 5) rather than an explicit preference for children of a particular sex, we collapse the possible permutations of sex into an indicator for whether a mother's first two children have the same sex.

We report the results that we obtain for the aggregate criteria in Table 7; and we report the results that we obtain for the conditional criteria—from which we subtract the aggregate criteria, in order to visualise these criteria on a common scale—in figs. 6 and 7. We draw attention to three things.

First, we draw attention to the posterior plausibility of the model (see Table 7). We assert that the model is plausible; despite the rarity of some outcomes (i.e., of some combinations of age and race, and of the occurrence of a multiple second birth), the model is not falsified in 86.9% and 5.1% of pseudo-samples depending upon the sub-sample that is considered.[22] Whilst we reiterate that this does not amount to verification of the model, it does provide some reassurance about our choice of economic variates and the appropriateness of the assumptions that we make about the data generating process.

Second, we draw attention to the wide and largely uninformative estimates that the model produces (see Table 7 and figs. 6 and 7). Family composition conveys very little information,

[22]The drop in plausibility between sub-sample S5 and sub-sample S6 is likely attributable to the relative sizes of the two sub-samples (sub-sample S5 comprises nearly twice as many mothers as sub-sample S6); combined with the rarity of some outcomes, fewer observations imply a lower chance of departure from uniformity.

TABLE 7
Estimates and α-proportion robust credible regions.

	Specification	Posterior plausibility	$\beta_1(\mu)$	$\beta_0(\mu)$	$\delta_\beta(\mu)$
Sub-sample S5	Baseline[1.]	100.0%	0.607 [0.598,0.616]	[0.420,0.800] [0.419,0.797]	[-0.189,0.187] [-0.198,0.196]
	Stochastic[2.]	86.9%	0.600 [0.590,0.609]	[0.420,0.795] [0.419,0.797]	[-0.196,0.179] [-0.205,0.189]
	Fixed[3.]	86.9%	0.562 [0.548,0.577]	[0.418,0.763] [0.416,0.765]	[-0.201,0.145] [-0.215,0.159]
Sub-sample S6	Baseline[1.]	100.0%	0.561 [0.548,0.573]	[0.410,0.763] [0.408,0.765]	[-0.203,0.151] [-0.215,0.163]
	Stochastic[2.]	5.1%	0.552 [0.539,0.564]	[0.410,0.763] [0.408,0.765]	[-0.211,0.142] [-0.224,0.154]
	Fixed[3.]	5.1%	0.518 [0.498,0.538]	[0.399,0.733] [0.397,0.736]	[-0.215,0.118] [-0.234,0.139]

Notes: All specifications use the occurrence of a multiple second birth in combination with family composition—specifically, whether a mother's first two children have the same sex. We draw 10,000 pseudo-distributions from the posterior. See eqs. (5.11) to (5.14) for definitions of the criteria of interest.
[1.] All mothers are allocated to a single category, to mimic a model absent covariates.
[2.] Conditional criteria are aggregated according to the shares of their corresponding categories in each pseudo-distribution.
[3.] Conditional criteria are aggregated according to the shares of their corresponding categories in the demographic statistics and projections (U.S. Census Bureau, 2020).

the preference for variety in the sex of children being only weak (see fig. 5); this negative result is offset somewhat by how common it is that a mother's first two children have the same sex, which leads to a low degree of statistical uncertainty. The occurrence of a multiple second birth conveys considerable information, which is sufficient to point identify one of the two conditional average structural functions; this positive result is offset somewhat by the rarity of the occurrence of a multiple second birth, which leads to a high degree of statistical uncertainty. That we are unable to draw any meaningful inference as to the sign of the effect of additional children upon maternal employment—or, failing this, a useful limit upon the absolute magnitude of the effect—is disappointing.

Third, we draw attention to the coincidence of the estimates that the model produces (see Table 7 and figs. 6 and 7). The estimates that we obtain by using the baseline specification versus the estimates that we obtain by using the stochastic specification broadly agree—a result that is compatible with the idea that the occurrence of a multiple second birth and family composition are truly random events (i.e., they are uncorrelated with age and race), and is something that we assert it is reasonable to assume (we refer the reader to Appendix C and fig. 12). We reiterate that the coincidence of the estimates that the model produces—and the idea that age and race can be ignored in the aggregate—does not mean that the effect of additional children upon maternal employment is uniform; there are some differences in the effect of additional children upon maternal employment for certain groups—-notably, black mothers with more than two children are between 6.0% and 21.6% more likely to work than their white counterparts. That we are able to detect some differences in the effect of additional children upon maternal employment for certain groups provides some reassurance about our choice of economic variates and, especially, about our inclusion of race—this appears to be an important influence upon maternal employ-

ment, whether directly or as a proxy for other economic variates such as income or wealth that it is correlated with.

8. Conclusion

We summarise our contributions to the economic literature as follows.

We propose and study a non-parametric instrumental variable model; we show that the model is partially identifying. We provide a sharp characterisation of the identification region for several criteria of interest, and provide testable implications that are necessary and sufficient to detect any observable distributions that are incompatible with the model. We extend our analysis to handle several conditional or full statistical independence conditions hold. We use the model to consider the question of whether and by how much additional children affect maternal employment at the extensive margin. We adopt a Bayesian posture towards the inferential problem; we follow the robust approach of Giacomini & Kitagawa (2021) and assume the Jeffreys prior over response, treatment and attributes—which we associate with whether a mother worked in the previous year, with whether a mother has more than two children, and with family composition and the occurrence of a multiple second birth as well as with age and race, respectively. We find that the Jeffreys prior favours rejection of the model—a property that is driven by how much density it ascribes to distributions at the extremities of the simplex.. We emphasise that this is a practical issue if data does not or insufficiently updates some parameters of the Jeffreys prior. We are, therefore, careful to allocate mothers to one of 15 categories based upon their age and race, so that our categorisation is sufficiently saturated by data. We are unable to draw any substantive conclusions as to the effect of additional children on maternal employment due to the wide and largely uninformative estimates that the model produces. We do, however, find that the model is plausible, and that we can—and, arguably, should—omit age and race.

Whether it is desirable that the Jeffreys prior favours rejection of the model is subjective, and depends upon the policymaker's risk aversion (Knightian or regular; Knight, 1921) and preference towards the status quo. Regardless of its desirability though, it is clear that the Jeffreys prior imposes a high burden of proof—something that is also true of the uniform prior, albeit to a lesser extent. Relative to some applications, we possess a vast amount of data; yet despite this abundance, the rarity with which some outcomes are observed—and the implied lack of divergence of the posterior from the Jeffreys prior—means that we frequently reject the model not because of the evidence that we have but because of the absence of more. The Jeffreys prior is a common choice of prior generally but, here, is also one that is in keeping with the theme that underpins the robust approach of Giacomini & Kitagawa (2021) and minimally restrictive structural economic modelling—an uninformative prior is a seemingly natural counterpart to the deliberate stripping back of the restrictions that a model embeds to the bare minimum as are needed to guarantee identification to some degree. We conclude that the usefulness of the model—when used in conjunction with the robust approach of Giacomini & Kitagawa (2021)—is limited to applications in which whether outcomes are sufficiently saturated by data is not a concern, or in which covariates can be omitted. We suggest that both scenarios require instruments that are balanced (i.e., are as likely to realise one level as another, or close to this) and truly random. We also wonder whether this property—that the Jeffreys prior favours rejection of the model—extends to other distinct economic environments in which there is a different natural conjugate prior.

We are unable to draw any meaningful inference as to the sign of the effect of additional children upon maternal employment—or, failing this, a useful limit upon the absolute magnitude

of the effect. *Sharp* instruments (i.e., instruments that set the probability of treatment equal to zero or one) are useful in partially identifying settings for delivering sets that have low Lebesgue volume. We conclude that, in the absence of such instruments—and even with them—the usefulness of the model is limited to understanding what additional restrictions purchase—something that we assert is often useful to know.

