Identification of Long-Term Treatment Effects via Temporal Links, Observational, and Experimental Data

Filip Obradović*

March 22, 2026

Abstract

Recent literature proposes combining short-term experimental and long-term observational data to provide alternatives to conventional observational studies for the identification of long-term average treatment effects (LTEs). This paper re-examines the identification problem and uncovers that assumptions restricting *temporal link functions* – relationships between short-term and mean long-term potential outcomes – are central in this context. The experimental data serve to *amplify* the identifying power of such assumptions; absent them, the combined data are no more informative than the observational data alone. Plausible inference thus hinges on justifiable restrictions in this class. Motivated by this, I introduce two *treatment response* assumptions that may be defensible based on economic theory or intuition. To utilize them and facilitate future developments, I develop a novel unifying identification framework that computationally produces sharp bounds on the LTE for a general class of temporal link function restrictions and accommodates imperfect experimental compliance – thereby also extending existing approaches. I illustrate the method by estimating the long-term effects of Head Start participation. The findings indicate that the effects on educational attainment, employment, and criminal involvement are lasting but smaller in magnitude than those established by sibling comparisons.

*UCLA, Department of Economics. Email: obradovicfilip@ucla.edu

I am deeply grateful to Charles Manski, Ivan Canay, and Federico Bugni for their guidance and support. I am also thankful to Nemanja Antić, Susan Athey, Eric Auerbach, Denis Chetverikov, Piotr Dworczak, Danil Fedchenko, Joel Horowitz, Matías Martinez, Jana Obradović, Andres Santos, Alexander Torgovitsky, and Gabriel Ziegler for valuable suggestions. I thank ICPSR for providing the data and the participants at the NBER Summer Institute, and seminars and conferences at Bocconi, Booth, Duke, Emory, Mannheim, Notre Dame, Michigan, Northwestern, Penn State, Princeton, UPenn, UCLA, USC, UVic, and Yale for comments. Financial support from the Robert Eisner Memorial Fellowship is gratefully acknowledged.

1 Introduction

Identifying long-term average treatment effects (LTEs) is an important goal in economics and other fields of science. For example, researchers may be interested in the effects of childhood interventions on earnings in adulthood; the impact of early-life conditional cash transfers on employment prospects; or the long-run adverse or protective effects of vaccination. LTEs are also often of interest in private sector research (Gupta et al. 2019).

Nevertheless, identifying LTEs is often challenging in practice. Long-term experimentation is frequently infeasible due to cost or institutional constraints.[1] Short-term experiments may be more accessible, but alone, they may not reveal outcomes of interest. As a result, researchers often rely on observational data (Currie and Almond 2011; Hoynes and Schanzenbach 2018). However, observational studies critically rely on identifying assumptions that may be challenging to justify.

This motivates a burgeoning strand of recent literature that seeks credible alternatives to conventional observational studies by combining (i) a long-run observational dataset with non-randomized treatment assignment and (ii) a short-run experimental dataset in which long-run outcomes are unobserved (Athey et al. 2025; Athey, Chetty, and Imbens 2025). However, follow-up work indicates that commonly-used modeling assumptions in this literature may also be challenging to justify in various economic settings, highlighting the need for alternative restrictions (Ghassami et al. 2022; Van Goffrier, Maystre, and Gilligan-Lee 2023; Imbens et al. 2024; Park and Sasaki 2024a).

This paper re-examines the identification problem by first characterizing which assumptions can exploit the specific data structure. For identification of the LTE, the experimental data serve only to potentially *amplify* the identifying power of a well-defined class of modeling assumptions, restrictions on *temporal link functions* – means of long-term potential outcomes conditional on short-term potential outcomes. Absent such restrictions, the combined data are necessarily equally informative about the LTE as the observational data alone. When restrictions on temporal link functions are imposed, however, combining data can provide additional identifying power. Therefore, plausible inference that leverages the experimental data hinges on imposing justifiable restrictions of this type.

Building on this characterization, I introduce novel modeling assumptions within this class that may be justified by economic theory or intuition. Both impose shape restrictions on temporal link functions without constraining the treatment selection in any of the datasets, and therefore constitute treatment response assumptions. The first stipulates that temporal link functions are monotonic. Intuitively, this means that, absent selection into treatment, average long-term

1. Institutions supporting RCTs in development economics frequently require phase-in designs with staggered rollout of treatment to the whole sample. This limits follow-up for the control group.

outcomes would be monotonic in short-term ones. Monotonicity of conditional means has been widely used in other settings since it may often be defensible on economic or intuitive grounds (see e.g. Manski and Pepper (2000) and Manski (2009), Mogstad, Santos, and Torgovitsky (2018), Torgovitsky (2019)). The second assumption postulates that the temporal link functions are invariant to treatment. Such invariance is implied by established mediation models, as in Heckman, Pinto, and Savelyev (2013) and García et al. (2020), or statistical surrogacy (Prentice (1989), Athey et al. (2025)).

Finally, the paper develops a unifying identification framework that accommodates: (i) a general class of restrictions on temporal link functions and (ii) imperfect compliance in the experiment. The framework delivers sharp bounds on the LTE computationally, by solving optimization problems in which the maintained restrictions enter as constraints. The general sharp characterization is obtained by extending arguments of Beresteanu, Molchanov, and Molinari (2012) and Chesher and Rosen (2017) to address a new technical challenge – jointly bounding distribution functions and conditional means of latent variables.

The optimization-based formulation of the identified set enables direct implementation of the proposed restrictions. Moreover, it facilitates the development of new restrictions on temporal link functions by eliminating the need to derive bounds algebraically or to prove sharpness on a case-by-case basis. The framework also nests existing point-identification results and extends them to allow for imperfect compliance in the experiment, which is of substantial practical relevance. Building on Shi and Shum (2015) and this characterization, I propose a consistent criterion-based estimator for the bounds.

I illustrate the method by estimating the long-term effects of Head Start participation, the largest federally funded early childhood education program in the United States. To do so, I combine data from the Head Start Impact Study, a short-term experiment, and the Child and Young Adult Supplement to the National Longitudinal Survey of Youth 1979 cohort, a longitudinal survey. I find evidence of beneficial program impacts on educational and labor-market outcomes, as well as criminal involvement in adulthood. Head Start is estimated to increase the probability of high school graduation by 1.9 to 3.2 percentage points (pp), and to decrease the probability of grade repetition by 1.1 to 5.3 pp. The program is also estimated to lower the probability of idleness (neither working nor in school) by 1.5 to 4.6 pp and criminal involvement by 1.2 to 4.0 pp. The results suggest that Head Start has lasting effects, though smaller in magnitude than reported by sibling comparison studies (Deming (2009)). More broadly, they illustrate that the proposed assumptions can yield informative estimated bounds in applied work.

This paper is related to several strands of literature. It contributes to the recent body of work that combines long-term observational and short-term experimental data to identify long-term treatment effects (see also García et al. (2020), Dynarski et al. (2021), Hu, Zhou, and Wu (2022), Chen and Ritzwoller (2023), Park and Sasaki (2024b), Aizer et al. (2024)), and to the broader literature on data combination (Cross and Manski (2002), Molinari and Peski (2006), Ridder and Moffitt (2007), Fan, Sherman, and Shum (2014), D'Haultfœuille, Gaillac, and Maurel (2024)). In doing so, it draws on and extends results from the partial identification literature based on random set theory (Galichon and Henry (2011), Beresteanu, Molchanov, and Molinari (2012), Molchanov and Molinari (2014), Chesher and Rosen (2017, 2020)). The paper also contributes to work on general identification frameworks that operationalize broad classes of assumptions via optimization (Mogstad, Santos, and Torgovitsky (2018), Torgovitsky (2019), Russell (2021), Kamat (2024)). Finally, it is related to research that combines experimental data with economic theory to identify parameters beyond what the experimental data alone reveal (Todd and Wolpin (2006), Attanasio, Meghir, and Santiago (2012), Todd and Wolpin (2023)).

Section 2 introduces the setting. Section 3 details the role of restrictions on temporal link functions and experimental data. Section 4 introduces new restrictions on temporal link functions. Section 5 develops the identification framework. Section 6 provides the empirical illustration. Section 7 concludes. Appendix A contains additional discussions; Appendix B proves the main results; the Supplemental Appendix discusses estimation and proves auxiliary results.

2 Setting and Basic Assumptions

I formalize the problem using the standard potential outcomes model. Let $Y(d) \in \mathcal{Y} \subseteq \mathbb{R}$ and $S(d) \in \mathcal{S} \subseteq \mathbb{R}^{d_s}$ denote the long-term and short-term potential outcomes under some binary treatment $d \in \{0,1\}$, respectively. Denote the realized treatment by $D \in \{0,1\}$. The observed outcomes are:

$$\begin{aligned} Y &= DY(1) + (1-D)Y(0) \\ S &= DS(1) + (1-D)S(0). \end{aligned} \tag{1}$$

Let $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ be a vector of observed covariates. Define the conditional *long-term* average treatment effect (CLTE) $\tau(x)$:

$$\tau(x) = E[Y(1) - Y(0)|X = x]. \tag{2}$$

The parameter of interest can be the CLTE itself or its weighted averages, such as the average *long-term* treatment effect (LTE) $E[\tau(X)]$. I focus on the former for generality, noting that it is sufficient for identification of the latter when the weights are identified or given.

Example 1. *(Head Start Participation)* In the empirical illustration, D denotes Head Start participation, $S(d)$ is a vector of potential cognitive test scores in childhood, and $Y(d)$ are potential outcomes in adulthood, such as high school degree status or earnings, for treatment d.

2.1 Observed Data

As in prior work, I maintain that the researcher observes: 1) a short-term experimental dataset; and 2) a long-term observational dataset.[2] The population is partitioned into two subpopulations that are randomly sampled to generate the two datasets. Let $G \in \{O, E\}$ denote the subpopulation indicator, where $G = O$ produces the observational and $G = E$ the experimental data.

Let $Z \in \mathcal{Z}$ denote an exogenous (i.e., randomly assigned) instrument in the experimental dataset that induces individuals into treatment. The identification analysis accommodates bounded $\mathcal{Z}$ with an arbitrary number of support points. Typically, $Z \in \{0, 1\}$, representing random assignment to treatment or control groups, which may differ from the realized treatment D. More generally, Z may have more than two support points, or even satisfy $Z \in [0, 1]$ as in Heckman and Vytlacil (1999). Since the binary case is predominant in practice, I adopt its terminology for expositional convenience. I refer to experiments with $P(D = Z|G = E) = 1$ as having perfect compliance, and to the remaining cases as exhibiting imperfect compliance.

The short-term experimental dataset reveals (S, D, X, Z), but not the long-term outcome Y. The long-term observational dataset reveals (Y, S, D, X), but contains no exogenous instrument Z. Note that the data structure does not support the direct use of well-established instrumental variable methods for identification of *long-term* treatment effects, since the outcome of interest Y is never observed together with an instrument Z (e.g. as in Imbens and Angrist (1994), Heckman and Vytlacil (1999) and Mogstad, Santos, and Torgovitsky (2018)).

Example 1 (continued). The observational dataset is the Child and Young Adult Supplement to the National Longitudinal Survey of Youth 79 Cohort (NLSY79) which reveals (Y, S, D). The experimental dataset is the Head Start Impact Study (HSIS) which reveals (S, D, Z), where $Z = 1$ if the individual is assigned to participation in Head Start and $Z = 0$ if assigned to non-participation. Puma et al. (2010) explain that some individuals may have $D \neq Z$.

Remark 1. Introducing Z in the experiment allows for (but does not require) imperfect compliance, which is of great practical relevance. In contrast, existing methods predominantly assume that D is randomly assigned, which need not hold under imperfect compliance. The identification framework in Section 5 nests these methods and extends them to allow this possibility.

2. This setting is increasingly common. See also García et al. (2020), Ghassami et al. (2022), Hu, Zhou, and Wu (2022), Van Goffrier, Maystre, and Gilligan-Lee (2023), Chen and Ritzwoller (2023), Park and Sasaki (2024a, 2024b), Aizer et al. (2024), Imbens et al. (2024), Athey, Chetty, and Imbens (2025).

I maintain the following assumptions throughout the paper.

Assumption RA. *(Random Assignment)* $Z \perp\!\!\!\perp (Y(d), S(d))|X, G = E$ *for* $d \in \{0, 1\}$.

Assumption EV. *(Experimental External Validity)* $G \perp\!\!\!\perp (Y(d), S(d))|X$ *for* $d \in \{0, 1\}$.

Assumption RA is standard in the program evaluation literature. It is satisfied if Z in the experimental data is randomly assigned, conditional on X. Assumption EV is a key assumption in the data combination literature, linking the two datasets (Ghassami et al. (2022), Chen and Ritzwoller (2023), Park and Sasaki (2024a, 2024b), Athey, Chetty, and Imbens (2025)). It states that, conditional on X, the subpopulations generating the datasets do not differ in their counterfactual distributions for each d. It may be plausible if the two datasets are representative of the same population, conditional on X. For example, in the empirical illustration, HSIS and NLSY79 are designed to be representative of the U.S. population and pertain to the same treatment. More broadly, a growing body of empirical work relies on related links across datasets in this context; see, for example, García et al. (2020), Dynarski et al. (2021), Hu, Zhou, and Wu (2022), Aizer et al. (2024).

It is worth emphasizing what is *not* assumed. Maintained assumptions allow for $D \not\perp\!\!\!\perp (Y(d), S(d))|X, G = g$ for any $g \in \{O, E\}$ and $d \in \{0, 1\}$. This is expected in the observational dataset, and in the experimental data when compliance is imperfect. The assumptions also do not require $P(D = 1|G = g) \in (0, 1)$ for any $g \in \{O, E\}$. Instead, they allow $P(D = 1|G = g) \in [0, 1]$, which is relevant when a treatment is available only in one dataset, typically in the experiment. This is the case with some "model" early childhood intervention programs or novel vaccines.

Under Assumption EV, CLTE is invariant to G, $E[Y(1) - Y(0)|X = x, G] = E[Y(1) - Y(0)|X = x] = \tau(x)$. Henceforth, I keep conditioning on X implicit. The following analysis should be understood as conditional-on-X; I write the parameter of interest $\tau(x)$ as:

$$\tau = E[Y(1) - Y(0)] \tag{3}$$

and I continue referring to it as the LTE, with the understanding that it represents the CLTE.

Remark 2. Quasi-experimental datasets with Z satisfying Assumptions RA and EV may also serve as the experimental dataset. I continue to refer to such data as experimental, following prevailing terminology in the literature.

Notation: $\mathcal{H}(\theta)$ denotes the identified set for a parameter θ. I write laws conditional on an event $\mathcal{E}$, $P(\cdot|\mathcal{E}, G = g)$, as $P_g(\cdot|\mathcal{E})$ for $g \in \{O, E\}$. Whenever $P_E(\cdot|\mathcal{E}) = P_O(\cdot|\mathcal{E})$, I omit the subscript g. This is inherited by their features, e.g. $E_g[\cdot|\mathcal{E}] := E[\cdot|\mathcal{E}, G = g]$. If necessary, I specify the random element using subscripts (e.g. $P_{S(d)}$ is the law of $S(d)$).

3 Role of Experimental Data

This section uncovers the role played by the experimental data in identifying τ. Rather than removing the need for additional assumptions, experimental data serve to potentially *amplify* the identifying power of a well-defined class of modeling restrictions. Section 4 leverages this insight to propose assumptions within this class that may be justifiable based on economic theory or intuition. Section 5 develops a general identification framework that enables tractable implementation of various assumptions in the class. Define for $s \in \mathcal{S}$ and $d \in \{0, 1\}$:

$$m_d(s) := E[Y(d)|S(d) = s], \quad \gamma_d := P_{S(d)}. \tag{4}$$

where $P_{S(d)}$ is the marginal law of $S(d)$. I refer to $m_0(s)$ and $m_1(s)$ as *temporal link functions*, since they "link" the short-term and long-term potential outcomes. Collect the two functions $m := (m_0, m_1) \in \mathcal{M}$ and the marginal laws $\gamma := (\gamma_0, \gamma_1) = (P_{S(0)}, P_{S(1)})$.[3] These objects are directly related to the parameter of interest:

$$\tau = E[Y(1) - Y(0)] = \int_{\mathcal{S}} m_1(s) d\gamma_1(s) - \int_{\mathcal{S}} m_0(s) d\gamma_0(s). \tag{5}$$

Consider the class of modeling assumptions defined by the following generic restriction.

Assumption MA. *(Modeling Assumption) $m \in \mathcal{M}^A \subseteq \mathcal{M}$ for a known or identified set $\mathcal{M}^A$.*

Let $\mathcal{H}(\tau)$ be the identified set for τ, that is, the set of all values of τ compatible with both datasets and Assumptions RA, EV, and MA. Note that Assumption MA subsumes the trivial case $\mathcal{M}^A = \mathcal{M}$, corresponding to the absence of additional modeling assumptions. Denote by $\mathcal{H}^O(\tau)$ the identified set for τ under the same assumptions when *only observational data* are used, and by $\subsetneq$ a *strict* subset. The following proposition elucidates the relationship between $\mathcal{H}(\tau)$, $\mathcal{H}^O(\tau)$, and the modeling assumptions represented by $\mathcal{M}^A$.

Proposition 1. *Let Assumptions RA, EV hold, and suppose $\mathcal{H}(\tau) \neq \emptyset$. If $\mathcal{H}(\tau) \subsetneq \mathcal{H}^O(\tau)$, then $\mathcal{M}^A \subsetneq \mathcal{M}$. Equivalently, if $\mathcal{M}^A = \mathcal{M}$, then $\mathcal{H}(\tau) = \mathcal{H}^O(\tau)$.*

Proposition 1 shows that it is *necessary* to restrict m by assumption in order to gain identifying power from the experimental data. Equivalently, when the temporal link functions m are left unrestricted, incorporating the experimental data provides *no additional identifying power* for τ. Once such restrictions are imposed, the inclusion of experimental data may yield additional

3. Let $(\Omega, \mathcal{F}, P)$ denote the probability space and $\mathcal{B}$ the Borel $\sigma-$algebra. $\mathcal{M}$ is the set of Borel-measurable functions $\mu : \mathcal{S} \times \mathcal{S} \to \mathcal{Y} \times \mathcal{Y}$ such that $\mu \circ \varsigma$ is P-integrable for some $\mathcal{F}/\mathcal{B}(\mathcal{S} \times \mathcal{S})$-measurable function $\varsigma : \Omega \to \mathcal{S} \times \mathcal{S}$.

identifying power, enabling $\mathcal{H}(\tau) \subsetneq \mathcal{H}^O(\tau)$. The proposition thus isolates the class of assumptions that could be made more informative about τ by the addition of experimental data. In this sense, adding experimental data *may amplify* the identifying power of restrictions on m.

To gain intuition for the result, note that because $S(d)$ is revealed in the observational data whenever $D = d$, the experimental data can *only* provide additional information about the distribution of $S(d)$ for those individuals with $D \neq d$. For these individuals, however, the data impose no restrictions on the relationship between $S(d)$ and the mean $Y(d)$. When this relationship is left unrestricted, additional information about the distribution of $S(d)$ does not translate into additional information about the average $Y(d)$, and thus about τ. In contrast, when the relationship between $S(d)$ and the average $Y(d)$ is restricted by assumption, additional information about the distribution of $S(d)$ can yield further information about the mean of $Y(d)$ and hence about τ.