APPENDIX A: PROOFS

REMARK 4 (Notation). In what follows, we adopt the following convention. We signify the power set over a collection by including an asterisk in the superscript position; for instance, we write $\mathfrak{F}_u^*$ to mean the power set over $\mathfrak{F}_u$. Where the collection is the image of a correspondence, we position the asterisk before the parentheses enclosing the arguments; for instance, we write $\mathfrak{S}^*(\mathbf{a})$ to mean the collection that is obtained by taking all possible unions (in the power set-sense) of sets in $\mathfrak{S}(\mathbf{a})$.

PROOF OF THEOREM 1. The class of admissible structures that induce a given observable distribution is, by definition, eq. (4.2). Recalling that Assumption 1 is implicit in the definition of the capacity functional, we can characterise those structures that deliver a given observable distribution by

$$\text{Artstein}(Q) \equiv \{P, \mathbf{h} : P(\mathcal{U}|\mathbf{a}) \leq \text{Capacity}(Q, \mathcal{U}, \mathbf{a}) \forall \mathcal{U} \subseteq \mathfrak{F}_u \,\&\, \mathbf{a} \in \mathfrak{F}_\mathbf{a}\} \tag{A.1}$$

which is sharp (Artstein, 1983). We recall that heterogeneity is finite. As such, the power set enumerates all closed sets and unions of closed sets on its support—as is required by the theory of random sets from which eq. (A.1) originates (Molchanov, 2005). We refine eq. (A.1) by limiting the class of populations that we consider to those that encode Assumption 2 (or Definition 1,Definition 2 or Assumption 4); the result is

$$\text{Artstein}(Q) \cap \{P, \mathbf{h} : P \in \mathfrak{P}\} = \text{IR}(\mathfrak{P}, Q) \tag{A.2}$$

which, by construction,[23] satisfies Assumptions 1 to 3 and is sharp.

Given that the capacity functional and population are each non-decreasing in heterogeneity, there exists a proper subset of the power set that is sufficient to characterise the sharp identification region that is elsewhere labelled the class of core-determining sets (Galichon & Henry, 2011). That is,

$$\text{Artstein}(Q) = \{P, \mathbf{h} : P(\mathcal{U}|\mathbf{a}) \leq \text{Capacity}(Q, \mathcal{U}, \mathbf{a}) \forall \mathcal{U}, \mathbf{a} \in \mathfrak{S}(\mathbf{a}), \mathfrak{F}_\mathbf{a}\} \tag{A.3}$$

with eq. (4.3) describing the appropriate class of core-determining sets (Chesher & Rosen, 2017). Importantly, the class of core-determining sets contains only images of the contour functional corresponding to the prescribed levels of attributes, which is a consequence of how the contour functional connects.[24] We, therefore, establish that eq. (A.2) holds with equality and so constitutes the sharp identification region. □

[23] We again emphasise that eq. (A.1) embeds Assumption 1 via the capacity functional.

[24] The capacity functional is the complement of the containment functional (Molchanov, 2005). It is straightforward to translate results from one to the other by, for example, substituting intersections for unions. The class of core-determining sets contains only images of the contour functional and their unions (Chesher & Rosen, 2017, §Lemma 1) that cannot be partitioned into disjoint images (Chesher & Rosen, 2017, §Theorem 3).

PROOF OF THEOREM 2. We split the proof into several parts, adapting the results and technology of Kédagni & Mourifié (2020), Richardson & Robins (2024). We show that the set of structures that are compatible with eqs. (4.9) to (4.11) constitutes a subset of the set of structures that belong to IR($\mathfrak{P}, Q$). We show that any structure that violates a necessary condition of Assumptions 1 to 3 (i.e., any structure that does not belong to IR($\mathfrak{P}, Q$)) violates eqs. (4.9) to (4.11). We show that any structure that is compatible with eqs. (4.9) to (4.11) satisfies the Manski bounds, the Pearl bounds, or both. In doing so, we establish both that our characterisation of IR($\mathfrak{P}, Q$) is sharp and that our testable implications derive from a necessary condition of Assumptions 1 to 3. We focus on the case of strong random assignment due to its difficulty; we leverage the relationship between random assignment and joint statistical independence (every joint distribution that exhibits random assignment also exhibits joint statistical independence) and the relationship between joint statistical independence and marginal statistical independence (every joint distribution that exhibits joint statistical independence also exhibits marginal statistical independence) to trivially extend the proof to these other cases. Throughout, we exploit the presence of Assumption 1 in the definition of the counterfactuals, to consider a structure as a probability distribution over the counterfactuals rather than as the combination of a population and structural equations.

First, we show that the set of structures that are compatible with eqs. (4.9) to (4.11) constitutes a subset of the set of structures that belong to IR($\mathfrak{P}, Q$). We propose a structure that encodes Assumptions 1, 3 and 4 provided that eqs. (4.9) to (4.11) hold; in doing so, we leverage the relationship between strong random assignment and random assignment (every joint distribution that exhibits strong random assignment also exhibits random assignment). We emphasise that this structure may or may not coincide with the data generating process that delivers a given observable distribution, but is constructed so as to induce this observable distribution; to stress this point, we (circumflex) accent the structural objects—the population and structural equations—that correspond to the proposed structure. We let

$$p(y_1, y_0, t|\mathbf{a}) \equiv P(y_u(1, \mathbf{e}_x^{\mathrm{I}}\mathbf{a}) = y_1, y_u(0, \mathbf{e}_x^{\mathrm{I}}\mathbf{a}) = y_0, t_u(\mathbf{a}) = t|\mathbf{a}) \tag{A.4}$$

for convenience. We propose

$$\hat{p}(1,1,1|\mathbf{a}) \equiv Q(1,1|\mathbf{a}) - \xi(\mathbf{e}_x^{\mathrm{I}}\mathbf{a}) + \hat{p}(1,0,0|\mathbf{a}) \tag{A.5}$$

$$\hat{p}(1,1,0|\mathbf{a}) \equiv \underline{\xi}(\mathbf{e}_x^{\mathrm{I}}\mathbf{a}) - Q(1,1|\mathbf{a}) - \hat{p}(1,0,0|\mathbf{a}) \tag{A.6}$$

$$\hat{p}(1,0,1|\mathbf{a}) \equiv \xi(\mathbf{e}_x^{\mathrm{I}}\mathbf{a}) - \hat{p}(1,0,0|\mathbf{a}) \tag{A.7}$$

(A.8)

$$\hat{p}(1,0,0|\mathbf{a}) \equiv \min(\underline{\xi}(\mathbf{e}_x^{\mathrm{I}}\mathbf{a}) - Q(1,1|\mathbf{a}), \xi(\mathbf{e}_x^{\mathrm{I}}\mathbf{a}), \overline{\xi}(\mathbf{e}_x^{\mathrm{I}}\mathbf{a}) + \xi(\mathbf{e}_x^{\mathrm{I}}\mathbf{a}) - Q(1,1|\mathbf{a}) - Q(1,0|\mathbf{a}), Q(0,0|\mathbf{a}))$$

$$\hat{p}(0,1,1|\mathbf{a}) \equiv \overline{\xi}(\mathbf{e}_x^{\mathrm{I}}\mathbf{a}) + \xi(\mathbf{e}_x^{\mathrm{I}}\mathbf{a}) - Q(1,1|\mathbf{a}) - Q(1,0|\mathbf{a}) - \hat{p}(1,0,0|\mathbf{a}) \tag{A.9}$$

$$\hat{p}(0,1,0|\mathbf{a}) \equiv Q(1,1|\mathbf{a}) + Q(1,0|\mathbf{a}) - \underline{\xi}(\mathbf{e}_x^{\mathrm{I}}\mathbf{a}) + \hat{p}(1,0,0|\mathbf{a}) \tag{A.10}$$

$$\hat{p}(0,0,1|\mathbf{a}) \equiv 1 - \overline{\xi}(\mathbf{e}_x^{\mathrm{I}}\mathbf{a}) - \xi(\mathbf{e}_x^{\mathrm{I}}\mathbf{a}) - Q(0,0|\mathbf{a}) + \hat{p}(1,0,0|\mathbf{a}) \tag{A.11}$$

$$\hat{p}(0,0,0|\mathbf{a}) \equiv Q(0,0|\mathbf{a}) - \hat{p}(1,0,0|\mathbf{a}) \tag{A.12}$$

as a conditional distribution, where

$$\begin{aligned}\xi(x) &\equiv \max(0, 1 - \underline{q}_{1+1}(x) - \underline{q}_{0+1}(x) - \underline{q}_{0+0}(x), \overline{q}_{11}(x) - \underline{q}_{1+1}(x), \overline{q}_{00}(x) - \underline{q}_{0+0}(x)) \\ \underline{\xi}(x) &\equiv \max(\overline{q}_{11}(x), 1 - \underline{q}_{0+0}(x) - \underline{q}_{0+1}(x)) \\ \overline{\xi}(x) &\equiv \min(1 - \overline{q}_{00}(x), \underline{q}_{1+1}(x) + \underline{q}_{0+1}(x))\end{aligned} \tag{A.13}$$

are envelopes. We obtain eqs. (A.5) to (A.13) by means of a trivial extension of the conditional distribution that is proposed in Kédagni & Mourifié (2020); as such, we rely on several of the results that are stated for that conditional distribution. Specifically, the proposed distribution is proper—it is non-negative and sums to one—provided that eqs. (4.9) to (4.11) hold (Kédagni & Mourifié, 2020, §Section 9).[25] Adapting Richardson & Robins (2024, §Lemma 4), we utilise

$$\hat{P}(u = \upsilon) = \frac{\prod_{\mathbf{a} \in \mathfrak{F}_{\mathbf{a}}} \hat{p}(h_y(1, \mathbf{e}_x^{\intercal}\mathbf{a}, \upsilon), h_y(0, \mathbf{e}_x^{\intercal}\mathbf{a}, \upsilon), h_t(\mathbf{a}, \upsilon) | \mathbf{a})}{\prod_{x \in \mathfrak{F}_x} [\sum_{t \in \mathfrak{F}_t} \hat{p}(h_y(1, \mathbf{e}_x^{\intercal}\mathbf{a}, \upsilon), h_y(0, \mathbf{e}_x^{\intercal}\mathbf{a}, \upsilon), t | \underline{\mathbf{a}}(x))]^{|\mathfrak{F}_{\mathbf{z}|x}|-1}} \tag{A.14}$$

to construct a joint distribution from the proposed conditional distribution, where

$$\underline{\mathbf{a}}(x) \equiv \{\mathbf{a} : \mathbf{a} \text{ is the first point of support in } \mathfrak{F}_{\mathbf{a}|x}\} \tag{A.15}$$

extracts one—specifically, the first—level of attributes whose covariate-element has the specified value. The proposed joint distribution is proper—inheriting this property from the proposed conditional distribution—and induces the observable distribution.[26] By design, the proposed joint distribution is statistically independent of attributes (since eq. (A.14) is the product of *all* conditional distributions) with this property preserved via the summation operation through which we obtain the unconditional probability of sets of heterogeneity. As such, the proposed joint distribution satisfies random assignment; we thereby establish that the proposed joint distribution belongs to IR($\mathfrak{P}, Q$).

Second, we show that any structure that violates a necessary condition of Assumptions 1 to 3 (i.e., any structure that does not belong to IR($\mathfrak{P}, Q$)) violates eqs. (4.9) to (4.11). We label heterogeneity according to the levels of the potential outcomes that it induces, such that each level of heterogeneity induces at least one response or treatment that is different from any other level of heterogeneity in at least one counterfactual state. That is, we label heterogeneity so that its support has the form

$$\mathfrak{F}_u = \{u : \underbrace{(h_y(1, x, u))}_{\text{(A)}}{}_{x \in \mathfrak{F}_x} = \mathbf{y}_1, \underbrace{(h_y(0, x, u))}_{\text{(B)}}{}_{x \in \mathfrak{F}_x} = \mathbf{y}_0, \underbrace{(h_t(\mathbf{a}, u))}_{\text{(C)}}{}_{\mathbf{a} \in \mathfrak{F}_{\mathbf{a}}} = \mathbf{t}\}_{\mathbf{y}_1, \mathbf{y}_0, \mathbf{t} \in \mathfrak{F}_y^{2|\mathfrak{F}_x|} \times \mathfrak{F}_t^{|\mathfrak{F}_{\mathbf{a}}|}} \tag{A.16}$$

reiterate that a structure constitutes a probability distribution over every level of heterogeneity (that sums to one). An alternative view of the identification region then is as the solution to a

[25] It is trivial to show that the proposed distribution sums to one. We emphasise that eq. (A.8) is designed so that eqs. (A.6), (A.7), (A.9) and (A.12) are non-negative; eqs. (4.9) to (4.11) ensure that eq. (A.13) are of appropriate magnitude to guarantee that eqs. (A.5) and (A.10) are non-negative.

[26] We note that eqs. (A.5) and (A.7), eqs. (A.6) and (A.10), eqs. (A.9) and (A.11), and eqs. (A.8) and (A.12) each sum to an observable quantity.

linear programme, in which the probability of constituents of eq. (A.16) are associated with observable probabilities. Specifically, each observable probability fixes the (probability of) the component of $\mathbf{y}_1$ that relates to Ⓐ or the component of $\mathbf{y}_0$ that relates to Ⓑ (i.e., the response counterfactual corresponding to the specified level of treatment and attributes) and fixes the component of $\mathbf{t}$ that relates to Ⓒ (i.e., the treatment counterfactual corresponding to the specified level of attributes) but does not fix any other components. If the observable distribution is compatible with Assumptions 1 to 3 then a solution to the linear programme exists; otherwise, we are able to find $\mathfrak{T}(\mathbf{a}) \mapsto \mathfrak{F}_u^*$—the image of which we refer to as a test set—that violates

$$\sum_{\mathbf{a}\in\mathfrak{F}_\mathbf{a}} \text{Capacity}(Q, \mathfrak{T}(\mathbf{a}), \mathbf{a}) \geq 1 \text{ subject to } \cup_{\mathbf{a}\in\mathfrak{F}_\mathbf{a}} \mathfrak{T}(\mathbf{a}) \supseteq \mathfrak{F}_u \tag{A.17}$$

such that there is no feasible solution.[27] The number of possible test sets is large, but we can reduce the number of sets that we need to consider substantially by exploiting the presence of the minimum operator (if a smaller test set is able to cover the support of heterogeneity then a larger test set is redundant) and the coverage constraint (only certain combinations of test sets are able to cover the support of heterogeneity) in eq. (A.17). Invoking the intermediate results of Lemmata 1 to 4 that effect these exploits, we establish that it is sufficient to consider test sets that are non-empty for exactly one level of covariates and that are either empty or else comprise constituents of

$$\begin{aligned} \mathfrak{S}_2^{\text{Ref.}}(\mathbf{a}) &\equiv \{\mathfrak{U}_1 \cup \mathfrak{U}_2 : \mathfrak{U}_1 = \text{Contour}(y_1, 1, \mathbf{a}), \mathfrak{U}_2 = \text{Contour}(y_2, 0, \mathbf{a})\}_{y_1,y_2\in\mathfrak{F}_y^2} \\ \mathfrak{S}_3(\mathbf{a}) &\equiv \{\mathfrak{U}_1 \cup \mathfrak{U}_2 \cup \mathfrak{U}_3 : \mathfrak{U}_1, \mathfrak{U}_2, \mathfrak{U}_3 \in \mathfrak{S}(\mathbf{a})\} \end{aligned} \tag{A.18}$$

which are a refinement of all pairs of core-determining sets and all triplets of core-determining sets, respectively. Inverting the capacity functional, we make the association that

$$\begin{aligned} \mathfrak{S}_2^{\text{Ref.}}(\mathbf{a}) &= \{u : h_y(1, \mathbf{e}_x^\intercal \mathbf{a}, u) = y_1, h_y(0, \mathbf{e}_x^\intercal \mathbf{a}, u) = y_0\}_{y_1,y_0\in\mathfrak{F}_y^2} \\ \mathfrak{S}_3(\mathbf{a}) &= \{u : h_y(t, \mathbf{e}_x^\intercal \mathbf{a}, u) = y\}_{y,t\in\mathfrak{F}_y\times\mathfrak{F}_t} \end{aligned} \tag{A.19}$$

(i.e., each element of eq. (A.18) is associated with particular counterfactual responses and, importantly, incurs the same capacity as these alternative test sets). It is trivial to show that, in view of eq. (A.19), the only combinations of core-determining sets that can and *just* cover the support of heterogeneity are those combinations that deliver eqs. (4.9) to (4.11) (absent the minima).