Remark 3. Existing approaches typically do not impose explicit restrictions on m. However, their identification arguments ultimately reduce to such restrictions implied by the maintained assumptions. These point-identification results are thus nested within the identification framework here. For example, Athey, Chetty, and Imbens (2025) maintain *latent unconfoundedness (LUC)* – $Y(d) \perp\!\!\!\perp D|S(d), G = O$ for $d \in \{0, 1\}$, while the identification result uses its implication $m_d(s) = E_O[Y|S = s, D = d]$ for $d \in \{0, 1\}$ and $s \in \mathcal{S}$. For Imbens et al. (2024, Theorem 1), let S_t be subvectors of S for $t \in \{1, 2, 3\}$. Maintained assumptions imply the restriction used for identification: $m_d(s_3, s_2) = h(s_3, s_2, d)$ where h solves $E_O[Y|S_2, S_1, D] = E_O[h(S_3, S_2, D)|S_2, S_1, D]$.

Two points are worth emphasizing. First, the addition of experimental data need not necessarily amplify the identifying power of restrictions on m. In particular, it is possible that $\mathcal{M}^A \subsetneq \mathcal{M}$ while $\mathcal{H}(\tau) = \mathcal{H}^O(\tau)$. Second, restrictions on m may have identifying power for τ even in the absence of experimental data. Whether either case arises depends on the underlying data distribution and on the specific restriction imposed. These observations highlight that restrictions on m are *central* for identifying τ in this setting, while the experimental data play an auxiliary role by potentially amplifying their identifying power.

Proposition S.1 in Appendix S.2 illustrates these points using the latent unconfoundedness assumption defined in Remark 3, and frequently invoked in this context (see, for example, Hu, Zhou, and Wu (2022), Park and Sasaki (2024b), and Aizer et al. (2024)). Specifically, the proposition shows that if the data distribution is such that $m_d(s) = E_O[Y \mid S = s, D = d]$ are constant in s, both the combined and observational data alone point identify τ under LUC. In this case, $\mathcal{M}^A \subsetneq \mathcal{M}$ but $\mathcal{H}(\tau) = \mathcal{H}^O(\tau)$, so the experimental data do not amplify the identifying power of LUC. The proposition further shows that for a broad class of empirically relevant data distributions, the observational-data bound $\mathcal{H}^O(\tau)$ under LUC is strictly more informative than the worst-case bounds of Manski (1990) that would be sharp absent LUC.

4 Modeling Assumptions

For the purpose of identifying τ, the experimental data serve only to potentially amplify the identifying power of restrictions imposed on m. Consequently, plausible inference hinges on the plausibility of the restrictions within this class, despite the use of experimental data. However, commonly used assumptions implying restrictions on m may be challenging to interpret or justify in economically relevant settings (see e.g. Ghassami et al. (2022), Van Goffrier, Maystre, and Gilligan-Lee (2023), Imbens et al. (2024) and Park and Sasaki (2024a)). Hence, I explore alternative restrictions within this class that may be defensible based on economic theory or intuition.

Assumption LIV. *(Latent Monotone Instrumental Variables) For every $m \in \mathcal{M}^A$ and each $d \in \{0, 1\}$, $m_d(s)$ is nondecreasing in s with respect to the product order. That is, for all $s, s' \in \mathcal{S}$ such that $s \leq s'$ (componentwise): $m_d(s) \leq m_d(s')$.*

Assumption LIV states that the conditional means of $Y(d)$ are nondecreasing in any individual short-term potential outcome $S(d)$.[4] Intuitively, this would be plausible if researchers are willing to maintain that long-term outcomes would be monotonic in short-term ones, *absent selection into treatment.*

Example 2. *(LIV and Head Start)* Suppose first that $S(d)$ consists of a single childhood cognitive test score. Assumption LIV then states that people with a higher *potential* test score $S(d)$ have a weakly higher average of *potential* high school degree status in adulthood $Y(d)$ than people with a lower *potential* test score. In other words, under an exogenously fixed treatment, high school graduation *rates* are nondecreasing in the childhood test score. When $S(d)$ comprises multiple test scores, the assumption asserts that the high school graduation rates are nondecreasing in each individual test score, holding the remaining ones constant.

LIV is related to the monotone instrumental variable (MIV) assumption of Manski and Pepper (2000, 2009). MIV posits the existence of a variable $V \in \mathcal{V}$ such that $E[Y(d)|V = v]$ is nondecreasing in $v \in \mathcal{V}$, where V is observed for *all* individuals. The critical distinction is that the conditioning variables in Assumption LIV are latent counterfactuals. This feature introduces additional complexity, which is addressed by the identification framework.

4. One can also assume that $m_d(s)$ is nonincreasing for any sub-collection of elements $\{s_j\}_{j \in \mathcal{J}}$ of s. Results follow directly by defining $\tilde{S}(d)$ with components $\tilde{S}_j(d) = \mathbb{1}[j \in \mathcal{J}]\,(-S_j(d)) + \mathbb{1}[j \notin \mathcal{J}]S_j(d)$ and observing that $E[Y(d)|\tilde{S}(d) = s]$ satisfies LIV.

Assumption TI. *(Treatment Invariance) For all* $m \in \mathcal{M}^A$ *and* $s \in \mathcal{S}$, $m_1(s) = m_0(s)$.

The assumption states that the relationship between the *potential* short-term outcomes $S(d)$ and the mean of the long-term *potential* outcomes $Y(d)$ does not vary with the underlying treatment d. Intuitively, it means that the treatment affects *average* long-term outcomes only through short-term ones. Importantly, TI follows from theoretical models previously used in empirical research, as outlined by the following example. It is also implied by the statistical surrogacy assumption of Prentice (1989) $Y \perp\!\!\!\perp D|S, G = E$ in the special case of perfect compliance.[5]

Example 3. *(TI and Head Start)* Consider the following separable model:

$$Y(d) = \phi_d(S(d)) + \varepsilon_d = \phi(S(d)) + \varepsilon_d, \;\; \varepsilon_d \sim \varepsilon, \;\; \varepsilon_{d'} \perp\!\!\!\perp S(d), \forall d, d' \in \{0, 1\} \tag{6}$$

where $S(d)$ is a vector of short-term potential outcomes, including test scores and measures of non-cognitive skills. $S(d)$ are inputs in the production function ϕ_d for $Y(d)$. The production function ϕ_d and the distributions of unobservables ε_d do not depend on Head Start participation d. Therefore, $E[Y(d)|S(d) = s] = \phi(s) + E[\varepsilon]$ which is invariant to d, so TI is implied by the model. Researchers may thus use TI whenever they find the model plausible. García et al. (2020) argue the plausibility of such a model in the context of identifying the long-term effects of an early childhood program.

Note that neither of the two assumptions imposes restrictions on selection, i.e., how individuals choose their treatment. As such, they represent *treatment response assumptions*, whereas related preceding work predominantly relies on selection assumptions.

5 Identification Framework

The preceding sections argue that restrictions on temporal link functions m are central, and that experimental data can make them more informative. Here, I introduce a novel framework that enables the identification of τ under a general class of restrictions on m.

Define the functional $T : \mathcal{M} \times \mathcal{P}^{\mathcal{S}} \times \mathcal{P}^{\mathcal{S}} \to \bar{\mathbb{R}}$, where $\mathcal{P}^{\mathcal{S}}$ is the set of distribution functions supported on $\mathcal{S}$:

$$T(m, \gamma) = \int_{\mathcal{S}} m_1(s) d\gamma_1(s) - \int_{\mathcal{S}} m_0(s) d\gamma_0(s). \tag{7}$$

5. See Lemma S.8 *i) in Appendix S.2*. An important distinction relative to work relying on the surrogacy assumption in this literature is that the comparability assumption $G \perp\!\!\!\perp Y|S$ is *not* imposed here (e.g. see Athey et al. (2025) and Chen and Ritzwoller (2023)). Together, comparability and surrogacy imply a selection assumption on m under Assumption EV; see Lemma S.8 *vi)*.

Recalling that $E[Y(d)] = \int_{\mathcal{S}} m_d(s) d\gamma_d(s)$ by (5), T produces the corresponding value of τ for a given (m, γ). One can then incorporate assumptions on m to identify τ in two steps. First, find all (m, γ) consistent with the data, Assumptions RA, EV, and any restriction MA on m. This yields their identified set $\mathcal{H}(m, \gamma)$. Then, the identified set $\mathcal{H}(\tau)$ is the set of all values of τ consistent with the feasible (m, γ), or:

$$\mathcal{H}(\tau) := \Big\{ T(m, \gamma) : (m, \gamma) \in \mathcal{H}(m, \gamma) \Big\}. \tag{8}$$

5.1 Identifying (m, γ)

This section shows how to represent all information about (m, γ). (m, γ) are features of *latent* random variables $Y(d)$ and $S(d)$. I exploit this fact to construct $\mathcal{H}(m, \gamma)$ for a large class of restrictions on m.

Because at least some potential outcomes are unobserved for each unit, the data and maintained assumptions are consistent with a set of *random vectors* $(S(0), S(1), Y(0), Y(1))$. Intuitively, let $\mathcal{Q}$ be the set of all such random vectors that are compatible with the data, and Assumptions RA and EV. To obtain $\mathcal{H}(m, \gamma)$, it suffices to collect the corresponding pairs (m, γ) that additionally satisfy the modeling restriction $m \in \mathcal{M}^A$. By definition:

$$\mathcal{H}(m, \gamma) = \left\{ (m, \gamma) : \begin{array}{c} \overbrace{m \in \mathcal{M}^A,}^{\textit{Modeling assumption}} \quad \overbrace{\exists (S(0), S(1), Y(0), Y(1)) \in \mathcal{Q}}^{\textit{Data + Assumptions RA/EV}} \\ \underbrace{\forall d \in \{0,1\} : \gamma_d \overset{d}{=} S(d),\, m_d(S(d)) = E[Y(d)|S(d)] \text{ a.s.}}_{(m,\gamma) \textit{ correspond to } S(d) \textit{ and } Y(d)} \end{array} \right\}. \tag{9}$$

Random set theory provides a convenient way to deliver a *sharp* characterization of (m, γ) based on this intuition. It first allows one to characterize $\mathcal{Q}$ and, consequently, $\mathcal{H}(m, \gamma)$. It then yields an *equivalent* representation of (9) via moment inequalities that exhaust *all* information contained in the data and the maintained assumptions. To formalize the argument, I introduce the necessary basic definitions specialized to finite-dimensional Euclidean spaces. I henceforth maintain that all random elements are defined on a non-atomic probability space $(\Omega, \mathcal{F}, P)$.[6]

Notation: B and K denote sets. $\mathcal{K}(B)$, and $\mathcal{C}(B)$ denote the families of all compact, and closed subsets of the set B, respectively. $co(B)$ is the closed convex hull of the set B.

Definition 1. A measurable map $\mathbf{R} : \Omega \to \mathcal{C}(\mathbb{R}^{d_R})$ is called a *random (closed) set*.[7]

6. That is, for any $A \in \mathcal{F}$ with positive measure there exists a measurable $B \subset A$ such that $0 < P(B) < P(A)$.
7. $\mathbf{R}$ is measurable if for every compact set $K \in \mathcal{K}(\mathbb{R}^{d_R})$: $\{\omega \in \Omega : \mathbf{R}(\omega) \cap K \neq \emptyset\} \in \mathcal{F}$. The codomain $\mathcal{C}(\mathbb{R}^{d_R})$ is equipped with the $\sigma-$algebra generated by the families of sets $\{B \in \mathcal{C}(\mathbb{R}^{d_R}) : B \cap K \neq \emptyset\}$ over $K \in \mathcal{K}(\mathbb{R}^{d_R})$.

Definition 2. A random vector $R : \Omega \to \mathbb{R}^{d_R}$ such that $R \in \mathbf{R}$ a.s. is called a *(measurable) selection* of $\mathbf{R}$. $Sel(\mathbf{R})$ and $Sel^1(\mathbf{R})$ are the sets of all selections, and all integrable selections of $\mathbf{R}$, respectively.

Define the following closed random sets for $d \in \{0,1\}$:

$$\mathbf{Y}_d := \begin{cases} \{Y\}, \text{ if } (D,G) = (d,O) \\ \mathcal{Y}, \text{ otherwise} \end{cases}, \quad \mathbf{S}_d := \begin{cases} \{S\}, \text{ if } (D,G) \in \{(d,E),(d,O)\} \\ \mathcal{S}, \text{ otherwise} \end{cases}. \tag{10}$$

By construction, random sets $\mathbf{Y}_d$ and $\mathbf{S}_d$ summarize all information about $Y(d)$ and $S(d)$ contained in *the data*, respectively. As Beresteanu, Molchanov, and Molinari (2012) explain, *all* information in the data about $(S(0), S(1), Y(0), Y(1))$ can be represented by stating that they can be any selection of the corresponding random sets. Additional assumptions, such as Assumptions RA and EV, may be imposed by further restricting the admissible selections, which are represented by the set $\mathcal{Q}$ in (9). The following lemma formalizes $\mathcal{Q}$ and uses it to equivalently characterize $\mathcal{H}(m,\gamma)$.

Lemma 1. *Let Assumptions RA, EV, and MA hold. The identified set for (m,γ) is:*

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\}, \ \exists S(d) \in Sel(\mathbf{S}_d) \cap I, \\ \exists Y(d) \in Sel(\mathbf{Y}_d), \ \gamma_d \overset{d}{=} S(d), \ m_d(S(d)) = E_O[Y(d)|S(d)] \ a.s. \end{array} \right\}. \tag{11}$$

where I is the set of random elements $E_1 \in \mathcal{S}$ such that $E_1 \perp\!\!\!\perp G$ and $E_1 \perp\!\!\!\perp Z|G = E$.

The lemma exploits the fact that Y is observed only in the observational sample to eliminate restrictions on (m,γ) imposed by Assumptions RA and EV that are redundant in this setting. In particular, the independence restrictions that remain relevant are $S(d) \perp\!\!\!\perp G$ and $S(d) \perp\!\!\!\perp Z|G = E$. Moreover, only observational data impose restrictions on m directly, as reflected by $m_d(S(d)) = E_O[Y(d)|S(d)]$. These observations enable the following sharp characterization of $\mathcal{H}(m,\gamma)$ via moment inequalities identified by the data.

Theorem 1. *Let Assumptions RA, EV and MA hold. Suppose that $E[|Y(d)|] < \infty$ for $d \in \{0,1\}$. The identified set for (m,γ) is:*

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\}, \forall B \in \mathcal{C}(\mathcal{S}), \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ \forall u \in \{-1,1\}: \ um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s)) \ \gamma_d\text{--}a.e. \end{array} \right\} \tag{12}$$

where $h_{co(\mathcal{Y})}(u) := \sup_{y \in co(\mathcal{Y})} uy$, $\mu_d(s) := E_O[Y|S = s, D = d]$, and $\pi_{\gamma_d} := dP_O(S, D = d)/d\gamma_d$.

If a collection of sets $\mathfrak{C}$ is a core determining class for the containment functional of $\mathbf{S}_d$, then the condition $\forall B \in \mathcal{C}(\mathcal{S})$ can be replaced with $\forall B \in \mathfrak{C}$.

The argument combines Artstein's theorem (Artstein (1983, Theorem 2.1)) with the conditional Aumann expectation, tools that are typically applied separately, to address a novel technical challenge: jointly bounding (i) the distribution function of a selection of a random set and (ii) a conditional mean whose conditioning σ-algebra is generated by that selection.

To build intuition for the bounds, consider conditions that are necessary for (m, γ) to be compatible with the data and maintained assumptions. The proof shows that these are also sufficient, and hence deliver sharp bounds. First, recalling $\gamma_d := P_{S(d)}$, it is immediate that for any closed subset $B \in \mathcal{C}(\mathcal{S})$: $\gamma_d(B) \geq P(S(d) \in B, D = d)$. By Assumptions RA and EV, $S(d) \perp\!\!\!\perp G$ and $S(d) \perp\!\!\!\perp Z|G = E$. It is therefore also necessary that each γ_d satisfies:

$$\forall B \in \mathcal{C}(\mathcal{S}): \;\; \gamma_d(B) \geq P_O(S(d) \in B, D = d) \quad \text{and} \quad \gamma_d(B) \geq P_E(S(d) \in B, D = d|Z) \text{ a.s.} \tag{13}$$

which coincides with the restrictions on γ_d in the theorem. This represents an intersection bound on γ_d reflecting that $S(d)$ is observed in both datasets when $D = d$.

Second, recalling that under Assumption EV $m_d(S(d)) = E_O[Y(d)|S(d)]$, by the law of iterated expectations:

$$m_d(S(d)) = E_O[Y|D = d, S(d)]P_O(D = d|S(d)) + E_O[Y(d)|D \neq d, S(d)]P_O(D \neq d|S(d)) \text{ a.s.}$$

Let $\pi_{\gamma_d}(s) := P_O(D = d|S(d) = s)$ denote the latent propensity score for a given $\gamma_d \overset{d}{=} S(d)$. Since $S(d) = S$ whenever $D = d$ and $S(d) \perp\!\!\!\perp G$, π_{γ_d} is the Radon-Nikodym derivative such that $\pi_{\gamma_d}(S(d)) = \frac{dP_O(S, D=d)}{d\gamma_d}$. Because $E_O[Y(d)|D \neq d, S(d)] \in \mathcal{Y}$ is never observed, the usual worst-case bounds arguments imply:

$$\begin{aligned} m_d(s) &\geq E_O[Y|D = d, S = s]\pi_{\gamma_d}(s) + \inf \mathcal{Y}(1 - \pi_{\gamma_d}(s)) \;\; \gamma_d\text{−a.e.} \\ m_d(s) &\leq E_O[Y|D = d, S = s]\pi_{\gamma_d}(s) + \sup \mathcal{Y}(1 - \pi_{\gamma_d}(s)) \;\; \gamma_d\text{−a.e.} \end{aligned} \tag{14}$$

which is identical to the restrictions on m_d in the theorem. Accordingly, these conditions represent worst-case bounds on m_d for a *given* γ_d. Notably, the bounds on m_d may change with the particular γ_d under consideration.

Finally, Theorem 1 also enables the use of a *core-determining class* (CDC) to remove redundant constraints on γ without loss of information. Intuitively, a CDC does so by eliminating constraints that are implied by the remaining ones.

5.2 Characterizing $\mathcal{H}(\tau)$

The sharp identified set $\mathcal{H}(\tau)$ follows as the image of the functional T over the set of possible (m, γ). The next result shows that under easily verifiable high-level conditions on $\mathcal{M}^A$, $\mathcal{H}(\tau)$ is an interval which can be characterized by solving two optimization problems.

Theorem 2. *Let Assumptions RA, EV, and MA hold. Suppose that $E[|Y(d)|] < \infty$ for $d \in \{0, 1\}$ and that $\mathcal{M}^A$ is closed and convex. Then the closure of $\mathcal{H}(\tau)$ is:*

$$\left[\inf_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m}, \tilde{\gamma}), \sup_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m}, \tilde{\gamma}) \right]. \tag{15}$$

Using optimization problems to characterize identified sets has become common in partial identification. Such representations typically follow from linear objectives and polyhedral, hence convex, constraint sets. Theorem 2 requires a different argument since T is bilinear – and thus separately continuous – and the constraint set is bilinearly constrained and therefore need not be convex. The proof shows that $\mathcal{H}(m, \gamma)$ is nevertheless convex under the assumptions of the theorem and that T is continuous along line segments in $\mathcal{H}(m, \gamma)$. Then $\mathcal{H}(\tau)$ is a line-continuous image of a convex set, hence a connected set in $\mathbb{R}$, i.e., an interval.