Third, we show that any structure that is compatible with eqs. (4.9) to (4.11) satisfies the Manski bounds, the Pearl bounds, or both. We exploit the (binary) support of the counterfactual response to write

$$\overline{q}_{yt}(x) \leq 1 - P(y(t,x) = 1 - y|x) = P(y(t,x) = y|x) \tag{A.20}$$

such that the Manski bounds can be seen to imply a collection of upper and lower bounds on the probability of each counterfactual response. Moreover,

$$1 = P(y(t,x) = 1|x) + P(y(t,x) = 0|x) \leq \overline{q}_{1t}(x) + \overline{q}_{0t}(x) \tag{A.21}$$

[27] The interpretation of eq. (A.17) that we favour is that the sum of any upper bound on the probability of an event and any upper bound on the probability of its complement must sum to at least one. We construct a cover of the event and a cover of its complement via the union of test sets that we then relate to observable probabilities using the capacity functional.

and so we obtain the equivalence of eq. (4.9) and the Manski bounds. We can similarly write

$$P(y(t,x)=y, y(1-t,x)=1|x) + P(y(t,x)=y, y(1-t,x)=0|x) = P(y(t,x)=y|x) \tag{A.22}$$

and so obtain eqs. (4.10) and (4.11) via suitable addition.

Finally, we note that eq. (A.2) reduces to

$$P(\mathfrak{U}|x) = P(\mathfrak{U}|\mathbf{a}) \leq \text{Capacity}(Q, \mathfrak{U}, \mathbf{a}) \tag{A.23}$$

for particular choices of test sets under Assumption 2. Under random assignment, these test sets include all $\mathfrak{U} \in \mathfrak{S}_2^{\text{Ref.}}(\mathbf{a})$ and $\mathfrak{U} \in \mathfrak{S}_3(\mathbf{a})$; under joint statistical independence, these test sets include all $\mathfrak{U} \in \mathfrak{S}_2^{\text{Ref.}}(\mathbf{a})$ and $\mathfrak{U} \in \mathfrak{S}_3(\mathbf{a})$; and under marginal statistical independence conditions, these test sets include all $\mathfrak{U} \in \mathfrak{S}_3(\mathbf{a})$. □

PROOF OF THEOREM 3. The first two steps of our proof of Theorem 2 do not rely upon whether we impose Assumption 2 or Assumption 4. We, therefore, resume the proof at the third step.

Third, we show that any structure that is compatible with eqs. (4.9) to (4.11) satisfies the Manski bounds, the Pearl bounds, or both. We exploit the (binary) support of the counterfactual response to write

$$\overline{q}_{yt}(x) \leq 1 - P(y(t,x)=1-y) = P(y(t,x)=y) \tag{A.24}$$

such that the Manski bounds can be seen to imply a collection of upper and lower bounds on the probability of each counterfactual response. Moreover,

$$1 = P(y(t,x)=1) + P(y(t,x)=0) \leq \overline{q}_{1t}(x) + \overline{q}_{0t}(x) \tag{A.25}$$

and so we obtain the equivalence of eq. (4.9) and the Manski bounds. We can similarly write

$$P(y(t,x)=y, y(1-t,x)=1) + P(y(t,x)=y, y(1-t,x)=0) = P(y(t,x)=y) \tag{A.26}$$

and so obtain eqs. (4.10) and (4.11) via suitable addition.

Finally, we note that eq. (A.2) reduces to

$$P(\mathfrak{U}) = P(\mathfrak{U}|\mathbf{a}) \leq \text{Capacity}(Q, \mathfrak{U}, \mathbf{a}) \tag{A.27}$$

for particular choices of test sets under Assumption 2. Under random assignment, these test sets include all $\mathfrak{U} \in \mathfrak{S}_2^{\text{Ref.}}(\mathbf{a})$ and $\mathfrak{U} \in \mathfrak{S}_3(\mathbf{a})$; under joint statistical independence, these test sets include all $\mathfrak{U} \in \mathfrak{S}_2^{\text{Ref.}}(\mathbf{a})$ and $\mathfrak{U} \in \mathfrak{S}_3(\mathbf{a})$; and under marginal statistical independence conditions, these test sets include all $\mathfrak{U} \in \mathfrak{S}_3(\mathbf{a})$.

□

APPENDIX B: AUXILIARY RESULTS AND PROOFS

We present several intermediate results that we use to limit the number of test sets that we need to consider in appendix A.

LEMMA 1. *If* $\mathfrak{T}(\mathbf{a}) \mapsto \mathfrak{F}_u^*$ *violates eq.* (A.17) *then* $\mathfrak{T}(\mathbf{a}) \neq \mathfrak{F}_u$ *for all* $\mathbf{a} \in \mathfrak{F}_\mathbf{a}$.

PROOF OF LEMMA 1. If $\mathfrak{T}(\mathbf{a}) = \mathfrak{F}_u$ for at least one $\mathbf{a} \in \mathfrak{F}_\mathbf{a}$ then the summation of eq. (A.17) cannot be less than one since $\text{Capacity}(Q, \mathfrak{F}_u, \mathbf{a}) = 1$ for all $\mathbf{a} \in \mathfrak{F}_\mathbf{a}$. □

LEMMA 2. *If $\mathfrak{T}(\mathbf{a}) \mapsto \mathfrak{F}_u^*$ violates eq. (A.17) then there exists an alternative formulation of test sets that attains the same capacity than $\mathfrak{T}(\mathbf{a}) \mapsto \mathfrak{S}^*(\mathbf{a})$ and covers the support of heterogeneity by combining core-determining sets only.*

PROOF OF LEMMA 2. We claim that $\hat{\mathfrak{T}}(\mathbf{a}) = \cap_s\{\mathfrak{U} : \mathfrak{T}(\mathbf{a}) \subseteq \mathfrak{U} \,|\, \mathfrak{U} \in \mathfrak{S}^*(\mathbf{a})\}$ has the same capacity as $\mathfrak{T}(\mathbf{a}) \mapsto \mathfrak{F}_u^*$ and covers the support of heterogeneity.[28] □

LEMMA 3. *If $\mathfrak{T}(\mathbf{a}) \mapsto \mathfrak{S}^*(\mathbf{a})$ violates eq. (A.17) and is such that $\cup_{\mathbf{a}\in\mathfrak{F}_{\mathbf{a}|x}} \mathfrak{T}(\mathbf{a}) \neq \varnothing$ for more than one $x \in \mathfrak{F}_x$ then there exists an alternative formulation of test sets that attains a lower capacity than $\mathfrak{T}(\mathbf{a}) \mapsto \mathfrak{S}^*(\mathbf{a})$ and covers the support of heterogeneity by combining core-determining sets only.*

PROOF OF LEMMA 3. Suppose that $\mathfrak{T}(\mathbf{a}) \mapsto \mathfrak{S}^*(\mathbf{a})$ violates eq. (A.17) and is such that $\cup_{\mathbf{a}\in\mathfrak{F}_{\mathbf{a}|x}} \mathfrak{T}(\mathbf{a}) \neq \varnothing$ for more than one $x \in \mathfrak{F}_x$. If $\cup_{\mathbf{a}\in\mathfrak{F}_{\mathbf{a}|x}} \mathfrak{T}(\mathbf{a}) \neq \mathfrak{F}_u$ for all $x \in \mathfrak{F}_x$ then there exists $\hat{\mathfrak{T}}(\mathbf{a}) \mapsto \mathfrak{G}(\mathbf{a})$ such that $\hat{\mathfrak{T}}(\mathbf{a}) \subseteq \mathfrak{F}_u \setminus \mathfrak{T}(\mathbf{a})$ for all $\mathbf{a} \in \mathfrak{F}_\mathbf{a}$ with $\cap_{\mathbf{a}\in\mathfrak{F}_{\mathbf{a}|x}} \hat{\mathfrak{T}}(\mathbf{a}) \neq \varnothing$ for all $x \in \mathfrak{F}_x$. Given that each core-determining set fixes exactly two components of eq. (A.16), however, it follows that $\cap_{\mathbf{a}\in\mathfrak{F}_{\mathbf{a}|x}} \hat{\mathfrak{T}}(\mathbf{a}) \neq \varnothing$ and so $\cup_{\mathbf{a}\in\mathfrak{F}_\mathbf{a}} \mathfrak{T}(\mathbf{a}) \neq \mathfrak{F}_u$. To make concrete this point, we emphasise that

$$\text{(B.1)} \qquad \cap_{\mathbf{a}\in\mathfrak{F}_{\mathbf{a}|x}} \hat{\mathfrak{T}}(\mathbf{a}) \supseteq \{u : (y_u(1,x), y_u(0,x)) = \mathbf{y}_x, (t_u(\mathbf{a}))_{\mathbf{a}\in\mathfrak{F}_{\mathbf{a}|x}} = \mathbf{t}_x \,|\, \mathbf{y}_x, \mathbf{t}_x \in \mathfrak{F}_y^2 \times \mathfrak{F}_t^{|\mathfrak{F}_{\mathbf{z}|x}|}\}$$

and

$$\begin{aligned}
&\cap_{\mathbf{a}\in\mathfrak{F}_\mathbf{a}} \hat{\mathfrak{T}}(\mathbf{a}) \\
\text{(B.2)} \quad &= \cap_{x\in\mathfrak{F}_x} \cap_{\mathbf{a}\in\mathfrak{F}_{\mathbf{a}|x}} \hat{\mathfrak{T}}(\mathbf{a}) \\
\text{(B.3)} \quad &\supseteq \cap_{x\in\mathfrak{F}_x}\{u : (y_u(1,x), y_u(0,x)) = \mathbf{y}_x, (t_u(\mathbf{a}))_{\mathbf{a}\in\mathfrak{F}_{\mathbf{a}|x}} = \mathbf{t}_x \,|\, \mathbf{y}_x, \mathbf{t}_x \in \mathfrak{F}_y^2 \times \mathfrak{F}_t^{|\mathfrak{F}_{\mathbf{z}|x}|}\} \\
\text{(B.4)} \quad &= \{u : (y_u(1,x), y_u(0,x))_{x\in\mathfrak{F}_x} = (\mathbf{y}_x)_{x\in\mathfrak{F}_x}, (t_u(\mathbf{a}))_{\mathbf{a}\in\mathfrak{F}_\mathbf{a}} = (\mathbf{t}_x)_{x\in\mathfrak{F}_x} \,|\, \mathbf{y}_x, \mathbf{t}_x \in \mathfrak{F}_y^2 \times \mathfrak{F}_t^{|\mathfrak{F}_{\mathbf{z}|x}|}\} \\
\text{(B.5)} \quad &= \{u : (y_u(1,x))_{x\in\mathfrak{F}_x} = \mathbf{y}_1, (y_u(0,x))_{x\in\mathfrak{F}_x} = \mathbf{y}_0, (t_u(\mathbf{a}))_{\mathbf{a}\in\mathfrak{F}_\mathbf{a}} = \mathbf{t} \,|\, \mathbf{y}_1, \mathbf{y}_0, \mathbf{t} \in \mathfrak{F}_y^{2|\mathfrak{F}_x|} \times \mathfrak{F}_t^{|\mathfrak{F}_\mathbf{a}|}\}
\end{aligned}$$

constitute several levels or a single level of heterogeneity that are non-empty according to eq. (A.16). We note that eq. (B.2) to eq. (B.3) substitutes eq. (B.1); we note that eq. (B.3) to eq. (B.4) incorporates the intersection into the conditioning event; and we note that eq. (B.4) to eq. (B.5) follows upon suitable definition of $\mathbf{y}_1$, of $\mathbf{y}_0$, and of $\mathbf{t}$. We conclude that $\cup_{\mathbf{a}\in\mathfrak{F}_{\mathbf{a}|x}} \mathfrak{T}(\mathbf{a}) = \mathfrak{F}_u$ for at least one $x \in \mathfrak{F}_x$, which we label x_i. We claim that

$$\text{(B.6)} \qquad \ddot{\mathfrak{T}}(\mathbf{a}) = \begin{cases} \mathfrak{T}(\mathbf{a}) & \text{if } \mathbf{a} \in \mathfrak{F}_{\mathbf{a}|x_i} \\ \varnothing & \text{if } \mathbf{a} \notin \mathfrak{F}_{\mathbf{a}|x_i} \end{cases}$$

attains a lower capacity than $\mathfrak{T}(\mathbf{a}) \mapsto \mathfrak{S}^*(\mathbf{a})$ and covers the support of heterogeneity by combining core-determining sets only. □

[28] The intersection operator (over the test sets that adhere to the definition) means that we capture the smallest (union of) core-determining sets that covers a test set; this union cannot be smaller than the test set itself.

LEMMA 4. *If* $\mathfrak{T}(\mathbf{a}) \mapsto \mathfrak{S}^*(\mathbf{a})$ *violates eq.* (A.17) *and is such that* $\mathfrak{T}(\mathbf{a}_i) \in \mathfrak{S}(\mathbf{a}_i)$ *for at least one* $n = 1, ..., |\mathfrak{F}_{\mathbf{a}}|$ *then there exists an alternative formulation of test sets that attains a lower capacity than* $\mathfrak{T}(\mathbf{a}) \mapsto \mathfrak{S}^*(\mathbf{a})$ *and covers the support of heterogeneity by combining certain unions of core-determining sets only.*

PROOF OF LEMMA 4. Suppose that $\mathfrak{T}(\mathbf{a}) \mapsto \mathfrak{S}^*(\mathbf{a})$ violates eq. (A.17). We eliminate individual core-determining sets and certain pairs of core-determining sets. To facilitate this discussion, we define

$$\mathfrak{S}_2(\mathbf{a}) \equiv \{\mathfrak{U}_1 \cup \mathfrak{U}_2 : \mathfrak{U}_1, \mathfrak{U}_2 \in \mathfrak{S}(\mathbf{a})\} \tag{B.7}$$

as all pairs of core-determining sets.