The restrictions $\mathcal{M}^A$ enter as constraints in the optimization problems. This offers three advantages. First, it enables the implementation of LIV and TI. Second, it nests existing point-identification results and generalizes them by allowing for imperfect compliance (for examples see Remark 3). Third, it provides a tool that facilitates development of new restrictions on m tailored for specific empirical settings by computationally producing the corresponding sharp bounds for τ. This removes the need to derive closed-form expressions and to prove sharpness for each set of assumptions.

Remark 4. Assumptions LIV and TI are representable via linear equality and inequality restrictions on m. Therefore, each resulting $\mathcal{M}^A$ is an intersection of (possibly infinitely many) affine subspaces and halfspaces in a linear space of functions on $\mathcal{S}$, and is thus convex. Moreover, because the restrictions are imposed pointwise (or γ_d−a.e.) and are preserved under limits, the corresponding $\mathcal{M}^A$ are closed. Furthermore, whenever m is identified by the data, such as under existing assumptions in Remark 3, $\mathcal{M}^A$ is a singleton (γ_d−a.e.) and hence closed and convex.

5.3 Implementation and Estimation

This section discusses how the optimization problems can be used to tractably characterize and estimate $\mathcal{H}(\tau)$. The problems become finite-dimensional when $\mathcal{S}$ is a finite set, as shown by the following corollary.

Corollary 1. *Suppose that assumptions of Theorem 2 hold and that $k := |\mathcal{S}| < \infty$. Let $\Delta(k)$ denote the $k-$dimensional simplex. Then the closure of $\mathcal{H}(\tau)$ is:*

$$\left[\inf_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}), \sup_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma})\right]. \tag{16}$$

where:

$$\mathcal{H}(m,\gamma) = \left\{\begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s, D=d|Z), P_O(S=s, D=d)\right), \\ (m_d(s) - \inf \mathcal{Y})\, \gamma_d(s) \geq (E_O[Y|S=s, D=d] - \inf \mathcal{Y})\, P_O(S=s, D=d)\ , \\ (\sup \mathcal{Y} - m_d(s))\, \gamma_d(s) \geq (\sup \mathcal{Y} - E_O[Y|S=s, D=d])\, P_O(S=s, D=d) \end{array}\right\} \tag{17}$$

When short-term outcomes have finite support, the infinite-dimensional problems in Theorem 2 reduce to finite-dimensional *generalized bilinear* optimization programs (Al-Khayyal (1992), for other examples of bilinear programs see Dutz et al. (2021) and Shea (2022)). Although such problems are generally nonconvex, modern general-purpose solvers can compute globally optimal solutions via spatial branch-and-bound methods (e.g. Gurobi Optimization (2024)). Section S.1 provides a consistent criterion-based estimator based on the corollary.

Focusing on cases where relevant variables are finitely supported or discretized has been the predominant approach in work that relies on Artstein's theorem (Galichon and Henry (2011), Russell (2021), Luo, Ponomarev, and Wang (2024)). Discretization is also common in related settings (Rambachan, Singh, and Viviano (2024), Park and Sasaki (2024b)). Appendix A.1 discusses the interpretation of results when S is discretized. General-purpose methods for solving bilinear infinite-dimensional programs remain an interesting open problem. In special cases, ongoing work shows that additional structure permits tractable implementation via optimal transport theory with approximation guarantees even when $\mathcal{S}$ is infinite; see Remark 5.

A few details are worth highlighting. First, $\mathcal{Y}$ is unrestricted. When it is unbounded, then the m_d constraints involving $\mathcal{Y}$ are trivially satisfied. Second, to reduce computational complexity, the corollary uses $\{\{s\} : s \in \mathcal{S}\}$ as a CDC. Table 1 illustrates the resulting reduction as a function of $|\mathcal{S}|$.[8] If $S(d)$ represents percentiles, then not using the CDC already results in a prohibitively complex constraint set.

Remark 5. Problems $\sup / \inf_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma})$ become linear in certain cases. This substan-

8. The number of constraints on m imposed by the data given γ is $4|\mathcal{S}|$; the total number depends on the modeling assumptions. The number of constraints may potentially be reduced further by adapting methods such as Luo, Ponomarev, and Wang (2024), as $\{\{s\} : s \in \mathcal{S}\}$ is not necessarily the smallest CDC.

Table 1: Number of constraints on γ in $\mathcal{H}(m, \gamma)$.

Constraint # for γ	$\lvert\mathcal{S}\rvert$				
	2	5	10	20	100
Without CDC	8	64	2048	2,097,152	$> 10^{30}$
With CDC $\{\{s\} : s \in \mathcal{S}\}$	2	8	18	38	198

tially simplifies computation and occurs under either of the following two conditions. First, if $\mathcal{H}(m, \gamma) = \{m\} \times \mathcal{H}(\gamma)$, the problems reduce to linear programs over γ. For example, assumptions that point-identify m independently of γ, such as LUC in Remark 3, yield $\mathcal{H}(m, \gamma) = \{m\} \times \mathcal{H}(\gamma)$. In this case, one may use optimal transport theory to characterize and estimate the identified set even when $\mathcal{S}$ is infinite (Voronin 2025). Second, if $\mathcal{H}(m, \gamma) = \mathcal{H}(m) \times \{\gamma\}$ and $\mathcal{M}^A$ is representable by linear constraints, the problems reduce to linear programs over m. This case arises under Assumptions LIV and TI when the right-hand sides of the constraints for each γ_d sum to one, as under perfect compliance or when the datasets jointly point-identify γ.

5.3.1 Identifying Power

Closed-form bounds, when available, can provide intuition about the sources of identifying power. Optimization-based bounds typically trade off this transparency for the ability to computationally deliver sharp bounds under a wide range of assumptions (see, e.g., Mogstad, Santos, and Torgovitsky (2018), Torgovitsky (2019)). Under Assumptions LIV and TI, however, one can still shed some light on the sources of identifying power by characterizing m^{UB} and m^{LB} that attain the upper and lower bounds on τ, given a feasible γ. To simplify exposition, focus on the bounds for $E[Y(1)]$ for a fixed feasible γ, as the intuition for $E[Y(0)]$ (and hence τ) is analogous.[9] Based on Lemma S.9 in Appendix S.2 and letting $s' \leq s$ be the product order, under Assumption LIV:

$$\begin{aligned} m_1^{LB}(s) &= \sup_{s' \leq s} \left[E_O[Y|S = s', D = 1] \pi_{\gamma_1}(s') + \inf \mathcal{Y} \left(1 - \pi_{\gamma_1}(s')\right) \right], \\ m_1^{UB}(s) &= \inf_{s' \geq s} \left[E_O[Y|S = s', D = 1] \pi_{\gamma_1}(s') + \sup \mathcal{Y} \left(1 - \pi_{\gamma_1}(s')\right) \right]. \end{aligned} \tag{18}$$

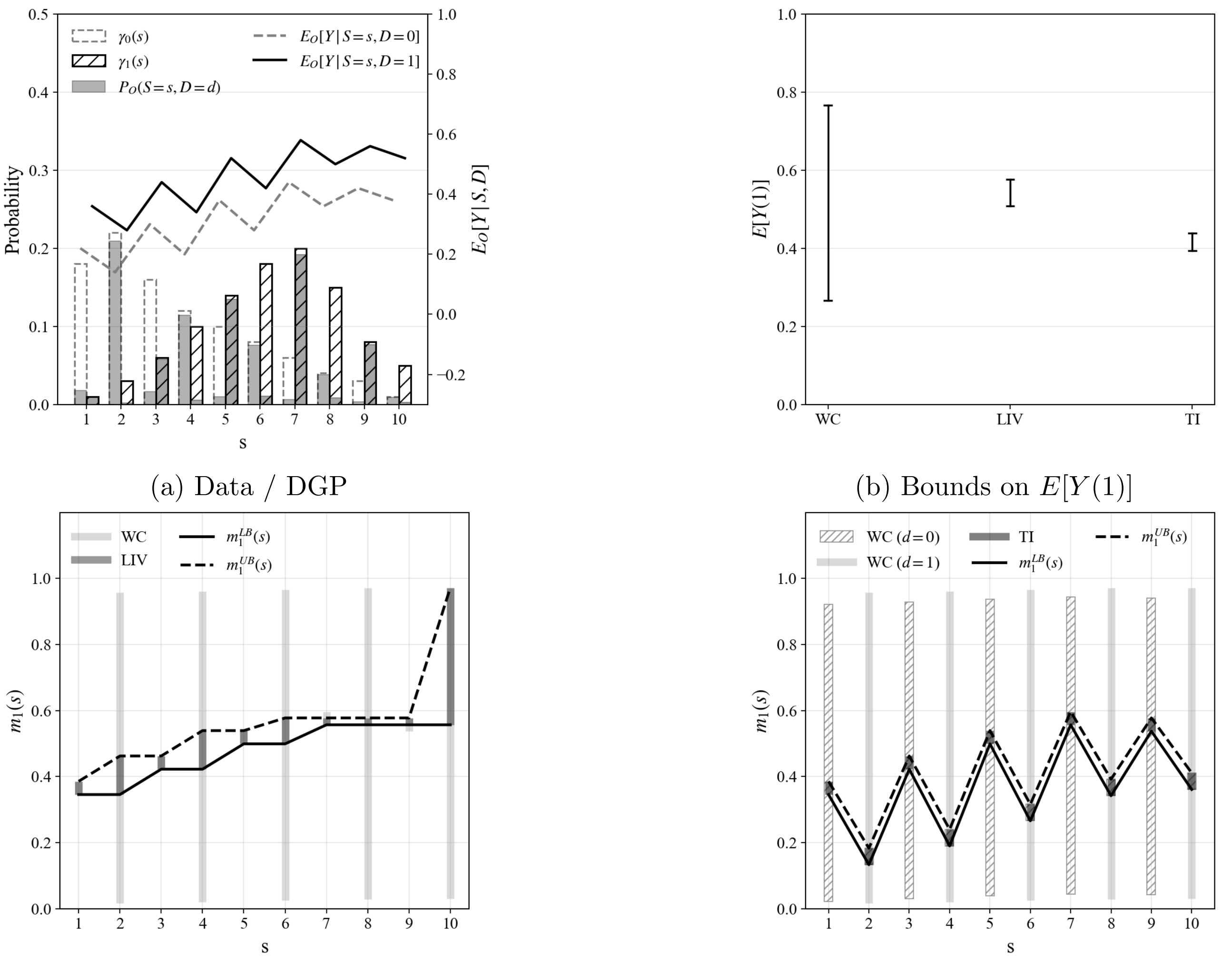
These represent the extremal monotonic m_1 that are compatible with worst-case (WC) bounds on m_1 given γ in (14). Similarly, under Assumption TI:

$$\begin{aligned} m_1^{LB}(s) &= \sup_{d \in \{0,1\}} \left[E_O[Y|S = s, D = d] \pi_{\gamma_d}(s) + \inf \mathcal{Y} \left(1 - \pi_{\gamma_d}(s)\right) \right], \\ m_1^{UB}(s) &= \inf_{d \in \{0,1\}} \left[E_O[Y|S = s, D = d] \pi_{\gamma_d}(s) + \sup \mathcal{Y} \left(1 - \pi_{\gamma_d}(s)\right) \right]. \end{aligned} \tag{19}$$

9. If compliance is perfect, there is a unique feasible γ. Otherwise, there may be a set of feasible γ, and the bounds are obtained by taking the union over all feasible γ.

Under TI, m_1 must satisfy the WC bounds for both m_1 and m_0. Hence the bound on it is given by the intersection of the two WC bounds at each support point s.

Based on these results, identifying power under Assumption LIV comes from the extent to which the WC bounds on m_1 violate monotonicity. Imposing monotonicity replaces the WC lower (upper) endpoint by its left-envelope $\sup_{s' \leq s}(\cdot)$ (right-envelope $\inf_{s' \geq s}(\cdot)$), which raises the lower bound (lowers the upper bound) wherever the WC endpoints are decreasing. Under Assumption TI, identifying power comes from how small the intersection of the WC bounds for m_1 and m_0 is at each s. The bounds tighten when those intervals overlap little, e.g., when some $\pi_{\gamma_d}(s)$ are large, so the corresponding WC interval is narrow. Both mechanisms are illustrated for a non-pathological DGP in Figure 1. Finally, bounded $\mathcal{Y}$ is necessary for either assumption to yield identifying power, since otherwise WC bounds are trivial for any feasible γ.



(a) Data / DGP

(b) Bounds on $E[Y(1)]$

(c) LIV bounds on $m_1(s)$

(d) TI bounds on $m_1(s)$

Figure 1: Illustrative example: LIV and TI bounds

Notes: Panel 1a specifies the DGP and γ. Panels 1c and 1d depict the worst-case (WC) bounds on m_d given by (14), m_1^{LB} and m_1^{UB} which deliver the lower and upper bounds on $E[Y(1)]$.

Remark 6. Both Assumptions LIV and TI can yield point identification of m, given γ. For Assumption TI, an important case is when D is constant in the observational data ($G = O$), for instance, when one treatment is not available in that dataset. For Assumption LIV, point identification arises when the WC endpoints are sufficiently non-monotonic that imposing non-decreasing m_d makes it constant (since the lower and upper monotone envelopes coincide). This may happen, for example, when the WC endpoints are strictly decreasing in s, or when the endpoints exhibit more pronounced oscillations than in Figure 1.

6 Empirical Illustration: Long-term Effects of Head Start Participation

Head Start is the largest early childhood education program in the United States, serving approximately 730,000 low-income preschool-age children in 2023.[10] Established in 1965 as part of the "War on Poverty," it was intended to help narrow gaps between disadvantaged and more advantaged children on a national scale. Its long-term effects have been studied extensively, primarily using observational designs. A common approach compares siblings within the same family who did and did not participate in Head Start (Currie and Thomas (1995), Garces, Thomas, and Currie (2002), Deming (2009), Bauer and Schanzenbach (2016)), while others exploit variation in program funding, income-based eligibility, or rollout timing to identify local average treatment effects (Ludwig and Miller (2007), Carneiro and Ginja (2014), Bailey, Sun, and Timpe (2021)). Kline and Walters (2016) instead monetize the experimental LATE of Head Start on test scores using estimates of the test-score–earnings relationship from administrative follow-up of Tennessee Project STAR participants (Chetty et al. (2011)). Despite this long history, the literature has yet to reach consensus on the program's long-term impacts (Gibbs, Ludwig, and Miller (2011), Pages et al. (2020)), as the assumptions and generalizability of parameters identified by existing approaches remain debated (Ludwig and Phillips (2008), Elango et al. (2015), Gonzalez (2020), García et al. (2020), Miller, Shenhav, and Grosz (2023)). I illustrate an alternative approach by applying the proposed method to estimate long-term average treatment effects of Head Start for eligible individuals under assumptions that do not restrict selection into treatment.

6.1 Data

I combine individual-level data from the Head Start Impact Study (HSIS) and the Child and Young Adult Supplement to the National Longitudinal Survey of Youth 1979 cohort (CNLSY).

10. Link: https://headstart.gov/program-data/article/head-start-program-facts-fiscal-year-2023 (Last accessed 01/14/2025).

HSIS was an experimental trial of Head Start mandated by the 105th US Congress as part of the program's 1998 reauthorization. In the fall of 2002, 4,667 children from nationally representative cohorts aged 3 and 4 were enrolled in the experiment. Across 383 randomly selected Head Start centers, participants were randomized either to a treatment group assigned to enroll in Head Start or to a control group barred from enrolling, yielding $Z \in \{0, 1\}$. Since participants were followed only through third grade, HSIS does not contain adolescence or adulthood outcomes that may be of interest. However, it does contain childhood math and reading test scores. I follow Kline and Walters (2016) and Kamat (2024) by pooling all children into a single cohort, then retaining only those with available scores, resulting in an experimental sample size of $n_E = 3{,}540$.

CNLSY is a biennial longitudinal survey introduced in 1986 that has tracked 11,545 children born to participants in the National Longitudinal Survey of Youth 1979 cohort (NLSY79), which, like HSIS, was designed to be nationally representative. It reveals long-term outcomes previously studied in the literature and non-randomized Head Start participation. As in Deming (2009), I consider eight long-term outcomes: grade repetition, diagnosis of a learning disability, high school graduation, "idleness", criminal involvement, teenage parenthood, self-reported health status, and average earnings. CNLSY also contains comparable childhood math and reading test scores that can be used as S, in conjunction with HSIS.[11] Following Carneiro and Ginja (2014, Section IIA), I construct the observational sample by selecting individuals who were either eligible for Head Start or reported having participated in the program. This yields an observational sample of size $n_O = 2{,}535$. Section A.2 provides additional details on data construction.

Table 2 presents descriptive statistics for the two samples. They have comparable gender compositions and similar shares of white and non-white individuals. Moreover, both NLSY79 and HSIS were designed to be nationally representative, and both pertain to the same treatment, lending support to Assumption EV. While lower than in the experiment, a significant proportion of CNLSY individuals participate in the program. The rate of compliance with the assigned treatment is 83.8% in the experiment. Imperfect compliance and the availability of the treatment in the observational population preclude direct application of existing data-combination methods that assume either perfect compliance or that the treatment is available only in the experiment. Nevertheless, the method developed here can still be used to estimate the effects under their corresponding identifying assumptions. I illustrate this in the following section by reporting bounds under relevant restrictions following from Athey, Chetty, and Imbens (2025) and García

11. Fadeout, i.e., disappearance of *average* effects on test scores over time, need not invalidate their use as S. Here, distributional effects are relevant, not only mean effects, as captured by γ (see also Bitler, Hoynes, and Domina (2014)). More generally, even if there are no distributional effects on S, m_d may vary across d for the scores.

et al. (2020).[12] Despite imperfect compliance, the estimated lower bounds on γ_d are close to summing to one, suggesting that the combined data are highly informative about the short-term potential outcome distributions.[13]

Table 2: Summary Statistics

	HSIS		CNLSY	
Variable	Mean	SD	Mean	SD
A *Individual Characteristics*				
Male	0.504	0.500	0.512	0.500
White	0.319	0.466	0.278	0.448
Math Score	50.921	24.649	40.963	26.066
Reading Score	55.225	24.413	52.500	25.853
Repeat Grade	-	-	0.320	0.467
Learning Disability	-	-	0.057	0.231
HS Graduate	-	-	0.847	0.360
Idle	-	-	0.173	0.379
Crime	-	-	0.389	0.488
Teen Pregnancy	-	-	0.244	0.430
Poor Health	-	-	0.166	0.373
Average Earnings (in 000)	-	-	22.166	17.237
B *Program Characteristics*				
D	0.531	0.499	0.443	0.497
Z	0.596	0.491	-	-
$\mathbb{1}[D = Z]$	0.838	0.369	-	-
Observations	3,540		2,535	

Notes: Summary statistics from the Head Start Impact Study (HSIS) and the Child and Young Adult Supplement of the National Longitudinal Survey of Youth 1979 cohort (CNLSY). D denotes Head Start participation and Z is experimental treatment assignment. $\mathbb{1}$ denotes the indicator variable, and SD the sample standard deviation.