We further suppose that $\mathfrak{T}(\mathbf{a}) \mapsto \mathfrak{S}^*(\mathbf{a})$ is such that $\mathfrak{T}(\mathbf{a}) \in \mathfrak{S}(\mathbf{a})$ or $\mathfrak{T}(\mathbf{a}) \in \mathfrak{S}_2(\mathbf{a})$ but $\mathfrak{T}(\mathbf{a}) \notin \mathfrak{S}_2^{\text{Ref.}}(\mathbf{a})$ for at least one $\mathbf{a} \in \mathfrak{F}_{\mathbf{a}}$, which we label $\mathbf{a}_i$ (comprising x_i and z_i). If $\cup_{\mathbf{a} \in \mathfrak{F}_{\mathbf{a}|x_i} \setminus \mathbf{a}_i} \mathfrak{T}(\mathbf{a}) \neq \mathfrak{F}_u$ then there exists $\hat{\mathfrak{T}}(\mathbf{a}) \mapsto \mathfrak{G}(\mathbf{a})$ such that $\hat{\mathfrak{T}}(\mathbf{a}) \subseteq \mathfrak{F}_u \setminus \mathfrak{T}(\mathbf{a})$ for all $\mathbf{a} \in \mathfrak{F}_{\mathbf{a}|x_i} \setminus \mathbf{a}_i$ with $\cap_{\mathbf{a} \in \mathfrak{F}_{\mathbf{a}|x_i} \setminus \mathbf{a}_i} \hat{\mathfrak{T}}(\mathbf{a}) \neq \varnothing$. Given that each core-determining set fixes exactly two components of eq. (A.16), however, it follows that $\cap_{\mathbf{a} \in \mathfrak{F}_{\mathbf{a}|x_i} \setminus \mathbf{a}_i} \hat{\mathfrak{T}}(\mathbf{a}) \not\subseteq \mathfrak{T}(\mathbf{a}_i)$ and so $\cup_{\mathbf{a} \in \mathfrak{F}_{\mathbf{a}|x_i}} \mathfrak{T}(\mathbf{a}) \neq \mathfrak{F}_u$. To make concrete this point, we emphasise that

$$\cap_{\mathbf{a} \in \mathfrak{F}_{\mathbf{a}|x_i} \setminus \mathbf{a}_i} \hat{\mathfrak{T}}(\mathbf{a}) \supseteq \{u : y_u(1,x) = y_1, y_u(0,x) = y_0, (t(\mathbf{a}))_{\mathbf{a} \in \mathfrak{F}_{\mathbf{a}|x_i} \setminus \mathbf{a}_i} = \mathbf{t}, t(\mathbf{a}_i) = t \,|\, \text{Ⓓ}\} \tag{B.8}$$

constitutes a single level of heterogeneity that is non-empty according to eq. (A.16) and has non-empty intersection with exactly two individual core-determining sets—sets that are associated with different levels of treatment counterfactuals, and where

$$\text{Ⓓ} \equiv y_1, y_0, \mathbf{t}, t \in \mathfrak{F}_y^2 \times \mathfrak{F}_t^{|\mathfrak{F}_{z|x_i}|} \tag{B.9}$$

We conclude that $\cup_{\mathbf{a} \in \mathfrak{F}_{\mathbf{a}} \setminus \mathbf{a}_i} \mathfrak{T}(\mathbf{a}) = \mathfrak{F}_u$.

$$\ddot{\mathfrak{T}}(\mathbf{a}) = \begin{cases} \mathfrak{T}(\mathbf{a}) & \text{if } \mathbf{a} \in \mathfrak{F}_{\mathbf{a}} \setminus \mathbf{a}_i \\ \varnothing & \text{if } \mathbf{a} = \mathbf{a}_i \end{cases} \tag{B.10}$$

attains a lower capacity than $\mathfrak{T}(\mathbf{a}) \mapsto \mathfrak{S}^*(\mathbf{a})$ and covers the support of heterogeneity by combining unions of core-determining sets only. □

APPENDIX C: ADDITIONAL SUMMARY STATISTICS

We present summary statistics relating to fertility in fig. 8. The figure reports estimates of the mean and associated 95% frequentist confidence interval of age at first birth—the calculated age at which a woman gave birth to the eldest child that she reports having that is present in the household. We decompose this age according to year, and also report a comparable statistic for the husbands of married women. Age at first birth increases over all of the survey years.

We present summary statistics relating to employment in figs. 9 to 11. Each figure reports estimates of the mean and associated 95% frequentist confidence interval of answers to several survey questions. Figure 9 details the number of hours per week that a woman reports as typically working during the previous 12 months; fig. 10 details whether a woman reports having worked

FIG 8. Age at first birth amongst women and their husbands (if applicable) in sub-samples S5 and S6.

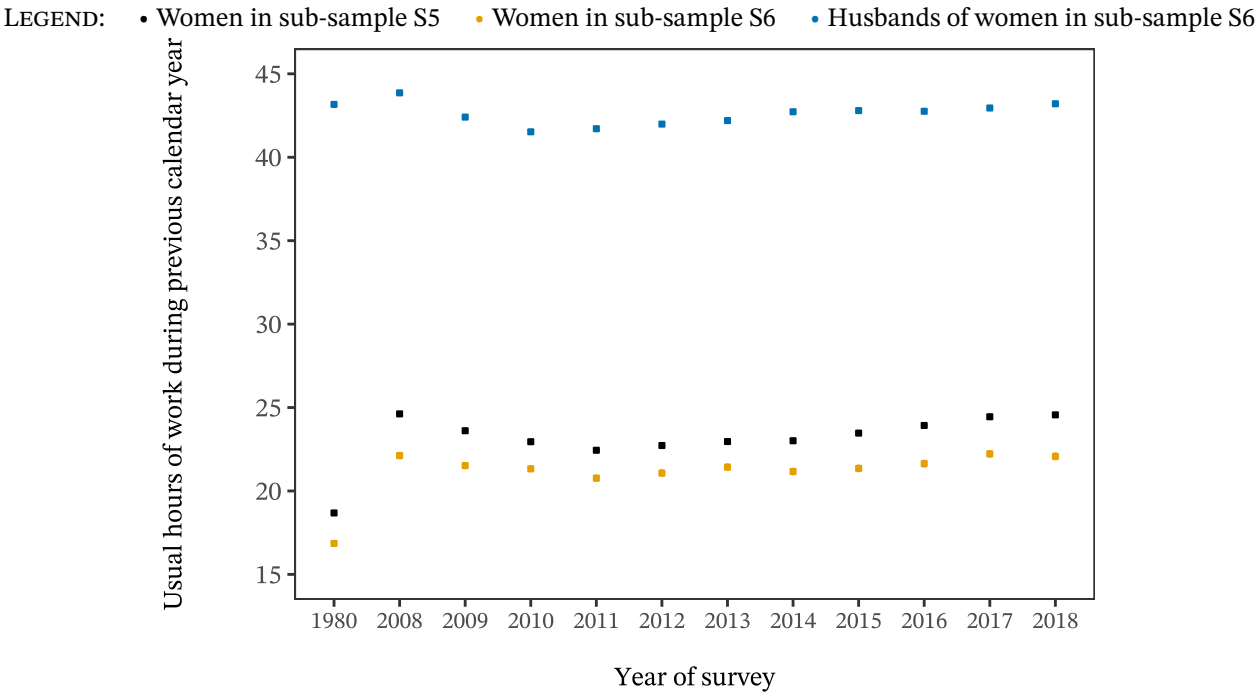


FIG 9. Intensive labour supply margin amongst women and their husbands (if applicable) in sub-samples S5 and S6.

at all for profit, pay, or as an unpaid family worker during the previous 12 months; and fig. 11 details the income that a woman reports as earning (point and line-range) or as total household income (ribbon; capturing the 95% frequentist confidence interval only). We decompose these

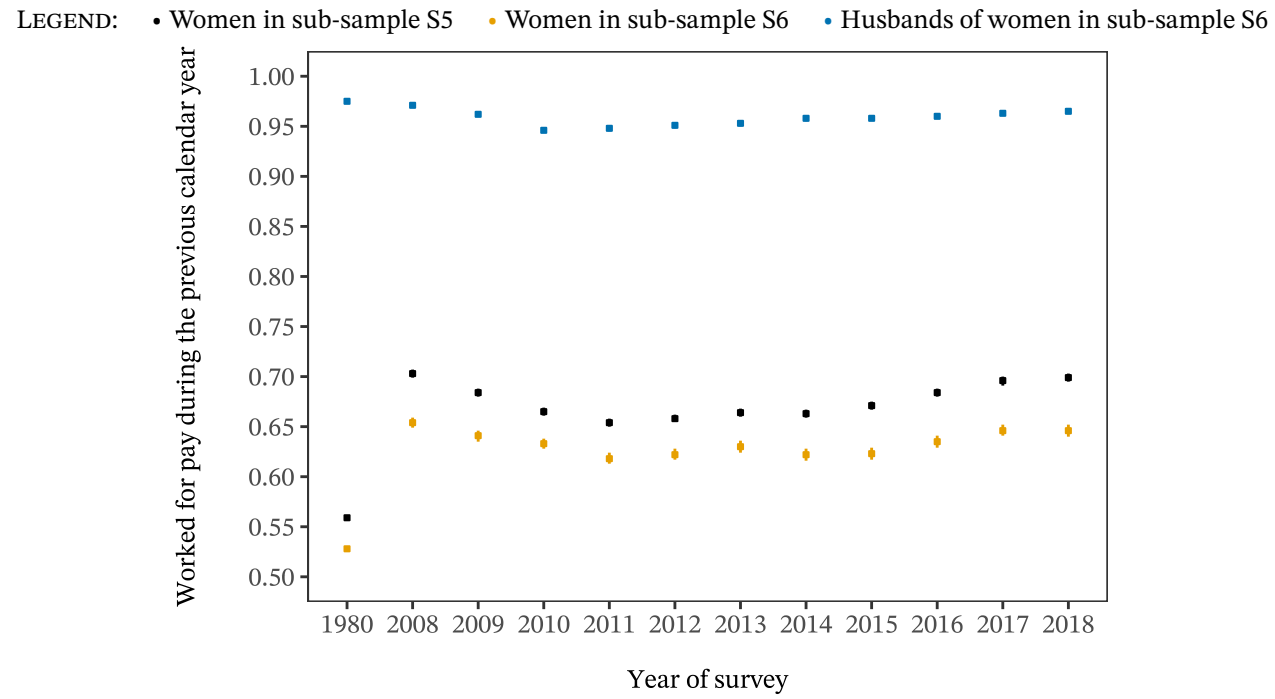


FIG 10. Extensive labour supply margin amongst women and their husbands (if applicable) in sub-samples S5 and S6.

answers according to year, and also report a comparable statistic for the husbands of married women. Female labour supply—but not male labour supply—increases between the 1980 survey year and the 2008 through 2018 survey years, but there is no discernible trend year-on-year during the 2008 through 2018 survey years; earned income and total household income increase between the 1980 survey year and the 2008 through 2018 survey years, but there is no discernible trend year-on-year during the 2008 through 2018 survey years aside from a reduction in income during the 2008 through 2011 survey years that coincides with a wider economic recession.

We present summary statistics relating to family composition in fig. 12. The figure reports estimates of the probability and associated 95% frequentist confidence interval that the first two children that a woman reports having that are present in the household are female or male. We decompose this probability according to year. Male children are slightly more likely than female children, but there is no difference in the probability of a particular family composition over all of the survey years.

Acknowledgments

A preliminary version of this paper appeared as a chapter in Rowley (2024, §Chapter E); the author would like to thank his supervisors—Andrew Chesher and Toru Kitagawa—for their guidance, and his examiners—Karim Chalak and Dennis Kristensen—for their insightful comments. The author would also like to thank Áureo de Paula, Adam Rosen, and Debbie Yang for their helpful discussion of the topic at various junctures.

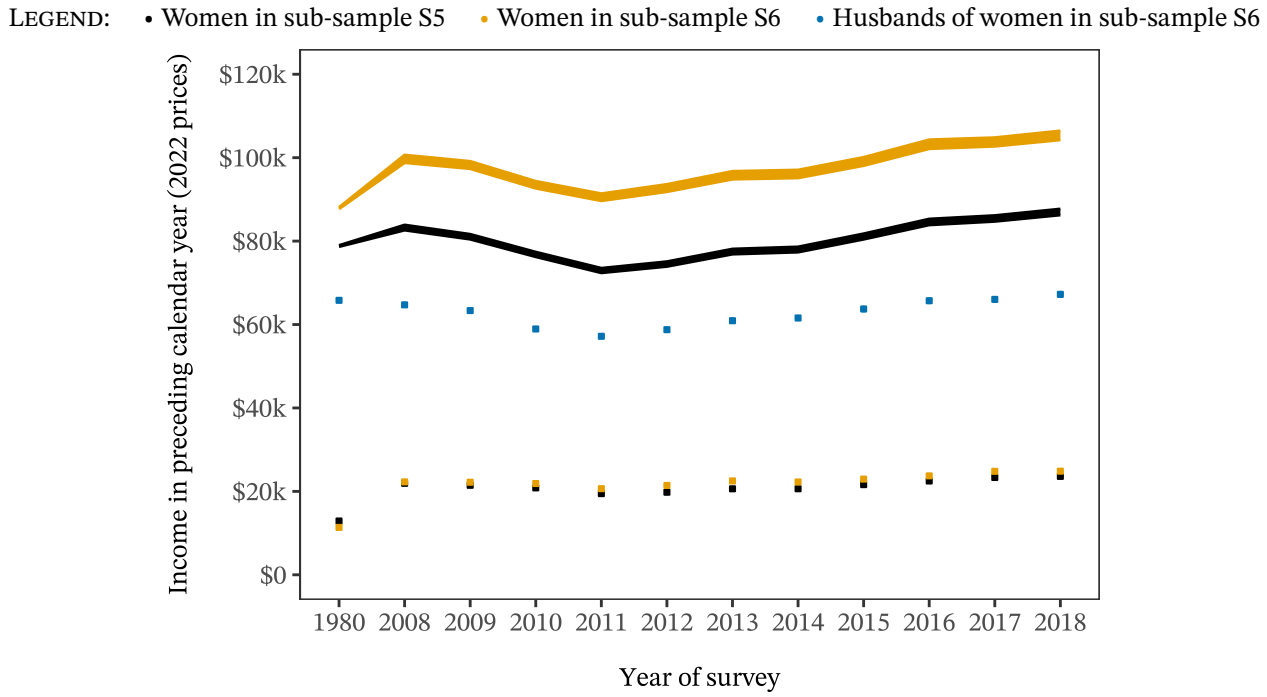


FIG 11. Wage income (points) and total family income (ribbons) amongst women and their husbands (if applicable) in sub-samples S5 and S6.

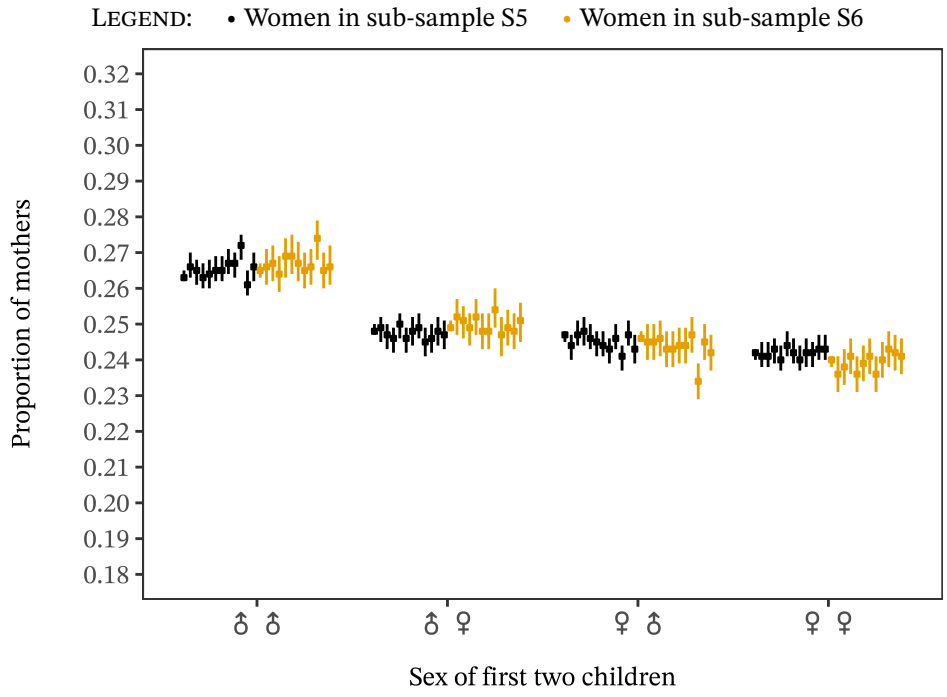


FIG 12. Fertility choice amongst women in sub-samples S5 and S6.