6.2 Results

Table 3 reports bound estimates under previously discussed assumptions, using the consistent estimator proposed in Section S.1. Worst-case bounds impose no restrictions on the temporal link functions m, coincide with the bounds of Manski (1990), as shown in Section 3, and therefore

12. Both assume perfect compliance. In García et al. (2020), the intervention is available only in the experiment, which is typical of "model" early childhood intervention programs such as the Carolina Abecedarian and Perry Preschool Projects, but not of large-scale programs such as Head Start.

13. In the empirical implementation, I fix γ_d at these bounds. This linearizes the optimization programs and substantially simplifies computation without resorting to coarse discretization; see Appendix S.1.1.

necessarily include zero. Thus, identifying even the sign of the effect requires additional assumptions. For all outcomes, each of the previously mentioned assumptions substantially reduces the width of the estimated bounds.

Table 3: Bounds Estimates

		Modeling Assumption			
Outcome	n_O	Worst-case	LIV	TI	LUC
Repeat Grade	2,455	−0.476 0.524	−0.053 −0.011	−0.075 0.050	−0.008 −0.008
Learning Disability	2,527	−0.448 0.552	−0.008 0.008	−0.076 0.050	−0.001 −0.001
HS Graduate	2,031	−0.525 0.475	0.019 0.032	−0.079 0.050	−0.002 −0.002
Idle	2,517	−0.474 0.526	−0.046 −0.015	−0.072 0.051	−0.030 −0.030
Crime	2,517	−0.499 0.501	−0.040 −0.012	−0.074 0.048	−0.032 −0.032
Teen Pregnancy	2,517	−0.460 0.540	0.003 0.022	−0.072 0.050	0.013 0.013
Poor Health	2,517	−0.466 0.534	−0.050 −0.029	−0.078 0.044	−0.010 −0.010
Average earnings (in 000)	2,395	−110.895 131.208	−3.841 1.621	−19.413 11.583	−2.449 −2.449

Notes: Estimated bounds, represented as $\frac{\text{Lower Bound}}{\text{Upper Bound}}$, for different long-term outcomes. Worst-case bounds impose no restrictions on m. The remaining bounds impose only the noted modeling assumption.

Assumption LIV maintains that temporal link functions are monotonic in each of the two test score components. For high school graduation and earnings, I impose weak monotonicity increasing in each potential test score. For grade repetition, learning disability diagnosis, "idleness", crime, teen pregnancy, and poor health, I impose weak monotonicity decreasing in each potential test score. This assumption may be particularly appealing for outcomes related to education, employment, crime, and earnings. Taking high school graduation as an example, it would hold if, fixing any Head Start participation, individuals with higher math or reading scores are equally or more likely to graduate from high school than individuals with lower scores.

Estimates under Assumption LIV reveal the sign of the effect for all but two outcomes. They further indicate that Head Start participation reduces the probability of grade repetition by at least 1.1 percentage points (pp), idleness by at least 1.5 pp, criminal involvement by at least 1.2 pp, and reporting poor health by at least 2.9 pp. The corresponding upper bounds are also quite informative. Overall, the estimates suggest that beneficial effects on grade repetition, learning disability, high school graduation, idleness, and poor health may be more modest than reported by the sibling study in Deming (2009). Compared to the same study, the effect on criminal involvement found here has the expected sign, while the impact on teen pregnancy does not. Moreover, because the present data allow for longer follow-up than in Deming (2009), they reveal earnings for a larger fraction of individuals. However, the sign of the effect on earnings remains unidentified.

Next, I turn to illustrative results relying on temporal link function restrictions that follow from previously proposed methods. These methods cannot be directly utilized here because of imperfect compliance and the availability of the treatment in CNLSY. However, as argued above, the proposed framework can be combined with the relevant assumptions, thereby extending their applicability. Estimates under Assumption TI also lead to a substantial reduction in the width of the bounds, though none exclude zero. Assumption TI would hold if the model proposed by García et al. (2020) is credible. That said, HSIS does not contain all of the short-term outcomes included by the authors in their model of the Carolina Abecedarian and CARE programs, nor the medium-term experimental outcomes they use to validate their choice of short-term outcomes. The corresponding results should therefore be interpreted with caution. The final column reports estimates under Assumption LUC from Athey, Chetty, and Imbens (2025). Since this assumption point-identifies m and the γ_d are fixed by empirical bounds, the estimated bounds for the LTE collapse to points. However, the estimated signs are not aligned with previous findings for some outcomes. As emphasized by Imbens et al. (2024), one reason for this may be that the estimates are biased by so-called long-term confounders–unobservables that relate Head Start participation, the short-term test scores, and the long-term outcome.

7 Conclusion

Recent literature proposes augmenting long-term observational studies with short-term experiments to provide alternatives to conventional long-term observational studies. This paper shows that such data combination is not a substitute for credible modeling assumptions. Nevertheless, it remains appealing for this purpose. Assumptions relating short-term to long-term potential outcomes may be defensible based on economic theory or intuition, and thus conducive to plausible inference. Data combination may be used to amplify the identifying power of such restrictions and thereby may yield more informative plausible inference than observational data alone.

This paper introduces two assumptions that exploit this feature of data combination. It also provides a general identification approach that enables computational derivation of bounds under new modeling assumptions, facilitating further developments. Tailor-made assumptions that are plausible in specific empirical settings are an interesting topic for future research, which may benefit from these results.

References

Aizer, Anna, Nancy Early, Shari Eli, Guido Imbens, Keyoung Lee, Adriana Lleras-Muney, and Alexander Strand. 2024. "The Lifetime Impacts of the New Deal's Youth Employment Program." *The Quarterly Journal of Economics.*

Al-Khayyal, Faiz A. 1992. "Generalized bilinear programming: Part I. Models, applications and linear programming relaxation." *European Journal of Operational Research* 60 (3): 306–314.

Artstein, Zvi. 1983. "Distributions of random sets and random selections." *Israel Journal of Mathematics* 46:313–324.

Athey, Susan, Raj Chetty, and Guido Imbens. 2025. "Using Experiments to Correct for Selection in Observational Studies." *arXiv preprint arXiv:2006.09676.*

Athey, Susan, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2025. "The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely." *Review of Economic Studies,* rdaf087.

Attanasio, Orazio P, Costas Meghir, and Ana Santiago. 2012. "Education choices in Mexico: using a structural model and a randomized experiment to evaluate Progresa." *The Review of Economic Studies* 79 (1): 37–66.

Bailey, Martha J, Shuqiao Sun, and Brenden Timpe. 2021. "Prep school for poor kids: The long-run impacts of Head Start on human capital and economic self-sufficiency." *American Economic Review* 111 (12): 3963–4001.

Bauer, Lauren, and Diane Whitmore Schanzenbach. 2016. "The long-term impact of the Head Start program." *The Hamilton Project.*

Beresteanu, Arie, Ilya Molchanov, and Francesca Molinari. 2012. "Partial identification using random set theory." *Journal of Econometrics* 166 (1): 17–32.

Bitler, Marianne P, Hilary W Hoynes, and Thurston Domina. 2014. *Experimental evidence on distributional effects of Head Start.* Technical report. National Bureau of Economic Research.

Carneiro, Pedro, and Rita Ginja. 2014. "Long-term impacts of compensatory preschool on health and behavior: Evidence from Head Start." *American Economic Journal: Economic Policy* 6 (4): 135–173.

Chen, Jiafeng, and David M Ritzwoller. 2023. "Semiparametric estimation of long-term treatment effects." *Journal of Econometrics* 237 (2): 105545.

Chesher, Andrew, and Adam M Rosen. 2017. "Generalized instrumental variable models." *Econometrica* 85 (3): 959–989.

———. 2020. "Generalized instrumental variable models, methods, and applications." In *Handbook of Econometrics,* 7:1–110. Elsevier.

Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *The Quarterly journal of economics* 126 (4): 1593–1660.

Cross, Philip J, and Charles F Manski. 2002. "Regressions, short and long." *Econometrica* 70 (1): 357–368.

Currie, Janet, and Douglas Almond. 2011. "Human capital development before age five." In *Handbook of labor economics,* 4:1315–1486. Elsevier.

Currie, Janet, and Duncan Thomas. 1995. "Does Head Start Make a Difference?" *The American Economic Review,* 341–364.

Deming, David. 2009. "Early childhood intervention and life-cycle skill development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1 (3): 111–134.

Dutz, Deniz, Ingrid Huitfeldt, Santiago Lacouture, Magne Mogstad, Alexander Torgovitsky, and Winnie Van Dijk. 2021. *Selection in surveys: Using randomized incentives to detect and account for nonresponse bias.* Technical report. National Bureau of Economic Research.

Dynarski, Susan, CJ Libassi, Katherine Michelmore, and Stephanie Owen. 2021. "Closing the gap: The effect of reducing complexity and uncertainty in college pricing on the choices of low-income students." *American Economic Review* 111 (6): 1721–1756.

D'Haultfœuille, Xavier, Christophe Gaillac, and Arnaud Maurel. 2024. "Partially linear models under data combination." *Review of Economic Studies,* rdae022.

Elango, Sneha, Jorge Luis García, James J Heckman, and Andrés Hojman. 2015. "Early childhood education." In *Economics of Means-Tested Transfer Programs in the United States, Volume 2,* 235–297. University of Chicago Press.

Fan, Yanqin, Robert Sherman, and Matthew Shum. 2014. "Identifying treatment effects under data combination." *Econometrica* 82 (2): 811–822.

Galichon, Alfred, and Marc Henry. 2011. "Set identification in models with multiple equilibria." *The Review of Economic Studies* 78 (4): 1264–1298.

Garces, Eliana, Duncan Thomas, and Janet Currie. 2002. "Longer-term effects of Head Start." *American economic review* 92 (4): 999–1012.

García, Jorge Luis, James J Heckman, Duncan Ermini Leaf, and María José Prados. 2020. "Quantifying the life-cycle benefits of an influential early-childhood program." *Journal of Political Economy* 128 (7): 2502–2541.

Ghassami, AmirEmad, Alan Yang, David Richardson, Ilya Shpitser, and Eric Tchetgen Tchetgen. 2022. "Combining experimental and observational data for identification and estimation of long-term causal effects." *arXiv preprint arXiv:2201.10743.*

Gibbs, Chloe, Jens Ludwig, and Douglas L Miller. 2011. *Does Head Start do any lasting good?* Technical report. National Bureau of Economic Research.

Gonzalez, Kathryn E. 2020. "Within-family differences in Head Start participation and parent investment." *Economics of Education Review* 74:101950.

Griffen, Andrew S, and Petra E Todd. 2017. "Assessing the performance of nonexperimental estimators for evaluating Head Start." *Journal of Labor Economics* 35 (S1): S7–S63.

Gupta, Somit, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al. 2019. "Top challenges from the first practical online controlled experiments summit." *ACM SIGKDD Explorations Newsletter* 21 (1): 20–35.

Gurobi Optimization. 2024. *Gurobi Optimizer Reference Manual.* https://www.gurobi.com.

Heckman, James, Neil Hohmann, Jeffrey Smith, and Michael Khoo. 2000. "Substitution and dropout bias in social experiments: A study of an influential social experiment." *The Quarterly Journal of Economics* 115 (2): 651–694.

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the mechanisms through which an influential early childhood program boosted adult outcomes." *American Economic Review* 103 (6): 2052–2086.

Heckman, James J, and Edward J Vytlacil. 1999. "Local instrumental variables and latent variable models for identifying and bounding treatment effects." *Proceedings of the national Academy of Sciences* 96 (8): 4730–4734.

Hoynes, Hilary W, and Diane Whitmore Schanzenbach. 2018. *Safety net investments in children.* Technical report. National Bureau of Economic Research.

Hu, Wenjie, Xiaohua Zhou, and Peng Wu. 2022. "Identification and estimation of treatment effects on long-term outcomes in clinical trials with external observational data." *arXiv preprint arXiv:2208.10163.*

Imbens, Guido, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. 2024. "Long-term causal inference under persistent confounding via data combination." *arXiv preprint arXiv:2202.07234.*

Imbens, Guido W, and Joshua D Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–475.

Kamat, Vishal. 2024. "Identifying the effects of a program offer with an application to head start." *Journal of Econometrics* 240 (1): 105679.

Kline, Patrick, and Christopher R Walters. 2016. "Evaluating public programs with close substitutes: The case of Head Start." *The Quarterly Journal of Economics* 131 (4): 1795–1848.

Ludwig, Jens, and Douglas L Miller. 2007. "Does Head Start improve children's life chances? Evidence from a regression discontinuity design." *The Quarterly journal of economics* 122 (1): 159–208.

Ludwig, Jens, and Deborah A Phillips. 2008. "Long-term effects of Head Start on low-income children." *Annals of the New York Academy of Sciences* 1136 (1): 257–268.

Luo, Ye, Kirill Ponomarev, and Hai Wang. 2024. *Selecting Inequalities for Sharp Identification in Models with Set-Valued Predictions.* Technical report. Working Paper, University of Chicago.

Manski, Charles F. 2009. *Identification for prediction and decision.* Harvard University Press.

———. 1990. "Nonparametric bounds on treatment effects." *The American Economic Review* 80 (2): 319–323.

Manski, Charles F, and John V Pepper. 2000. "Monotone Instrumental Variables: With an Application to the Returns to Schooling." *Econometrica* 68 (4): 997–1010.

———. 2009. "More on monotone instrumental variables." *The Econometrics Journal* 12 (suppl_1): S200–S216.

Miller, Douglas L, Na'ama Shenhav, and Michel Grosz. 2023. "Selection into identification in fixed effects models, with application to Head Start." *Journal of Human Resources* 58 (5): 1523–1566.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using instrumental variables for inference about policy relevant treatment parameters." *Econometrica* 86 (5): 1589–1619.

Molchanov, Ilya. 2017. *Theory of Random Sets.* 2nd ed. Vol. 87. Probability Theory and Stochastic Modelling. Springer.

Molchanov, Ilya, and Francesca Molinari. 2014. "Applications of random set theory in econometrics." *Annu. Rev. Econ.* 6 (1): 229–251.

———. 2018. *Random Sets in Econometrics.* Vol. 60. Cambridge University Press.

Molinari, Francesca, and Marcin Peski. 2006. "Generalization of a Result on 'Regressions, Short and Long." *Econometric Theory* 22 (1): 159–163.

Pages, Remy, Dylan J Lukes, Drew H Bailey, and Greg J Duncan. 2020. "Elusive longer-run impacts of head start: Replications within and across cohorts." *Educational Evaluation and Policy Analysis* 42 (4): 471–492.

Park, Yechan, and Yuya Sasaki. 2024a. "A Bracketing Relationship for Long-Term Policy Evaluation with Combined Experimental and Observational Data." *arXiv preprint arXiv:2401.12050.*

———. 2024b. "The Informativeness of Combined Experimental and Observational Data under Dynamic Selection." *arXiv preprint arXiv:2403.16177.*

Prentice, Ross L. 1989. "Surrogate endpoints in clinical trials: definition and operational criteria." *Statistics in medicine* 8 (4): 431–440.

Puma, Michael, Stephen Bell, Ronna Cook, Camilla Heid, Gary Shapiro, Pam Broene, Frank Jenkins, Philip Fletcher, Liz Quinn, Janet Friedman, et al. 2010. "Head Start Impact Study. Final Report." *Administration for Children & Families.*

Rambachan, Ashesh, Rahul Singh, and Davide Viviano. 2024. "Program evaluation with remotely sensed outcomes." *arXiv preprint arXiv:2411.10959.*

Ridder, Geert, and Robert Moffitt. 2007. "The econometrics of data combination." *Handbook of econometrics* 6:5469–5547.

Rockafellar, Ralph Tyrell. 1970. *Convex Analysis.* Princeton: Princeton University Press. ISBN: 9781400873173. https://doi.org/doi:10.1515/9781400873173. https://doi.org/10.1515/9781400873173.

Russell, Thomas M. 2021. "Sharp bounds on functionals of the joint distribution in the analysis of treatment effects." *Journal of Business & Economic Statistics* 39 (2): 532–546.

Shea, Joshua. 2022. "Testing for racial bias in police traffic searches." *University of Illinois, Champaign Urbana, USA.*

Shi, Xiaoxia, and Matthew Shum. 2015. "Simple two-stage inference for a class of partially identified models." *Econometric Theory* 31 (3): 493–520.

Todd, Petra E, and Kenneth I Wolpin. 2006. "Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility." *American economic review* 96 (5): 1384–1417.

———. 2023. "The best of both worlds: combining randomized controlled trials with structural modeling." *Journal of Economic Literature* 61 (1): 41–85.

Torgovitsky, Alexander. 2019. "Nonparametric inference on state dependence in unemployment." *Econometrica* 87 (5): 1475–1505.

Van Goffrier, Graham, Lucas Maystre, and Ciarán Mark Gilligan-Lee. 2023. "Estimating long-term causal effects from short-term experiments and long-term observational data with unobserved confounding." In *Conference on Causal Learning and Reasoning,* 791–813. PMLR.

Voronin, Andrei. 2025. "Generalized Optimal Transport." *arXiv preprint arXiv:2507.22422.*

Appendices

Appendix A Additional Discussions

A.1 Discretization of Short-term Outcomes

In this section, I clarify the implications of discretizing short-term outcomes. To this end, let a researcher pose a surjective discretization function $\lambda : \mathcal{S} \to \mathcal{S}^D := \{1, 2, \dots, k\}$ for some $k < \infty$, and define $S^D(d) = \lambda(S(d))$. Note that this subsumes the case in which $S(d)$ is finitely supported, since then $\lambda(s) = s$ for all $s \in \mathcal{S}$. I introduce λ to clarify the subtle differences in applications of results of Section 5.3 when $S(d)$ is finitely supported and discretized. Similarly define discretized temporal link functions $m_d^D : \mathcal{S}^D \to \mathcal{Y}$, given by $m_d^D = E[Y(d)|S^D(d)] = E[Y(d)|\lambda(S(d))]$, and let $m^D = (m_0^D, m_1^D)$. Pose the following analog of Assumption MA under the discretization.

Assumption MA:D. *Suppose* $\mathcal{M}^A$ *and* $\mathcal{M}^D$ *are known or identified sets, and that* $m \in \mathcal{M}^A \subseteq \mathcal{M}$. *Then* λ *is such that* $m^D \in \mathcal{M}^D$.

Assumptions MA and MA:D are closely related. The former maintains that the researcher imposes some modeling assumption that will restrict feasible m, as in Section 4. The latter strengthens this notion and assumes that additionally m^D satisfies known restrictions after discretization. Of course, if Assumption MA holds for a finitely supported $S(d)$, then Assumption MA:D trivially follows by taking λ to be an identity function up to necessary relabeling of $S(d)$ values, if any. The remark below explains that for some modeling assumptions and discretization functions, MA:D follows immediately from MA, but that it may be restrictive for others.

Remark 7. Consider Assumption LIV when $S(d)$ is a scalar stating that $E[Y(d)|S(d) = s]$ are nondecreasing functions. Then $E[Y(d)|S^D = s]$ must also be nondecreasing for any order-preserving λ, so Assumption MA:D holds for an appropriately chosen λ. However, LUC states that $m_d(s) = E_O[Y|S = s, D = d]$, which does not directly imply that $m_d^D(s) = E[Y|S^D = s, D = d]$. A similar remark can be made for treatment invariance.