Funding

The author gratefully acknowledges financial support from the Economic and Social Research Council (ESRC studentship number 1329842; Application of minimally restrictive econometric models).

REFERENCES

Abrevaya, J., Hsu, Y.-C. and Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics* **33** 485–505.

Al-Khaja, A. J. A. (2016). Essays on Female Empowerment and Women's Status, PhD thesis, UCL (University College London).

Andrews, D. W. K. and Han, S. (2009). Invalidity of the bootstrap and the m out of n bootstrap for confidence interval endpoints defined by moment inequalities. *The Econometrics Journal* **12** S172-S199.

Andrews, D. W. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica* **78** 119–157.

Angrist, J. D. and Evans, W. N. (1998). Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *The American Economic Review* **88** 450–477.

Artstein, Z. (1983). Distributions of random sets and random selections. *Israel Journal of Mathematics* **46** 313–324.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* **92** 1171–1176.

Benny, L. (2021). Essays in Applied Labour Economics, PhD thesis, University of Essex.

Beresteanu, A., Molchanov, I. and Molinari, F. (2012). Partial identification using random set theory. *Journal of Econometrics* **166** 17–32.

Blau, F. D., Kahn, L. M., Brummund, P., Cook, J. and Larson-Koester, M. (2020). Is there still son preference in the United States? *Journal of Population Economics* **33** 709–750.

Blundell, R. and Powell, J. L. (2003). *Endogeneity in Nonparametric and Semiparametric Regression Models*. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress. Econometric Society Monographs* **2** 312–357. Cambridge University Press.

Bonet, B. (2001). Instrumentality Tests Revisited. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence* 48–55.

Bugni, F. A. (2010). Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica* **78** 735–753.

U. S. Census Bureau (2020). National Population by Characteristics: Annual Estimates of the Resident Population by Sex, Age, Race, and Hispanic Origin for the United States [NC-EST2023-ASR6H]. Accessed September, 2024.

Chalak, K. (2017). Instrumental variables methods with heterogeneity and mismeasured instruments. *Econometric Theory* **33** 69–104.

Chernozhukov, V., Lee, S. and Rosen, A. M. (2013). Intersection bounds: Estimation and inference. *Econometrica* **81** 667–737.

Chesher, A. and Rosen, A. M. (2013). What do instrumental variable models deliver with discrete dependent variables? *American Economic Review* **103** 557–562.

Chesher, A. and Rosen, A. M. (2017). Generalized instrumental variable models. *Econometrica* **85** 959–989.

Chesher, A. and Rosen, A. M. (2020). Generalized instrumental variable models, methods, and applications. In *Handbook of Econometrics*, **7** 1–110. Elsevier.

Chesher, A. and Rosen, A. M. (2021). Counterfactual worlds. *Annals of Economics and Statistics* **142** 311–335.

Frisch, R. (1995). *Autonomy of economic relations: Statistical versus theoretical relations in economic macrodynamics.* Cambridge university press Reprint of original presentation at the League of Nations, 1938. Editors: D.F. Hendry and M.S. Morgan.

Galichon, A. and Henry, M. (2011). Set identification in models with multiple equilibria. *The Review of Economic Studies* **78** 1264–1298.

Giacomini, R. and Kitagawa, T. (2021). Robust Bayesian inference for set-identified models. *Econometrica* **89** 1519–1556.

Gronau, R. (1977). Leisure, home production, and work–the theory of the allocation of time revisited. *Journal of political economy* **85** 1099–1123.

Gunsilius, F. (2019). A path-sampling method to partially identify causal effects in instrumental variable models. *arXiv:1910.09502*.

Heckman, J. J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica* **46** 931–959.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association* **81** 945–960.

Hurwicz, L. (1950). Generalisation of the Concept of Identification. In *Statistical inference in dynamic economic models* (T. C. Koopmans, ed.) John Wiley & Sons, Inc.

Iacovou, M. (2001). Fertility and female labour supply Technical Report, ISER Working Paper Series.

IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica* **62** 467–475.

IPUMS: MINNEAPOLIS, MN (2024). IPUMS USA: Version 15.0 [dataset]. Accessed September, 2024. Compiled and maintained by Steven Ruggles, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rogers and Megan Schouweiler.

JIANG, Z. and DING, P. (2020). Measurement errors in the binary instrumental variable model. *Biometrika* **107** 238–245.

KAIDO, H., MOLINARI, F. and STOYE, J. (2019). Confidence intervals for projections of partially identified parameters. *Econometrica* **87** 1397–1432.

KÉDAGNI, D. and MOURIFIÉ, I. (2020). Generalized instrumental inequalities: testing the instrumental variable independence assumption. *Biometrika* **107** 661–675.

KITAGAWA, T. (2012). Estimation and inference for set-identified parameters using posterior lower probability.

KITAGAWA, T. (2021). The identification region of the potential outcome distributions under instrument independence. *Journal of Econometrics* **225** 231–253.

KITAGAWA, T., MONTIEL OLEA, J. L., PAYNE, J. and VELEZ, A. (2020). Posterior distribution of nondifferentiable functions. *Journal of Econometrics* **217** 161-175.

KLINE, B. and TAMER, E. (2016). Bayesian inference in a class of partially identified models. *Quantitative Economics* **7** 329–366.

KNIGHT, F. H. (1921). Risk, uncertainty and profit. *Hart, Schaffner and Marx*.

LEWBEL, A. (2000). Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics* **97** 145-177.

MANSKI, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review* **80** 319–323.

MANSKI, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

MOLCHANOV, I. (2005). *Theory of Random Sets*. Springer.

MOON, H. R. and SCHORFHEIDE, F. (2012). Bayesian and frequentist inference in partially identified models. *Econometrica* **80** 755–782.

MOURIFIÉ, I. (2015). Sharp bounds on treatment effects in a binary triangular system. *Journal of Econometrics* **187** 74–81.

NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. *Statistical Science* 465–472. Translation from original Polish appearing in Roczniki Nauk Rolniczych (Annals of Agricultural Statistics), 1923. Editors: D. M. Dabrowska and T. P. Speed.

NORETS, A. and TANG, X. (2014). Semiparametric Inference in Dynamic Binary Choice Models. *Review of Economic Studies* **81**.

THE NEW YORK TIMES (2024). Care Policies Take Center Stage in Harris's Economic Message. Accessed September, 2024.

PEARL, J. (1995). On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* 435–443.

PEARL, J. (2009). *Causality*. Cambridge university press.

RICHARDSON, T. S. and ROBINS, J. M. (2014). ACE bounds; SEMs with equilibrium conditions. *Statistical Science* **29** 363–366.

RICHARDSON, T. S. and ROBINS, J. M. (2024). Assumptions and bounds in the instrumental variable model. *arXiv:2401.13758*.

ROWLEY, J. (2024). Econometric models of treatment with application to labour market outcomes, PhD thesis, UCL (University College London).

ROY, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford economic papers* **3** 135–146.

SHAIKH, A. M. and VYTLACIL, E. J. (2011). Partial identification in triangular systems of equations with binary dependent variables. *Econometrica* **79** 949–955.

STROTZ, R. H. and WOLD, H. O. (1960). Recursive vs. nonrecursive systems: An attempt at synthesis (part I of a triptych on causal chain systems). *Econometrica: Journal of the Econometric Society* 417–427.

VYTLACIL, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica* **70** 331–341.

WHITE, H. and CHALAK, K. (2009). Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning. *Journal of Machine Learning Research* **10**.

ZERMELO, E. (1904). Beweis, dass jede Menge wohlgeordnet werden kann: Aus einem an Herrn Hilbert gerichteten Briefe. *Mathematische Annalen* **59** 514–516.