If $S(d)$ is finitely supported, MA and MA:D are equivalent and Section 5.3 characterizes the identified set. If $S(d)$ is discretized and Assumption MA:D holds as a direct consequence of Assumption MA, such as under LIV, then results characterize the identified set $\mathcal{H}(\tau)$ that is sharp *under finitely-supported short-term outcomes.*[14] This is also the case if the researcher believes the modeling assumption holds under discretized data, i.e., is willing to maintain MA:D directly. Otherwise, the results in Section 5.3 should be viewed as providing an approximation of the identified set.

A.2 Data Creation

The construction of the two samples follows previous related work. The definitions of the long-term outcomes in CNLSY follow Deming (2009). Grade retention and diagnosis of a learning disability are defined as having reported being retained in any grade in school and being diagnosed with a learning disability, respectively. High school graduation is defined as reporting having graduated from high school or completing General Educational Development certification. Individuals are classified as idle if, in their most recent interview year, they report neither wages nor school attendance. Criminal involvement is defined as ever reporting having been convicted of a crime, placed on probation, sentenced by a judge, or incarcerated. Health status is measured by averaging responses to a Likert scale item on self-reported health status and generating an

14. Note that this set may be larger than the intractable identified set that would have been obtained using non-discretized data.

indicator equal to one if it is below three on a five-point scale. Finally, earnings are averaged across all reported values and converted to 2020 dollars using the Bureau of Labor Statistics Consumer Price Index.

Program participation in CNLSY is determined by the question of whether the child has ever attended Head Start, while HSIS contains indicators for true participation.[15] Eligibility of non-participants in CNLSY must be inferred. For this, I rely on eligibility variables constructed by Carneiro and Ginja (2014, Section IIA). Eligibility is inferred by determining whether the child met the contemporaneous program requirements based on survey responses. Children ages three to five have been eligible if their family income is below the federal poverty line, or if their family has been eligible for any of the following public assistance programs: Aid to Families with Dependent Children (AFDC) or Temporary Assistance for Needy Families (TANF) after 1996, or Supplemental Security Income (SSI). Poverty status is verified by comparing the reported family income at ages three to five with the relevant federal poverty line, which is dependent on the family size and year. Eligibility for AFDC/TANF is determined based on two family income tests: the gross income test and the countable income test, as well as other pertinent categorical requirements. The income tests have state-specific thresholds that may vary by year and family size. Additionally, AFDC requires a specific family structure: either it must be female-headed, or with an unemployed main earner. The observational sample consists of individuals who were either determined to have been eligible for Head Start, or have participated in the program based on the relevant responses.

HSIS contains Woodcock-Johnson III (WJ-III) cognitive assessment scores, which are used as short-term outcomes. For compatibility with the observational data, following Griffen and Todd (2017), I create composite scores for math and reading ability by averaging the national-level percentile scores on the corresponding components of the cognitive ability test and using them as a two-dimensional S, each binned to 20 support points, giving a total of 400 support points. As corresponding measures of math and reading ability in CNLSY, I take the percentile scores on the Peabody Individual Achievement Math and Reading Recognition subtests, also binned to a total of 400 support points.

Appendix B Proofs

This section contains the proofs of the main results. It begins by summarizing notation. Supplemental Appendix S.2 contains auxiliary lemmas and propositions, with their proofs.

15. In this paper, I abstract away from substitution bias (Heckman et al. (2000)), as in the main analysis of García et al. (2020). I consider the effects of Head Start participation compared to non-participation, irrespective of the take-up of alternatives.

Preliminaries and Notation

Equality of distribution of two random elements or a random element and a law is denoted by $\stackrel{d}{=}$ (e.g. $R \stackrel{d}{=} P_R$ and $R \stackrel{d}{=} R'$). I denote random sets with boldface letters (e.g. $\mathbf{R}$) and use $\mathbf{R}, Z$ to denote the random set $\mathbf{R} \times \{Z\}$. $\mathbb{E}(\mathbf{R}|X)$ is used for the conditional Aumann expectation of a random set $\mathbf{R}$ given a sigma-algebra generated by a random vector X. The set of all selections of $\mathbf{R}$ is denoted by $Sel(\mathbf{R})$. The set of all random vectors $R \in Sel(\mathbf{R})$ such that $E[||R||] < \infty$ is denoted by $Sel^1(\mathbf{R})$. I use $\stackrel{d}{=}$ to denote that a random element has a law, or an equivalent distribution-determining functional. A, B and K represent sets. $\mathcal{C}(A)$ and $\mathcal{B}(A)$ are the families of all closed and Borel subsets of the set A, respectively. $co(A)$ is the closed convex hull of the set A. The identified set for a generic parameter θ is $\mathcal{H}(\theta)$. The set of distribution functions of random vectors with support $\mathcal{R}$ is $\mathcal{P}^{\mathcal{R}}$. I assume throughout that $\mathcal{Y} \times \mathcal{S}$ is a locally compact, second countable Hausdorff space, more precisely $\mathbb{R}^{1+d}$ endowed with its natural topology, while any of its subspaces inherit their relative topologies.

In the proofs for simpler notation, I will use the following random variable:

$$\tilde{Z} = \mathbb{1}[G = E]Z + \mathbb{1}[G = O](sup\mathcal{Z} + 1). \tag{20}$$

I use LIE to refer to the "law of iterated expectations".

B.1 Main Results

Proof of Proposition 1. If only Assumptions RA, EV hold, by Lemma S.4 *i*), $\mathcal{H}(P_{Y(0)}, P_{Y(1)}) = \mathcal{H}^O(P_{Y(0)}, P_{Y(1)})$. Then, $\mathcal{H}(\tau) = \mathcal{H}^O(\tau)$, since τ is a functional of $(P_{Y(0)}, P_{Y(1)})$. Therefore, if $\mathcal{H}(\tau) \subsetneq \mathcal{H}^O(\tau)$ and RA, EV hold, additional assumptions must be imposed. If $\mathcal{H}(\tau) \neq \emptyset$, these assumptions are not refuted. Henceforth, denote by $\mathcal{H}^A(\cdot)$ and $\mathcal{H}(\cdot)$ the identified sets with and without the additional assumptions. Let $\mathcal{H}^{A,O}(\cdot)$ be identified sets with additional assumptions, using only observational data.

Given that $\mathcal{H}^A(\tau)$ and $\mathcal{H}^{O,A}(\tau)$ are images of T over $\mathcal{H}^A(m,\gamma)$ and $\mathcal{H}^{O,A}(m,\gamma)$, respectively, the additional assumptions must directly or indirectly restrict m, γ or (m,γ). Otherwise, $\mathcal{H}^O(m,\gamma) = \mathcal{H}^{A,O}(m,\gamma)$ and $\mathcal{H}(m,\gamma) = \mathcal{H}^A(m,\gamma)$, so $\mathcal{H}^A(\tau) = \mathcal{H}(\tau) = \mathcal{H}^O(\tau) = \mathcal{H}^{A,O}(\tau)$ where the second equality is by Lemma S.4 *i*).

By way of contradiction, suppose that $\emptyset \neq \mathcal{H}^A(\tau) \subsetneq \mathcal{H}^{A,O}(\tau)$ and that only γ is further restricted by the assumptions, or equivalently $\gamma \in \Gamma^A$, for some $\Gamma^A \subsetneq (\mathcal{P}^{\mathcal{S}})^2$. Then:

$$\begin{aligned}\mathcal{H}^A(P_{Y(0)}, P_{Y(1)}, \gamma_0, \gamma_1) &= \mathcal{H}(P_{Y(0)}, P_{Y(1)}, \gamma_0, \gamma_1) \cap \left((\mathcal{P}^{\mathcal{Y}})^2 \times \Gamma^A\right) \\ &= \mathcal{H}(P_{Y(0)}, P_{Y(1)}) \times \left(\mathcal{H}(\gamma_0, \gamma_1) \cap \Gamma^A\right)\end{aligned} \tag{21}$$

where the first equality is by the fact that only γ is restricted by assumption, and the second is by Lemma S.4 *ii*). It is then immediate that:

$$\mathcal{H}^A(P_{Y(0)}, P_{Y(1)}) = \mathcal{H}(P_{Y(0)}, P_{Y(1)}) = \mathcal{H}^O(P_{Y(0)}, P_{Y(1)})$$

where the first equality is by (21) and the definition of a projection, and the second is by Lemma S.4 *i*). Thus, $\mathcal{H}^A(\tau) = \mathcal{H}^O(\tau)$. But since $\mathcal{H}^A(\tau) \subseteq \mathcal{H}^{A,O}(\tau) \subseteq \mathcal{H}^O(\tau)$, then it must also be that $\mathcal{H}^A(\tau) = \mathcal{H}^{A,O}(\tau)$, contradicting $\mathcal{H}^A(\tau) \subsetneq \mathcal{H}^{A,O}(\tau)$. Therefore, the additional assumptions must restrict m or (m, γ). In either case, $m \in \mathcal{M}^A \subsetneq \mathcal{M}$. □

Proof of Lemma 1. Recall by (20) that $\tilde{Z} = \mathbb{1}[G = E]Z + \mathbb{1}[G = O](\sup \mathcal{Z} + 1) \in \tilde{\mathcal{Z}}$. Note that $\tilde{Z} = Z$ when $G = E$ and $\tilde{Z}$ equals a distinct constant when $G = O$. Therefore, Assumptions RA and EV hold if and only if $\tilde{Z} \perp\!\!\!\perp (Y(d), S(d))$ for all $d \in \{0, 1\}$. Let $\tilde{I}$ be the set of random elements (E_1, E_2, E_3) such that $(E_1, E_2, E_3) \in \mathcal{Y} \times \mathcal{S} \times \tilde{\mathcal{Z}}$ and $(E_1, E_2) \perp\!\!\!\perp E_3$, i.e. that satisfy the two assumptions. Define the random set:

$$(\mathbf{Y}_0, \mathbf{S}_0, \mathbf{Y}_1, \mathbf{S}_1) := \begin{cases} \{(Y, S)\} \times \mathcal{Y} \times \mathcal{S}, \text{ if } (D, G) = (0, O) \\ \mathcal{Y} \times \mathcal{S} \times \{(Y, S)\}, \text{ if } (D, G) = (1, O) \\ \mathcal{Y} \times \{S\} \times \mathcal{Y} \times \mathcal{S}, \text{ if } (D, G) = (0, E) \\ \mathcal{Y} \times \mathcal{S} \times \mathcal{Y} \times \{S\}, \text{ if } (D, G) = (1, E) \\ \mathcal{Y} \times \mathcal{S} \times \mathcal{Y} \times \mathcal{S}, \text{ otherwise} \end{cases}. \tag{22}$$

As in the proof of Beresteanu, Molchanov, and Molinari (2012, Proposition 2.3), by definition of $(\mathbf{Y}_0, \mathbf{S}_0, \mathbf{Y}_1, \mathbf{S}_1)$, all information on $(Y(0), S(0), Y(1), S(1))$ in the observed data can be summarized by $(Y(0), S(0), Y(1), S(1)) \in Sel((\mathbf{Y}_0, \mathbf{S}_0, \mathbf{Y}_1, \mathbf{S}_1))$. Also, by definition of the random set, $(\mathbf{Y}_0, \mathbf{S}_0, \mathbf{Y}_1, \mathbf{S}_1) = (\mathbf{Y}_0, \mathbf{S}_0) \times (\mathbf{Y}_1, \mathbf{S}_1)$, where for $d \in \{0, 1\}$:

$$(\mathbf{Y}_d, \mathbf{S}_d) = \begin{cases} \{(Y, S)\}, \text{ if } (D, G) = (d, O) \\ \mathcal{Y} \times \{S\}, \text{ if } (D, G) = (d, E) \\ \mathcal{Y} \times \mathcal{S}, \text{ otherwise} \end{cases}. \tag{23}$$

By applying Lemma S.1 twice, $Sel((\mathbf{Y}_0, \mathbf{S}_0, \mathbf{Y}_1, \mathbf{S}_1)) = Sel((\mathbf{Y}_0, \mathbf{S}_0)) \times Sel((\mathbf{Y}_1, \mathbf{S}_1)) = Sel(\mathbf{Y}_0) \times Sel(\mathbf{S}_0) \times Sel(\mathbf{Y}_1) \times Sel(\mathbf{S}_1)$. All information about $(Y(0), S(0), Y(1), S(1))$ in the data, Assumptions RA EV and $E[|Y(d)|] < \infty$ for $d \in \{0,1\}$ can thus equivalently be expressed as $(Y(d), S(d), \tilde{Z}) \in Sel(\mathbf{Y}_d) \times Sel(\mathbf{S}_d) \times \{\tilde{Z}\} \cap \tilde{I}$ for $d \in \{0,1\}$. If Assumptions RA and EV hold, the intersection is non-empty. Let $\bar{I}$ be the set of random elements (E_1, E_2) such that $(E_1, E_2) \in \mathcal{S} \times \tilde{\mathcal{Z}}$ and $E_1 \perp\!\!\!\perp E_2$. The set of (m, γ) consistent with the data and the two assumptions follows by definition as:

$$
\begin{aligned}
&\mathcal{H}^{EV/RA}(m,\gamma) \\
&:= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M} \times \left(\mathcal{P}^{\mathcal{S}}\right)^2 : \forall d \in \{0,1\},\ \exists(\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel(\mathbf{Y}_d) \times Sel(\mathbf{S}_d) \times \{\tilde{Z}\} \cap \tilde{I}, \\ \gamma_d \stackrel{d}{=} \varsigma_d,\ m_d(\varsigma_d) = E[\upsilon_d | \varsigma_d] \text{ a.s.} \end{array} \right\} \\
&= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M} \times \left(\mathcal{P}^{\mathcal{S}}\right)^2 : \forall d \in \{0,1\},\ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}, \\ \exists \upsilon_d \in Sel(\mathbf{Y}_d),\ (\upsilon_d, \varsigma_d) \perp\!\!\!\perp \tilde{Z},\ \gamma_d \stackrel{d}{=} \varsigma_d,\ m_d(\varsigma_d) = E[\upsilon_d|\varsigma_d] \text{ a.s.} \end{array} \right\} \\
&= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M} \times \left(\mathcal{P}^{\mathcal{S}}\right)^2 : \forall d \in \{0,1\},\ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}, \\ \exists \upsilon_d \in Sel(\mathbf{Y}_d),\ (\upsilon_d, \varsigma_d) \perp\!\!\!\perp \tilde{Z},\ \gamma_d \stackrel{d}{=} \varsigma_d,\ m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d] \text{ a.s.} \end{array} \right\} \\
&= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M} \times \left(\mathcal{P}^{\mathcal{S}}\right)^2 : \forall d \in \{0,1\},\ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}, \\ \exists \upsilon_d \in Sel(\mathbf{Y}_d),\ \gamma_d \stackrel{d}{=} \varsigma_d,\ m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d] \text{ a.s.} \end{array} \right\}
\end{aligned} \tag{24}
$$

where the second equality is by Lemma S.1, and the third is by $(\upsilon_d, \varsigma_d) \perp\!\!\!\perp \tilde{Z}$ and the fourth is by Lemma S.5. It only remains to impose Assumption MA. Then, the identified set is:

$$
\begin{aligned}
\mathcal{H}(m,\gamma) &= \mathcal{H}^{EV/RA}(m,\gamma) \cap (\mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2) \\
&= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 :\ \forall d \in \{0,1\},\ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}, \\ \exists \upsilon_d \in Sel(\mathbf{Y}_d),\ \gamma_d \stackrel{d}{=} \varsigma_d,\ m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d] \text{ a.s.} \end{array} \right\} \\
&= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 :\ \forall d \in \{0,1\},\ \exists \varsigma_d \in Sel(\mathbf{S}_d) \cap I, \\ \exists \upsilon_d \in Sel(\mathbf{Y}_d),\ \gamma_d \stackrel{d}{=} \varsigma_d,\ m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d] \text{ a.s.} \end{array} \right\}.
\end{aligned} \tag{25}
$$

where the first equality is by observation and the second is since $\varsigma_d \in Sel(\mathbf{S}_d) \cap I$ can be equivalently stated as $(\varsigma_d, \tilde{Z}) \in Sel(\mathbf{S}_d, \tilde{Z}) \cap \bar{I}$.

Next note that for every $(m, \gamma) \in \mathcal{H}(m, \gamma)$, there exist $(\upsilon_0, \varsigma_0, \upsilon_1, \varsigma_1)$ that generate them and that are consistent with the data, modeling assumption, Assumptions RA and EV. Therefore, $\mathcal{H}(m, \gamma)$ is sharp. □

Proof of Theorem 1. Recall that $\bar{I}$ is the set of random elements (E_1, E_2) such that $(E_1, E_2) \in \mathcal{S} \times \tilde{\mathcal{Z}}$ and $E_1 \perp\!\!\!\perp E_2$. Likewise I is defined by Lemma 1 and $\tilde{Z}$ by (20). The proof then proceeds through a series of steps. I first show that $\mathcal{H}(m,\gamma)=\tilde{\mathcal{H}}(m,\gamma)$ where:

$$\tilde{\mathcal{H}}(m,\gamma) \\ := \left\{ \begin{array}{l} (m,\gamma)\in\mathcal{M}^A\times(\mathcal{P}^{\mathcal{S}})^2:\forall d\in\{0,1\},\ \exists(\varsigma_d,\tilde{Z})\in Sel((\mathbf{S}_d,\tilde{Z}))\cap\bar{I},\ \gamma_d\overset{d}{=}\varsigma_d, \\ \forall u\in\{-1,1\}:\ \ um_d(s)\le u\mu_d(s)\pi_{\gamma_d}(s)+h_{co(\mathcal{Y})}(u)(1-\pi_{\gamma_d}(s)),\ \gamma_d\text{−a.e.} \end{array} \right\} \tag{26}$$

I then show that $\tilde{\mathcal{H}}(m,\gamma)$ is equivalent to (12). Observing that $\varsigma_d\in Sel(\mathbf{S}_d)\cap I$ can be equivalently stated as $(\varsigma_d,\tilde{Z})\in Sel(\mathbf{S}_d,\tilde{Z})\cap\bar{I}$, by Lemma 1:

$$\mathcal{H}(m,\gamma)=\left\{ \begin{array}{l} (m,\gamma)\in\mathcal{M}^A\times(\mathcal{P}^{\mathcal{S}})^2:\forall d\in\{0,1\},\ \ \exists(\varsigma_d,\tilde{Z})\in Sel((\mathbf{S}_d,\tilde{Z}))\cap\bar{I}, \\ \exists\upsilon_d\in Sel(\mathbf{Y}_d),\ \ \gamma_d\overset{d}{=}\varsigma_d,\ \ m_d(\varsigma_d)=E_O[\upsilon_d|\varsigma_d]\text{ a.s.} \end{array} \right\} \tag{27}$$

noting that $\mathcal{M}^A$ is such that for any (m,γ) under consideration, each m_d must be γ_d-integrable since $E[|Y(d)|]<\infty$ implies $\int|m_d|d\gamma_d=E[|E[Y(d)|S(d)]|]\le E[|Y(d)|]<\infty$. It is immediate that then also:

$$\mathcal{H}(m,\gamma)=\left\{ \begin{array}{l} (m,\gamma)\in\mathcal{M}^A\times(\mathcal{P}^{\mathcal{S}})^2:\forall d\in\{0,1\},\ \ \exists(\varsigma_d,\tilde{Z})\in Sel((\mathbf{S}_d,\tilde{Z}))\cap\bar{I}, \\ \exists\upsilon_d\in Sel^1(\mathbf{Y}_d),\ \ \gamma_d\overset{d}{=}\varsigma_d,\ \ m_d(\varsigma_d)=E_O[\upsilon_d|\varsigma_d]\text{ a.s.} \end{array} \right\}. \tag{28}$$

<u>$\mathcal{H}(m,\gamma)=\tilde{\mathcal{H}}(m,\gamma)$</u>

To show $\mathcal{H}(m,\gamma)\subseteq\tilde{\mathcal{H}}(m,\gamma)$, fix $(m,\gamma)\in\mathcal{H}(m,\gamma)$ and $d\in\{0,1\}$. Then there exist $(\varsigma_d,\tilde{Z})\in Sel((\mathbf{S}_d,\tilde{Z}))\cap\bar{I}$ and $\upsilon_d\in Sel^1(\mathbf{Y}_d)$ such that $\gamma_d\overset{d}{=}\varsigma_d$ and $m_d(\varsigma_d)=E_O[\upsilon_d\mid\varsigma_d]$ a.s. Let $\sigma(\varsigma_d|G=O)$ be the sub-σ-algebra generated by ς_d given $\{\omega\in\Omega:G=O\}$. Let $\mathbb{E}_O[\mathbf{Y}_d|\varsigma_d]:=cl\{E_O[\upsilon_d|\varsigma_d]:\upsilon_d\in Sel^1(\mathbf{Y}_d)\}$, where the closure is taken in L^1 space of all $\sigma(\varsigma_d|G=O)$-measurable functions. $\mathbb{E}_O[\mathbf{Y}_d|\varsigma_d]$ exists, is a unique random set, and has at least one integrable selection (Molchanov (2017, Theorem 2.1.71)). By definition of $\mathbb{E}_O[\mathbf{Y}_d|\varsigma_d]$, it is immediate that:

$$\exists\upsilon_d\in Sel^1(\mathbf{Y}_d):\ \ m_d(\varsigma_d)=E_O[\upsilon_d|\varsigma_d]\text{ a.s.}\Rightarrow m_d(\varsigma_d)\in\mathbb{E}_O[\mathbf{Y}_d|\varsigma_d]\text{ a.s.} \tag{29}$$

By assumption, the probability space is non-atomic. By Lemma S.2, P has no atoms over $\sigma(\varsigma_d|G=O)$ for any measurable selection ς_d. Since $E[|Y(d)|]<\infty$ for all $d\in\{0,1\}$, $\mathbf{Y}_d$ is integrable. Thus, $\mathbb{E}_O[\mathbf{Y}_d|\varsigma_d]$ is almost surely convex and equal to $\mathbb{E}_O[co(\mathbf{Y}_d)|\varsigma_d]$ (Molchanov (2017, Theorem 2.1.77)). Therefore, $h_{\mathbb{E}_O[\mathbf{Y}_d|\varsigma_d]}(u)=h_{\mathbb{E}_O[co(\mathbf{Y}_d)|\varsigma_d]}(u)$ a.s. for all $u\in\mathbb{R}$ by definition of the support function h. By $\mathbb{E}_O[co(\mathbf{Y}_d)|\varsigma_d]=\mathbb{E}_O[\mathbf{Y}_d|\varsigma_d]$ and integrability of the latter, the former set is also integrable. It then follows that $h_{\mathbb{E}_O[co(\mathbf{Y}_d)|\varsigma_d]}(u)=E_O[h_{co(\mathbf{Y}_d)}(u)|\varsigma_d]$ a.s. for all $u\in\mathbb{R}$ (Molchanov (2017, Theorem 2.1.72)). Hence, recalling that $\mathbb{E}_O[co(\mathbf{Y}_d)|\varsigma_d]=\mathbb{E}_O[\mathbf{Y}_d|\varsigma_d]$, also $h_{\mathbb{E}_O[\mathbf{Y}_d|\varsigma_d]}(u) = h_{\mathbb{E}_O[co(\mathbf{Y}_d)|\varsigma_d]}(u) = E_O[h_{co(\mathbf{Y}_d)}(u)|\varsigma_d]$ a.s. for all $u \in \mathbb{R}$. Then:

$$\begin{aligned} m_d(\varsigma_d) \in \mathbb{E}_O[\mathbf{Y}_d|\varsigma_d] \text{ a.s.} \Leftrightarrow & \forall u \in \{-1,1\}: \quad u m_d(\varsigma_d) \leq h_{\mathbb{E}_O[\mathbf{Y}_d|\varsigma_d]}(u) \text{ a.s.} \\ \Leftrightarrow & \forall u \in \{-1,1\}: \quad u m_d(\varsigma_d) \leq E_O[h_{co(\mathbf{Y}_d)}(u)|\varsigma_d] \text{ a.s.} \end{aligned} \tag{30}$$

where the first line is by Rockafellar (1970, Theorem 13.1) and almost sure convexity of $\mathbb{E}_O[\mathbf{Y}_d|\varsigma_d]$, and the second is by $h_{\mathbb{E}_O[\mathbf{Y}_d|\varsigma_d]}(u) = E_O[h_{co(\mathbf{Y}_d)}(u)|\varsigma_d]$ a.s. for all $u \in \mathbb{R}$. Moreover:

$$\begin{aligned} E_O[h_{co(\mathbf{Y}_d)}(u)|\varsigma_d] &= E_O[h_{co(\mathbf{Y}_d)}(u)|\varsigma_d, D = d] P_O(D = d|\varsigma_d) \\ &\quad + E_O[h_{co(\mathbf{Y}_d)}(u)|\varsigma_d, D \neq d] P_O(D \neq d|\varsigma_d) \\ &= u E_O[Y|\varsigma_d, D = d] P_O(D = d|\varsigma_d) + h_{co(\mathcal{Y})}(u) P_O(D \neq d|\varsigma_d) \\ &= u E_O[Y|S, D = d] P_O(D = d|\varsigma_d) + h_{co(\mathcal{Y})}(u) P_O(D \neq d|\varsigma_d) \\ &= u \mu_d(\varsigma_d) P_O(D = d|\varsigma_d) + h_{co(\mathcal{Y})}(u) P_O(D \neq d|\varsigma_d) \\ &= u \mu_d(\varsigma_d) \pi_{\gamma_d}(\varsigma_d) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(\varsigma_d)) \end{aligned} \tag{31}$$

where the first equality is by LIE. The second follows because $co(\mathbf{Y}_d) = \{Y\}$ whenever $D = d$, $h_{\{Y\}}(u) = uY$, and $co(\mathbf{Y}_d) = co(\mathcal{Y})$ when $D \neq d$. The third is by observing that $P_O(\varsigma_d = S|D = d) = 1$ since $\varsigma_d \in Sel(\mathbf{S}_d)$ and $\mathbf{S}_d = \{S\}$ when $D = d$. The fourth is by definition of μ_d and $P_O(\varsigma_d = S|D = d) = 1$. The final equality is by Lemma S.6. Then observe that:

$$\begin{aligned} & \forall u \in \{-1,1\}: \quad u m_d(\varsigma_d) \leq E_O[h_{co(\mathbf{Y}_d)}(u)|\varsigma_d] \text{ a.s.} \\ \Leftrightarrow & \forall u \in \{-1,1\}: \quad u m_d(\varsigma_d) \leq u \mu_d(\varsigma_d) \pi_{\gamma_d}(\varsigma_d) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(\varsigma_d)) \text{ a.s.} \\ \Leftrightarrow & \forall u \in \{-1,1\}: \quad u m_d(s) \leq u \mu_d(s) \pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s)) \ \gamma_d\text{−a.e.} \end{aligned} \tag{32}$$

where the second line follows by (31) and the third by $\varsigma_d \overset{d}{=} \gamma_d$. Since ς_d was an arbitrary selection such that $(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}$ and $\varsigma_d \overset{d}{=} \gamma_d$ for any $d \in \{0,1\}$, by (29), (30) and (32) then $\mathcal{H}(m,\gamma) \subseteq \tilde{\mathcal{H}}(m,\gamma)$.

For the converse, for any $d \in \{0,1\}$ fix any ς_d such that $(\varsigma_d, \tilde{Z}) \in Sel(\mathbf{S}_d) \times \{\tilde{Z}\} \cap \bar{I}$, letting $\varsigma_d \overset{d}{=} \gamma_d$. Let $m \in \mathcal{M}^A$ be a pair of arbitrary m_d such that for $d \in \{0,1\}$:

$$\forall u \in \{-1,1\}: \quad u m_d(s) \leq u \mu_d(s) \pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s)) \ \gamma_d\text{−a.e.} \tag{33}$$

Since $E[|Y(d)|] < \infty$ and $P(G = O) > 0$, $E_O[|Y(d)|] < \infty$. Given that $\mathcal{M}^A$ must be such that m_d is γ_d integrable for any $d \in \{0,1\}$, by Lemma 2 then for $d \in \{0,1\}$ there exist $\upsilon_d \in Sel^1(\mathbf{Y}_d)$ such that $m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d]$ a.s. It is then immediate that $\tilde{\mathcal{H}}(m,\gamma) \subseteq \mathcal{H}(m,\gamma)$.

<u>$\tilde{\mathcal{H}}(m,\gamma)$ is equivalent to (12)</u>

By Artstein (1983, Theorem 2.1), a distribution function characterizes a selection in $Sel((\mathbf{S}_d, \tilde{Z}))$ if and only if:

$$\forall B \in \mathcal{C}(\mathcal{S} \times \tilde{\mathcal{Z}}): \quad P((S(d), \tilde{Z}) \in B) \geq P((\mathbf{S}_d, \tilde{Z}) \subseteq B) \tag{34}$$

$$\Leftrightarrow \forall B \in \mathcal{C}(\mathcal{S}): \quad P(S(d) \in B|\tilde{Z}) \geq P(\mathbf{S}_d \subseteq B|\tilde{Z}) \text{ a.s.} \tag{35}$$

where the second line follows by Molchanov and Molinari (2018, Theorem 2.33). Now consider the containment functional $P(\mathbf{S}_d \subseteq B|\tilde{Z})$. If $B = \mathcal{S}$, $P(\mathbf{S}_d \subseteq B|\tilde{Z}) = 1$. If $B \subsetneq \mathcal{S}$, then $P(\mathbf{S}_d \subseteq B|\tilde{Z}) = P(S \in B, D = d|\tilde{Z})$. Hence, $\exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z}))$ such that $\gamma_d \in \mathcal{P}^{\mathcal{S}}$ and $\gamma_d \overset{d}{=} \varsigma_d$ if and only if:

$$\forall B \in \mathcal{C}(\mathcal{S}): \quad P(\varsigma_d \in B|\tilde{Z}) \geq P(S \in B, D = d|\tilde{Z}) \text{ a.s.} \tag{36}$$

Since $(\varsigma_d, \tilde{Z}) \in \bar{I}$, (36) is equivalent to:

$$\begin{aligned} \forall B \in \mathcal{C}(\mathcal{S}): \quad P(\varsigma_d \in B) &\geq ess \sup_{\tilde{Z}} P(S \in B, D = d|\tilde{Z}) && (37) \\ &= \max\left(ess \sup_{Z} P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right) && (38) \end{aligned}$$

where the first line follows since, by definition of $\bar{I}$, $\varsigma_d \perp\!\!\!\perp \tilde{Z}$. The second is by definition of $\tilde{Z}$ and $P(G = O) > 0$ given that two datasets are observed. Therefore:

$$\begin{aligned} &\exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I} \text{ s.t. } \gamma_d \overset{d}{=} \varsigma_d \\ \Leftrightarrow &\forall B \in \mathcal{C}(\mathcal{S}): \quad P(\varsigma_d \in B) \geq \max\left(ess \sup_{Z} P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right) \end{aligned} \tag{39}$$

By definition of $\tilde{\mathcal{H}}(m, \gamma)$ and (39):

$$\tilde{\mathcal{H}}(m, \gamma) = \left\{ \begin{array}{l} (m, \gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0, 1\}, \forall B \in \mathcal{C}(\mathcal{S}), \\ \gamma_d(B) \geq \max\left(ess \sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ \forall u \in \{-1, 1\}: \quad u m_d(s) \leq u \mu_d(s) \pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s)), \ \gamma_d\text{–a.e.} \end{array} \right\} \tag{40}$$

By its definition, if $\mathfrak{C}$ is a core determining class, (39) is equivalent to:

$$\begin{aligned} &\exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I} \text{ s.t. } \gamma_d \overset{d}{=} \varsigma_d \\ \Leftrightarrow &\forall B \in \mathfrak{C}: \quad P(\varsigma_d \in B) \geq \max\left(ess \sup_{Z} P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right) \end{aligned} \tag{41}$$

It is then immediate that in (40) the condition $\forall B \in \mathcal{C}(\mathcal{S})$ can be replaced by $\forall B \in \mathfrak{C}$.

The result then follows by $\mathcal{H}(m,\gamma) = \tilde{\mathcal{H}}(m,\gamma)$. $\square$

Proof of Theorem 2. Recall $\mathcal{H}(\tau) = \{T(m,\gamma) : (m,\gamma) \in \mathcal{H}(m,\gamma)\}$. The proof proceeds in three steps.

Step 1: $\mathcal{H}(m,\gamma)$ is convex.

Take any $(m,\gamma), (m',\gamma') \in \mathcal{H}(m,\gamma)$ and any $a \in (0,1)$. Define:

$$m^a := am + (1-a)m', \qquad \gamma^a := a\gamma + (1-a)\gamma'. \tag{42}$$

Since $\mathcal{M}^A$ is convex by assumption, $m^a \in \mathcal{M}^A$. $\mathcal{P}^{\mathcal{S}}$ is the space of probability measures and is therefore convex, so $\gamma^a_d \in \mathcal{P}^{\mathcal{S}}$ for each d. It remains to verify that set constraints on m_d and γ_d from Theorem 1 are satisfied for $d \in \{0,1\}$.

For γ_d constraints, note that for any $d \in \{0,1\}$ and $B \in \mathcal{C}(\mathcal{S})$:

$$\begin{aligned}
\gamma^a_d(B) &= a\gamma_d(B) + (1-a)\gamma'_d(B) \\
&\geq a \max\Big(ess\sup_Z P_E(S \in B, D = d \mid Z), P_O(S \in B, D = d)\Big) \\
&\quad + (1-a)\max\Big(ess\sup_Z P_E(S \in B, D = d \mid Z), P_O(S \in B, D = d)\Big) \\
&= \max\Big(ess\sup_Z P_E(S \in B, D = d \mid Z), P_O(S \in B, D = d)\Big),
\end{aligned} \tag{43}$$

where the first line is by definition of γ^a, the second is by $(m,\gamma), (m',\gamma') \in \mathcal{H}(m,\gamma)$ and Theorem 1, and the third is by observation.

For the m_d constraint, fix any $d \in \{0,1\}$. For any B with $\gamma_d(B) = 0$, $P_O(S \in B, D = d) \leq \gamma_d(B) = 0$. Hence, $P_O(S \in \cdot, D = d) \ll \gamma_d$, and $P_O(S \in \cdot, D = d) \ll \gamma'_d$. Since $\gamma^a_d = a\gamma_d + (1-a)\gamma'_d$, also $P_O(S \in \cdot, D = d) \ll \gamma^a_d$. Then by the Radon-Nikodym theorem there exist measurable functions:

$$\pi_{\gamma_d} := \frac{dP_O(S \in \cdot, D = d)}{d\gamma_d}, \qquad \pi_{\gamma'_d} := \frac{dP_O(S \in \cdot, D = d)}{d\gamma'_d}, \qquad \pi^a := \frac{dP_O(S \in \cdot, D = d)}{d\gamma^a_d}. \tag{44}$$

Since $\gamma^a_d = a\gamma_d + (1-a)\gamma'_d$ with $a \in (0,1)$, $\gamma_d \ll \gamma^a_d$ and $\gamma'_d \ll \gamma^a_d$. Therefore, similarly, the following exist:

$$\rho := \frac{d\gamma_d}{d\gamma^a_d}, \qquad \rho' := \frac{d\gamma'_d}{d\gamma^a_d}. \tag{45}$$

Then $\rho, \rho' \geq 0$ by nonnegativity of measures, and:

$$a\rho + (1-a)\rho' = 1 \qquad \gamma_d^a - \text{a.e.} \tag{46}$$

by $\gamma_d^a = a\gamma_d + (1-a)\gamma_d'$ and linearity of the Radon–Nikodym derivative. By the chain rule for Radon–Nikodym derivatives, set:

$$\pi^a = \pi_{\gamma_d}\rho = \pi_{\gamma_d'}\rho' \qquad \gamma_d^a - \text{a.e.} \tag{47}$$

with the convention $\pi_{\gamma_d} = 0$ on the event $\{\rho = 0\}$ and similarly $\pi_{\gamma_d'} = 0$ on $\{\rho' = 0\}$.

Since $(m, \gamma) \in \mathcal{H}(m, \gamma)$, for each $u \in \{-1, 1\}$:

$$um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)\big(1 - \pi_{\gamma_d}(s)\big) \qquad \gamma_d - \text{a.e.} \tag{48}$$

by Theorem 1, and similarly with (m', γ'). Since (48) holds γ_d-a.e. and $\gamma_d(B) = \int_B \rho(s) d\gamma_d^a(s)$ for every measurable B, it follows that (48) also holds γ_d^a-a.e. on $\{\rho > 0\}$. Analogously, the corresponding inequality for (m_d', γ_d') holds γ_d^a-a.e. on $\{\rho' > 0\}$.

Fix $u \in \{-1, 1\}$. On $\{\rho > 0\} \cap \{\rho' > 0\}$, by (47) rewrite (48) as:

$$um_d(s) \leq u\mu_d(s)\frac{\pi^a(s)}{\rho(s)} + h_{co(\mathcal{Y})}(u)\Big(1 - \frac{\pi^a(s)}{\rho(s)}\Big) \qquad \gamma_d^a - \text{a.e. on } \{\rho > 0\}, \tag{49}$$

and similarly:

$$um_d'(s) \leq u\mu_d(s)\frac{\pi^a(s)}{\rho'(s)} + h_{co(\mathcal{Y})}(u)\Big(1 - \frac{\pi^a(s)}{\rho'(s)}\Big) \qquad \gamma_d^a - \text{a.e. on } \{\rho' > 0\}. \tag{50}$$

Then define $A(s) := \frac{a}{\rho(s)} + \frac{1-a}{\rho'(s)}$ on the event $\{\rho > 0\} \cap \{\rho' > 0\}$ and note:

$$\begin{aligned}
um_d^a(s) &= aum_d(s) + (1-a)um_d'(s) \\
&\leq a\Big(u\mu_d(s)\frac{\pi^a(s)}{\rho(s)} + h_{co(\mathcal{Y})}(u)\Big(1 - \frac{\pi^a(s)}{\rho(s)}\Big)\Big) \\
&\quad + (1-a)\Big(u\mu_d(s)\frac{\pi^a(s)}{\rho'(s)} + h_{co(\mathcal{Y})}(u)\Big(1 - \frac{\pi^a(s)}{\rho'(s)}\Big)\Big) \\
&= u\mu_d(s)\pi^a(s)\Big(\frac{a}{\rho(s)} + \frac{1-a}{\rho'(s)}\Big) + h_{co(\mathcal{Y})}(u)\Big(1 - \pi^a(s)\Big(\frac{a}{\rho(s)} + \frac{1-a}{\rho'(s)}\Big)\Big) \\
&= u\mu_d(s)\pi^a(s)A(s) + h_{co(\mathcal{Y})}(u)\big(1 - \pi^a(s)A(s)\big), \text{ on } \{\rho > 0\} \cap \{\rho' > 0\}
\end{aligned} \tag{51}$$

where the first line is by definition of m_d^a, the second is by the two inequalities above, the third is by rearrangement and factoring out $\pi^a(s)$, and the fourth is by definition of $A(s)$. Observe that:

$$\begin{aligned}
A(s) - 1 &= \frac{a}{\rho(s)} + \frac{1-a}{\rho'(s)} - 1 \\
&= \frac{a\rho'(s) + (1-a)\rho(s)}{\rho(s)\rho'(s)} - \frac{1}{a\rho(s) + (1-a)\rho'(s)} \\
&= \frac{(a\rho'(s) + (1-a)\rho(s))(a\rho(s) + (1-a)\rho'(s)) - \rho(s)\rho'(s)}{\rho(s)\rho'(s)(a\rho(s) + (1-a)\rho'(s))} \\
&= \frac{a(1-a)(\rho(s) - \rho'(s))^2}{\rho(s)\rho'(s)(a\rho(s) + (1-a)\rho'(s))} \\
&= \frac{a(1-a)(\rho(s) - \rho'(s))^2}{\rho(s)\rho'(s)} \geq 0 \qquad \text{on } \{\rho > 0\} \cap \{\rho' > 0\},
\end{aligned} \tag{52}$$

where the first line is by definition of $A(s)$, the second is by (46), the third and fourth are by observation, and the fifth is by (46). Note that $\mu_d(s) = E_O[Y|S = s, D = d]$ is well-defined and finite $P_O(S \in \cdot|D = d)$−a.e. because $E_O[|Y|\mathbb{1}[D = d]] < \infty$ by assumption. Moreover, since $Y \in \mathcal{Y}$ a.s., $\mu_d(s) \in \text{co}(\mathcal{Y})$ wherever it is defined and hence $u\mu_d(s) \leq h_{\text{co}(\mathcal{Y})}(u)$, for $u \in \{-1, 1\}$. Since $A(s) \geq 1$ and $u\mu_d(s) - h_{\text{co}(\mathcal{Y})}(u) \leq 0$, by (51):

$$um_d^a(s) \leq h_{\text{co}(\mathcal{Y})}(u) + (u\mu_d(s) - h_{\text{co}(\mathcal{Y})}(u))\pi^a(s)A(s) \leq u\mu_d(s)\pi^a(s) + h_{\text{co}(\mathcal{Y})}(u)(1 - \pi^a(s)) \tag{53}$$

on $\{\rho > 0\} \cap \{\rho' > 0\}$.

On $\{\rho = 0\} \cup \{\rho' = 0\}$, note that $\gamma_d^a(\{\rho = 0\} \cap \{\rho' = 0\}) = 0$ by (46), so the constraint holds trivially there. On the remaining events $\{\rho > 0\} \cap \{\rho' = 0\}$ and $\{\rho = 0\} \cap \{\rho' > 0\}$, (47) implies $\pi^a = 0$ γ_d^a-a.e., so the constraint reduces to $um_d^a(s) \leq h_{\text{co}(\mathcal{Y})}(u)$. Since $(m, \gamma), (m', \gamma') \in \mathcal{H}(m, \gamma)$ implies $m \in \mathcal{M}^A \subseteq \mathcal{M}$ and $m' \in \mathcal{M}^A \subseteq \mathcal{M}$, we have $m_d(s), m'_d(s) \in \mathcal{Y} \subseteq \text{co}(\mathcal{Y})$ everywhere. Therefore $m_d^a(s) = am_d(s) + (1-a)m'_d(s) \in \text{co}(\mathcal{Y})$ everywhere by convexity, and hence $um_d^a(s) \leq h_{\text{co}(\mathcal{Y})}(u)$ holds everywhere.

Combining the cases $\{\rho > 0\} \cap \{\rho' > 0\}$ and $\{\rho = 0\} \cup \{\rho' = 0\}$, for every $u \in \{-1, 1\}$:

$$um_d^a(s) \leq u\mu_d(s)\pi^a(s) + h_{co(\mathcal{Y})}(u)\big(1 - \pi^a(s)\big) \qquad \gamma_d^a - \text{a.e.} \tag{54}$$

Since $\pi^a = dP_O(S \in \cdot, D = d)/d\gamma_d^a = \pi_{\gamma_d^a}$, the moment constraints hold for (m^a, γ^a). Thus $(m^a, \gamma^a) \in \mathcal{H}(m, \gamma)$ and $\mathcal{H}(m, \gamma)$ is convex.

Step 2: T is continuous on line segments over $\mathcal{H}(m, \gamma)$.

Fix arbitrary $(m^1, \gamma^1), (m^0, \gamma^0) \in \mathcal{H}(m, \gamma)$ and define for $t \in [0, 1]$:

$$m^t := tm^1 + (1-t)m^0, \qquad \gamma^t := t\gamma^1 + (1-t)\gamma^0. \tag{55}$$

By Step 1, $(m^t, \gamma^t) \in \mathcal{H}(m, \gamma)$ for every $t \in [0,1]$, so $T(m^t, \gamma^t)$ is well-defined and finite for every t by $E[|Y(d)|] < \infty$. Fix any $t \in (0,1)$ and $d \in \{0,1\}$. Since $(m^t, \gamma^t) \in \mathcal{H}(m, \gamma)$:

$$\int |m_d^t| d\gamma_d^t < \infty. \tag{56}$$

Then:

$$\int |m_d^t| d\gamma_d^0 \leq \frac{1}{1-t} \int |m_d^t| d\gamma_d^t < \infty, \qquad \int |m_d^t| d\gamma_d^1 \leq \frac{1}{t} \int |m_d^t| d\gamma_d^t < \infty, \tag{57}$$

where the inequalities are by $\gamma_d^t = t\gamma_d^1 + (1-t)\gamma_d^0 \geq (1-t)\gamma_d^0$ and $\gamma_d^t \geq t\gamma_d^1$, and the finiteness is by (56). Moreover, by $|tx + (1-t)y| \geq t|x| - (1-t)|y|$:

$$|m_d^t| = |tm_d^1 + (1-t)m_d^0| \geq t|m_d^1| - (1-t)|m_d^0|. \tag{58}$$

Rearranging yields $t|m_d^1| \leq |m_d^t| + (1-t)|m_d^0|$. Integrating with respect to γ_d^0 gives:

$$t \int |m_d^1| d\gamma_d^0 \leq \int |m_d^t| d\gamma_d^0 + (1-t) \int |m_d^0| d\gamma_d^0 < \infty, \tag{59}$$

where the finiteness is by (57) and $\int |m_d^0| d\gamma_d^0 < \infty$, since $(m^0, \gamma^0) \in \mathcal{H}(m, \gamma)$. Since $t > 0$, $\int |m_d^1| d\gamma_d^0 < \infty$. The argument for $\int |m_d^0| d\gamma_d^1 < \infty$ is analogous. Hence for each d, the integrals $\int m_d^1 d\gamma_d^1$, $\int m_d^1 d\gamma_d^0$, $\int m_d^0 d\gamma_d^1$, $\int m_d^0 d\gamma_d^0$ are finite. Therefore:

$$\begin{aligned} \int m_d^t d\gamma_d^t &= \int (tm_d^1 + (1-t)m_d^0) d(t\gamma_d^1 + (1-t)\gamma_d^0) \\ &= t^2 \int m_d^1 d\gamma_d^1 + t(1-t) \int m_d^1 d\gamma_d^0 + t(1-t) \int m_d^0 d\gamma_d^1 + (1-t)^2 \int m_d^0 d\gamma_d^0, \end{aligned} \tag{60}$$

where the first line is by definition of m_d^t and γ_d^t, and the second by bilinearity of the integral and the fact that individual integrals are finite. Then $f(t) := \int m_1^t d\gamma_1^t - \int m_0^t d\gamma_0^t$ is a quadratic polynomial in t, hence continuous on $[0,1]$. Because $f(t) = T(m^t, \gamma^t)$, T is continuous on line segments.

Step 3: $cl(\mathcal{H}(\tau))$ is a closed interval.

Step 1 implies $\mathcal{H}(m, \gamma)$ is convex, hence path-connected by line segments: for any two points $(m, \gamma), (m', \gamma') \in \mathcal{H}(m, \gamma)$, the line segment $\{(m^t, \gamma^t) : t \in [0,1]\}$ lies entirely in $\mathcal{H}(m, \gamma)$. By Step 2, T is continuous along every such line segment. Therefore, for any two values $\tau_0 = T(m, \gamma), \tau_1 = T(m', \gamma') \in \mathcal{H}(\tau)$, the set $\mathcal{H}(\tau)$ contains the continuous image $\{T(m^t, \gamma^t) : t \in [0,1]\}$ of $[0,1]$ connecting τ_0 and τ_1. By the intermediate value theorem, this image contains the whole interval $[\min(\tau_0, \tau_1), \max(\tau_0, \tau_1)]$. Hence $\mathcal{H}(\tau) \subseteq \mathbb{R}$ is an interval. The closure of an interval in $\mathbb{R}$ is the closed interval with endpoints given by its infimum and supremum. Thus:

$$cl(\mathcal{H}(\tau)) = \left[\inf_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}), \sup_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}) \right]. \tag{61}$$

□

Proof of Corollary 1. Since $\mathcal{S} = \{1, 2, \ldots, k\}$, represent γ_d as an element of the $k-$dimensional simplex $\Delta(k)$ and $m_d \in \mathcal{Y}^k$. Let $\gamma_d(s)$ and $m_d(s)$ denote the $s-$th element of the corresponding vectors. Then:

$$\begin{aligned}\mathcal{H}(m,\gamma) &= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\}, \forall B \in \mathcal{C}(\mathcal{S}), \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S\in B, D=d|Z), P_O(S\in B, D=d)\right), \\ \forall u \in \{-1,1\}\text{: } um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1-\pi_{\gamma_d}(s))\ \gamma_d\text{−a.e.} \end{array} \right\} \\ &= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\}, \forall B \in \mathcal{C}(\mathcal{S}), \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S\in B, D=d|Z), P_O(S\in B, D=d)\right), \forall u \in \{-1,1\}\text{:} \\ um_d(s) \leq u\mu_d(s)\frac{P_O(S=s,D=d)}{\gamma_d(s)} + h_{co(\mathcal{Y})}(u)\left(1 - \frac{P_O(S=s,D=d)}{\gamma_d(s)}\right)\ \gamma_d\text{−a.e.} \end{array} \right\} \end{aligned} \tag{62}$$

where the first line is by Theorem 1. The second is by definition of $\pi_{\gamma_d}(s)$ and γ_d being supported on $\mathcal{S}$ with $|\mathcal{S}| < \infty$. $\mathcal{S}$ is closed by definition. Since it is finite, it is bounded. Hence, $\mathbf{S}_d$ is almost surely compact, by definition. Then, by Beresteanu, Molchanov, and Molinari (2012, Lemma

B.1) $\{\{s\} : s \in \mathcal{S}\}$ is a core-determining class for the containment functional of $\mathbf{S}_d$. Then:

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\}, \forall s \in \mathcal{S}, \forall u \in \{-1,1\}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s, D=d|Z), P_O(S=s,D=d)\right), \\ um_d(s) \leq u\mu_d(s)\frac{P_O(S=s,D=d)}{\gamma_d(s)} + h_{co(\mathcal{Y})}(u)\left(1 - \frac{P_O(S=s,D=d)}{\gamma_d(s)}\right) \gamma_d\text{−a.e.} \end{array} \right\}$$
$$= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\}, \forall s \in \mathcal{S}, \forall u \in \{-1,1\}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s, D=d|Z), P_O(S=s,D=d)\right), \\ um_d(s) \leq uE[Y|S=s,D=d]\frac{P_O(S=s,D=d)}{\gamma_d(s)} + h_{co(\mathcal{Y})}(u)\left(1 - \frac{P_O(S=s,D=d)}{\gamma_d(s)}\right) \gamma_d\text{−a.e.} \end{array} \right\}$$
$$= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\}, \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s, D=d|Z), P_O(S=s,D=d)\right), \\ (m_d(s) - \inf \mathcal{Y})\,\gamma_d(s) \geq (E_O[Y|S=s,D=d] - \inf\mathcal{Y})\,P_O(S=s,D=d)\ \gamma_d\text{−a.e.}, \\ (\sup\mathcal{Y} - m_d(s))\,\gamma_d(s) \geq (\sup\mathcal{Y} - E_O[Y|S=s,D=d])\,P_O(S=s,D=d)\ \gamma_d\text{−a.e.} \end{array} \right\}$$
$$= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\}, \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s, D=d|Z), P_O(S=s,D=d)\right), \\ (m_d(s) - \inf \mathcal{Y})\,\gamma_d(s) \geq (E_O[Y|S=s,D=d] - \inf\mathcal{Y})\,P_O(S=s,D=d), \\ (\sup\mathcal{Y} - m_d(s))\,\gamma_d(s) \geq (\sup\mathcal{Y} - E_O[Y|S=s,D=d])\,P_O(S=s,D=d) \end{array} \right\}. \tag{63}$$

where the first line is by Theorem 1 and (62), the second line is by definition of $\mu_d(s)$, the third is by definition of $h_{co(\mathcal{Y})}(u)$ and rearrangement, and the fourth is by observation. The result then follows from Theorem 1 and Theorem 2. □

Lemma 2. *Let $\bar{I}$ be the set of random elements (E_1, E_2) such that $(E_1, E_2) \in \mathcal{S}\times\tilde{\mathcal{Z}}$ and $E_1 \perp\!\!\!\perp E_2$. Suppose that $E_O[|Y(d)|] < \infty$ for any $d \in \{0,1\}$. For any (m,γ) such that $\forall d \in \{0,1\}$: $\exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}$ with $\gamma_d \overset{d}{=} \varsigma_d$, and m_d that is γ_d integrable satisfying:*

$$\forall u \in \{-1,1\}: \ \ um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1-\pi_{\gamma_d}(s)), \ \gamma_d\text{−}a.e. \tag{64}$$

there exist $\upsilon_d \in Sel^1(\mathbf{Y}_d)$ such that $m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d]$ a.s. for $d \in \{0,1\}$.

Proof. Fix any (m,γ) satisfying the conditions. For $d \in \{0,1\}$, let $(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}$ be such that $\varsigma_d \overset{d}{=} \gamma_d$. Note that $P_O(\varsigma_d \in \cdot) = \gamma_d(\cdot)$ since $\varsigma_d \overset{d}{=} \gamma_d$ and $(\varsigma_d, \tilde{Z}) \in \bar{I}$. Recall that $\mu_d(s) = E_O[Y|S=s, D=d] = E_O[Y|\varsigma_d = s, D=d]$ γ_d−a.e., where the second equality follows by observing that $P_O(\varsigma_d = S|D=d) = 1$ since $\varsigma_d \in Sel(\mathbf{S}_d)$ and $\mathbf{S}_d = \{S\}$ when $D=d$. $E_O[|Y\mathbb{1}[D=d]|] \in \mathbb{R}$ since $E_O[|Y\mathbb{1}[D=d]|] \leq E_O[|Y(d)|] < \infty$ where the final inequality is by assumption. Also, γ_d−a.e:

$$E_O\big(Y\mathbb{1}[D=d]|\varsigma_d=s\big)=E_O(Y|\varsigma_d=s,D=d)P_O(D=d|\varsigma_d=s)=\mu_d(s)\pi_{\gamma_d}(s), \tag{65}$$

and thus $\mu_d(s)\pi_{\gamma_d}(s)\in\mathbb{R}$ γ_d−a.e. Fix an arbitrary $y_0\in\mathcal{Y}$. Define:

$$c_d(s):=\begin{cases}\dfrac{m_d(s)-\pi_{\gamma_d}(s)\mu_d(s)}{1-\pi_{\gamma_d}(s)}, & \text{if } \pi_{\gamma_d}(s)<1,\\ y_0, & \text{if } \pi_{\gamma_d}(s)=1.\end{cases} \tag{66}$$

Note that $c_d(s)\in\mathbb{R}$ γ_d−a.e., since $c_d(s)=y_0\in\mathbb{R}$ on $\{\pi_{\gamma_d}=1\}$, and on $\{\pi_{\gamma_d}<1\}$ m_d is finite γ_d−a.e. and $\mu_d(s)\pi_{\gamma_d}(s)\in\mathbb{R}$ γ_d−a.e. By (64), $\mu_d(s)=m_d(s)$ whenever $\pi_{\gamma_d}(s)=1$, so then:

$$(1-\pi_{\gamma_d})c_d=m_d-\pi_{\gamma_d}\mu_d \qquad \gamma_d-\text{a.e.} \tag{67}$$

For any s such that $\pi_{\gamma_d}(s)<1$ note that $\forall u\in\{-1,1\}$:

$$um_d(s)\leq u\mu_d(s)\pi_{\gamma_d}(s)+h_{co(\mathcal{Y})}(u)(1-\pi_{\gamma_d}(s))\Rightarrow uc_d(s)\leq h_{co(\mathcal{Y})}(u). \tag{68}$$

Hence, by Rockafellar (1970, Theorem 13.1), $c_d(s)\in co(\mathcal{Y})$ whenever $\pi_{\gamma_d}(s)<1$. Thus also $c_d(s)\in co(\mathcal{Y})$, since $c_d(s)=y_0\in\mathcal{Y}$ when $\pi_{\gamma_d}(s)=1$. Define for $t\in\mathbb{R}$:

$$y_-(t):=\sup(\mathcal{Y}\cap(-\infty,t]), \qquad y_+(t):=\inf(\mathcal{Y}\cap[t,\infty)). \tag{69}$$

For $t\in co(\mathcal{Y})\cap\mathbb{R}$, $y_-(t),y_+(t)\in\mathcal{Y}$ and $y_-(t)\leq t\leq y_+(t)$. Moreover, y_- and y_+ are monotone, hence Borel measurable. $\mathcal{Y}$ is closed, by definition. Fix values:

$$a:=\begin{cases}\max(\mathcal{Y}\cap(-\infty,0]), & \text{if } \mathcal{Y}\cap(-\infty,0]\neq\emptyset,\\ \min\mathcal{Y}, & \text{otherwise,}\end{cases} \quad b:=\begin{cases}\min(\mathcal{Y}\cap[0,\infty)), & \text{if } \mathcal{Y}\cap[0,\infty)\neq\emptyset,\\ \max\mathcal{Y}, & \text{otherwise.}\end{cases} \tag{70}$$

Let $C_{\mathcal{Y}}:=|a|+|b|$. Observe that $a,b\in\mathbb{R}\cap\mathcal{Y}$ and hence $C_{\mathcal{Y}}<\infty$. Define on the event $\{\pi_{\gamma_d}<1\}$:

$$(y_1(s),y_2(s)):=\begin{cases}(y_-(c_d(s)),a), & \text{if } c_d(s)\leq a,\\ (a,b), & \text{if } a<c_d(s)<b,\\ (b,y_+(c_d(s))), & \text{if } c_d(s)\geq b,\end{cases} \tag{71}$$

On the event $\{\pi_{\gamma_d}=1\}$, set $y_1(s)=y_2(s)=y_0$. Since $y_-(t),y_+(t)\in\mathcal{Y}$ for any $t\in co(\mathcal{Y})$,

$c_d(s) \in co(\mathcal{Y})$, and $a, b \in \mathbb{R} \cap \mathcal{Y}$, then $P_O(y_i(\varsigma_d) \in \mathcal{Y}) = 1$ for $d \in \{0,1\}$ and $i \in \{1,2\}$. Finally, define:

$$p(s) := \begin{cases} \dfrac{y_2(s) - c_d(s)}{y_2(s) - y_1(s)}, & \text{if } y_2(s) > y_1(s), \\ 1, & \text{if } y_2(s) = y_1(s). \end{cases} \tag{72}$$

Recall that for any $t \in co(\mathcal{Y}) \cap \mathbb{R}$, $y_-(t) \leq t \leq y_+(t)$. Since $c_d(s) \in co(\mathcal{Y})$ and $c_d(s) \in \mathbb{R}$, then it must be that $y_-(c_d(s)) \leq c_d(s) \leq y_+(c_d(s))$ and therefore by (71) $y_1(s) \leq c_d(s) \leq y_2(s)$ γ_d-a.e. on the event $\{\pi_{\gamma_d} < 1\}$. On $\{\pi_{\gamma_d} = 1\}$, $y_1(s) = y_2(s) = c_d(s) = y_0$. Then $p(s) \in [0,1]$ and γ_d-a.e:

$$p(s)y_1(s) + (1 - p(s))y_2(s) = c_d(s). \tag{73}$$

After possibly enlarging the probability space, there exists a random variable $U : \Omega \to [0,1]$ which is uniformly distributed on $[0,1]$ with $U \perp\!\!\!\perp (Y, D, \varsigma_d)|G = O$. Define:

$$\tilde{Y}_d := \Big(y_1(\varsigma_d)\mathbb{1}[U \leq p(\varsigma_d)] + y_2(\varsigma_d)\mathbb{1}[U > p(\varsigma_d)]\Big)\mathbb{1}[G = O] + y_0\mathbb{1}[G = E]. \tag{74}$$

Since $P_O(y_i(\varsigma_d) \in \mathcal{Y}) = 1$ for $d \in \{0,1\}$ and $i \in \{1,2\}$, it is immediate that $\tilde{Y}_d \in \mathcal{Y}$ a.s. It is also direct that $\tilde{Y}_d \perp\!\!\!\perp D|\varsigma_d, G = O$ since $U \perp\!\!\!\perp (Y, D, \varsigma_d)|G = O$. Then:

$$E_O[\tilde{Y}_d|\varsigma_d] = p(\varsigma_d)y_1(\varsigma_d) + (1 - p(\varsigma_d))y_2(\varsigma_d) = c_d(\varsigma_d) \qquad \text{a.s.} \tag{75}$$

where the first equality is by $U \perp\!\!\!\perp (Y, D, \varsigma_d)|G = O$ and the definition of $\tilde{Y}_d$, and the second is by (73). Finally, define:

$$\upsilon_d := Y\mathbb{1}[D = d, G = O] + \tilde{Y}_d\mathbb{1}[D \neq d \vee G \neq O]. \tag{76}$$

By construction, $\upsilon_d = Y$ on $\{D = d, G = O\}$ and $\upsilon_d \in \mathcal{Y}$ on $\{D \neq d \vee G \neq O\}$, hence $\upsilon_d \in Sel(\mathbf{Y}_d)$. Then γ_d-a.e.:

$$\begin{aligned} E_O[\upsilon_d|\varsigma_d = s] &= E_O\big[Y\mathbb{1}[D = d]|\varsigma_d = s\big] + E_O\big[\tilde{Y}_d\mathbb{1}[D \neq d]|\varsigma_d = s\big] \\ &= \pi_{\gamma_d}(s)\mu_d(s) + (1 - \pi_{\gamma_d}(s))c_d(s) = m_d(s), \end{aligned} \tag{77}$$

where the first equality is by observation, the second is by (65), $\tilde{Y}_d \perp\!\!\!\perp D|\varsigma_d, G = O$ and (75), and the last equality is by (67). Therefore, $m_d(\varsigma_d) = E_O(\upsilon_d|\varsigma_d)$ a.s.

It remains to show that $E[|\upsilon_d|] < \infty$ so that $\upsilon_d \in Sel^1(\mathbf{Y}_d)$. To that end, observe that whenever $c_d(s) \leq a$ or $c_d(s) \geq b$, $y_1(s)$ and $y_2(s)$ have the same sign. If $\mathcal{Y} \cap (-\infty, 0] = \emptyset$, then $\min(\mathcal{Y}) > 0$ and by definition $a = \min \mathcal{Y}$ and $b = \min(\mathcal{Y} \cap [0, \infty)) = \min \mathcal{Y}$. Hence $a = b > 0$ and $\mathcal{Y} \subseteq [a, \infty) \subset (0, \infty)$. Therefore, since $y_1(s), y_2(s) \in \mathcal{Y}$, $y_1(s) \geq 0$ and $y_2(s) \geq 0$. If $\mathcal{Y} \cap [0, \infty) = \emptyset$, then $\max(\mathcal{Y}) < 0$ and similarly $\mathcal{Y} \subset (-\infty, 0)$, so $y_1(s) \leq 0$ and $y_2(s) \leq 0$. If both $\mathcal{Y} \cap (-\infty, 0] \neq \emptyset$ and $\mathcal{Y} \cap [0, \infty) \neq \emptyset$, then by definition $a = \max(\mathcal{Y} \cap (-\infty, 0]) \leq 0$ and $b = \min(\mathcal{Y} \cap [0, \infty)) \geq 0$. When $c_d(s) \leq a$, $(y_1(s), y_2(s)) = (y_-(c_d(s)), a)$, and since $y_-(c_d(s)) \leq c_d(s) \leq a \leq 0$ then $y_1(s) \leq 0$ and $y_2(s) \leq 0$. When $c_d(s) \geq b$, $(y_1(s), y_2(s)) = (b, y_+(c_d(s)))$, and since $y_+(c_d(s)) \geq c_d(s) \geq b \geq 0$ then $y_1(s) \geq 0$ and $y_2(s) \geq 0$.

Hence, for s with $\pi_{\gamma_d}(s) < 1$:

$$E_O[|\tilde{Y}_d| | \varsigma_d = s] \leq |c_d(s)| + C_{\mathcal{Y}}, \tag{78}$$

because if $c_d(s) \leq a$ or $c_d(s) \geq b$ then $y_1(s), y_2(s)$ have the same sign and $E_O[|\tilde{Y}_d| | \varsigma_d = s] = |E_O[\tilde{Y}_d | \varsigma_d = s]| = |c_d(s)|$; if $a < c_d(s) < b$ then $\tilde{Y}_d \in \{a, b\}$ so $E_O[|\tilde{Y}_d| | \varsigma_d = s] \leq |a| + |b| = C_{\mathcal{Y}}$. Then:

$$\begin{aligned}
E_O\big[|\tilde{Y}_d| \mathbb{1}[D \neq d]\big] &= E_O\Big[(1 - \pi_{\gamma_d}(\varsigma_d)) E_O[|\tilde{Y}_d| | \varsigma_d]\Big] \\
&\leq \int (1 - \pi_{\gamma_d}) |c_d| d\gamma_d + C_{\mathcal{Y}} P_O(D \neq d) \\
&= \int |m_d - \pi_{\gamma_d} \mu_d| d\gamma_d + C_{\mathcal{Y}} P_O(D \neq d) \\
&\leq \int |m_d| d\gamma_d + \int |\pi_{\gamma_d} \mu_d| d\gamma_d + C_{\mathcal{Y}} P_O(D \neq d) \\
&= \int |m_d| d\gamma_d + \int \big|E_O(Y \mathbb{1}[D = d] | \varsigma_d)\big| d\gamma_d + C_{\mathcal{Y}} P_O(D \neq d) \\
&\leq \int |m_d| d\gamma_d + E_O\big(|Y| \mathbb{1}[D = d]\big) + C_{\mathcal{Y}} P_O(D \neq d) < \infty
\end{aligned} \tag{79}$$

where the first line is by LIE, the second uses $(1 - \pi_{\gamma_d}(\varsigma_d)) = 0$ on $\{\pi_{\gamma_d}(\varsigma_d) = 1\}$ and (78) on $\{\pi_{\gamma_d}(\varsigma_d) < 1\}$, the third by (67) and $\pi_{\gamma_d} \in [0, 1]$, the fourth by the triangle inequality, the fifth by (65), the sixth by Jensen's inequality and LIE, and the final since m_d is γ_d integrable, $E_O[|Y \mathbb{1}[D = d]|] \leq E_O[|Y(d)|] < \infty$, and $C_{\mathcal{Y}} < \infty$. Then by (76), (79) and $E_O[|Y \mathbb{1}[D = d]|] < \infty$, $E_O[|\upsilon_d|] < \infty$. Moreover, $E_E[|\upsilon_d|] = |y_0| < \infty$ and hence $E[|\upsilon_d|] < \infty$ so $\upsilon_d \in Sel^1(\mathbf{Y}_d)$. $\square$

Identification of Long-Term Treatment Effects via Temporal Links, Observational, and Experimental Data

Supplemental Appendix

Filip Obradović*

This supplement: i) develops a consistent estimation procedure; ii) discusses computational simplifications; and iii) collects auxiliary technical results used in the main text.

S.1 Estimation

This section develops a consistent estimator building on Section 5. Suppose that the researcher observes experimental and observational samples $\{(S_j, D_j, Z_j)\}_{j=1}^{n_E}$ and $\{(Y_i, S_i, D_i)\}_{i=1}^{n_O}$, respectively. Define $n := \min\{n_O, n_E\}$. Let $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ be the subvector of (m, γ) that may be partially identified under maintained assumptions, and $\mathcal{H}(\theta)$ its identified set. Let ϑ be a (possibly empty) subvector collecting components of (m, γ) point identified by assumption.

The identified set $\mathcal{H}(\theta)$ is characterized by a collection of finitely many inequality and equality constraints, under appropriate structure on $\mathcal{M}^A$. Recall that $k = |\mathcal{S}|$ and let μ_d be a k-dimensional vector with components $\mu_d(s) = E_O[Y|S = s, D = d]$. Let η_d be a $k \times (|\mathcal{Z}| + 1)$ matrix with the element (s, z) being $\eta_d(s, z) = P_E(S = s, D = d|Z = z)$ for $z \leq |\mathcal{Z}|$ and $\eta_d(s, z) = P_O(S = s, D = d)$ for $z = |\mathcal{Z}| + 1$. Collect $\beta = (\mu_0, \mu_1, \eta_0, \eta_1, \vartheta, \tilde{\beta}) \in \mathfrak{B}$, where $\tilde{\beta} \in \tilde{\mathfrak{B}}$ is a (possibly empty) vector of other identified features of the population distribution defining $\mathcal{H}(\theta)$. For some known functions $\tilde{h}$ and $\tilde{g}$, one can then write:

$$\mathcal{H}(\theta) = \left\{\theta \in \Theta : \tilde{h}(\theta, \beta) \geq 0, \;\; \tilde{g}(\theta, \beta) = 0\right\}. \tag{1}$$

Example 1. Suppose that the researcher imposes Assumption LIV, which allows partial identification of m, and that γ may be partially identified due to imperfect compliance. Then $\theta = (m, \gamma)$, ϑ is empty and:

$$\mathcal{H}(\theta) = \mathcal{H}(m, \gamma) = \left\{ \begin{array}{l} (m, \gamma) \in \mathcal{Y}^{2k} \times (\Delta(k))^2 : \forall d \in \{0, 1\}, \forall s \in \mathcal{S}, \\ m_d(s') - m_d(s) \geq 0 \text{ for } s' \geq s \text{ in the product order}, \\ \gamma_d(s) - ess\sup_z \eta_d(s, z) \geq 0, \\ (m_d(s) - \inf \mathcal{Y})\, \gamma_d(s) - (\mu_d(s) - \inf \mathcal{Y})\, \eta_d(s, |\mathcal{Z}| + 1) \geq 0\,, \\ (\sup \mathcal{Y} - m_d(s))\, \gamma_d(s) - (\sup \mathcal{Y} - \mu_d(s))\, \eta_d(s, |\mathcal{Z}| + 1) \geq 0 \end{array} \right\}. \tag{2}$$

*UCLA, Department of Economics. Email: obradovicfilip@ucla.edu

If one additionally assumes that datasets jointly point-identify γ, such as under perfect compliance, then $\theta = m$ and $\vartheta = (\gamma_0, \gamma_1)$ where $\gamma_d(s) = ess\sup_z \eta_d(s,z)$ for any $d \in \{0,1\}$ and $s \in \mathcal{S}$. Additionally, $\mathcal{H}(m,\gamma) = \mathcal{H}(\theta) \times \{\gamma\}$ with $\mathcal{H}(\theta)$ defined by restrictions as in (2) with $\gamma_d(s) - ess\sup_z \eta_d(s,z) \geq 0$ omitted. If one imposes TI instead of LIV, constraints $m_0(s) - m_1(s) = 0$ replace $m_d(s') - m_d(s) \geq 0$ in the definitions of $\mathcal{H}(\theta)$.

Example 2. Suppose that the researcher imposes Assumption LUC from Athey, Chetty, and Imbens (2025), which point identifies m, and uses the framework to account for imperfect compliance, so that γ may be partially identified. Then $\theta = \gamma$, and $\vartheta = (m_0, m_1)$ with $m_d(s) = E_O[Y|S=s, D=d]$ for any $d \in \{0,1\}$ and $s \in \mathcal{S}$. Additionally:

$$\mathcal{H}(\theta) = \mathcal{H}(\gamma) = \left\{ \begin{array}{l} \gamma \in (\Delta(k))^2 : \forall d \in \{0,1\}, \forall s \in \mathcal{S}, \\ \gamma_d(s) - ess\sup_z \eta_d(s,z) \geq 0, \end{array} \right\}. \tag{3}$$

The identified set $\mathcal{H}(\theta)$ can be equivalently represented via a criterion function. Define:

$$Q(\theta,\beta) := \|\tilde{g}(\theta,\beta)\|_1 + \sum_{t=1}^{q} \max\left(-\tilde{h}_t(\theta,\beta), 0\right), \tag{4}$$

where q denotes the number of inequality constraints in $\tilde{h}(\theta,\beta)$. If the assumptions hold, $\mathcal{H}(\theta) \neq \emptyset$, so $\bar{Q} := \min_{\theta\in\Theta} Q(\theta,\beta) = 0$ and $\mathcal{H}(\theta) = \operatorname{argmin}_{\theta\in\Theta} Q(\theta,\beta) = \{\theta \in \Theta : Q(\theta,\beta) = 0\}$. Writing $T(\theta,\vartheta)$ for $T(m,\gamma)$ evaluated at the (m,γ) implied by (θ,ϑ):

$$\begin{aligned} \min_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m,\gamma) &= \min_{\theta\in\mathcal{H}(\theta)} T(\theta,\vartheta) = \min_{\theta\in\Theta} T(\theta,\vartheta) \quad \text{s.t.} \quad Q(\theta,\beta) \leq \bar{Q}, \\ \max_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m,\gamma) &= \max_{\theta\in\mathcal{H}(\theta)} T(\theta,\vartheta) = \max_{\theta\in\Theta} T(\theta,\vartheta) \quad \text{s.t.} \quad Q(\theta,\beta) \leq \bar{Q}. \end{aligned} \tag{5}$$

Let β_n be a consistent estimator of β obtained from the empirical analogs, and let ϑ_n denote the subvector of β_n corresponding to ϑ. Denote by $\bar{Q}_n := \min_{\theta\in\Theta} Q(\theta,\beta_n)$. The criterion estimator of $\mathcal{H}(\tau)$ is:

$$\begin{aligned} &\mathcal{H}_n(\tau) := \left[\tau_n^{LB}, \tau_n^{UB}\right], \\ &\tau_n^{LB} = \min_{\theta\in\Theta} T(\theta,\vartheta_n) \quad \text{s.t.} \quad Q(\theta,\beta_n) \leq \bar{Q}_n, \\ &\tau_n^{UB} = \max_{\theta\in\Theta} T(\theta,\vartheta_n) \quad \text{s.t.} \quad Q(\theta,\beta_n) \leq \bar{Q}_n. \end{aligned} \tag{6}$$

To establish Hausdorff consistency without requiring a tuning parameter, I introduce additional notation. Let $\mathcal{H}^{ie}(\theta) := \{\theta \in \Theta : \tilde{h}(\theta,\beta) \geq 0\}$ denote the inequality-constrained parameter space. Modeling assumptions involving only equalities, such as Assumptions TI and LUC do not affect $\mathcal{H}^{ie}(\theta)$ since they are collected by $\tilde{g}(\theta,\beta)=0$. Maintain the following assumption.

Assumption E. *(Estimation)*

i) $\{(S_j,D_j,Z_j)\}_{j=1}^{n_E}$ *and* $\{(Y_i,S_i,D_i)\}_{i=1}^{n_O}$ *are i.i.d. samples;*

ii) $\mathcal{Y}$ *is bounded and* $|\mathcal{S}|,|\mathcal{Z}|<\infty$*;*

iii) $\mathcal{M}^A$ *is defined through finitely many linear equality and weak inequality constraints which may depend on a consistently estimable vector of population parameters* $\beta\in\mathfrak{B}$ *where* $\mathfrak{B}$ *is compact. The Jacobian of all linear equality constraints, including any simplex equalities adjoined by reformulation, has full row rank.*

iv) $\mathrm{cl}\left(\mathrm{int}(\mathcal{H}^{ie}(\theta))\cap\mathcal{H}(\theta)\right)=\mathcal{H}(\theta)$ *or* $\mathcal{H}(\theta)$ *is a singleton.*

Assumption E *i)* is standard under random sampling. Condition *ii)* maintains that long-term outcomes are bounded and that short-term potential outcomes and the instrument Z are finitely supported. Assumption E *iii)* defines the class of modeling assumptions compatible with the estimation procedure. It is sufficiently general to encompass all previously stated modeling assumptions, but may be further weakened to allow for continuously differentiable restrictions on m if necessary. Condition *iv)* is a mild condition introduced by Shi and Shum (2015) that enables consistent estimation without requiring a tuning parameter. For example, it holds when $\mathrm{int}(\mathcal{H}(\theta))\neq\emptyset$, i.e. when components of θ are partially identified, or when $\mathcal{H}(\theta)$ is in the interior of $\mathcal{H}^{ie}(\theta)$. It may be further relaxed, as explained by Shi and Shum (2015, Section 2). Applying the arguments of Shi and Shum (2015, Theorem 2.1) yields the following.

Theorem S.1. *Let Assumptions RA, EV, MA, and E hold. Then as* $n\to\infty$*:*

$$d_H(\mathcal{H}_n(\tau),\mathcal{H}(\tau)):=\max\left\{\sup_{\tau_0\in\mathcal{H}(\tau)}\inf_{\hat{\tau}\in\mathcal{H}_n(\tau)}||\tau_0-\hat{\tau}||,\sup_{\hat{\tau}\in\mathcal{H}_n(\tau)}\inf_{\tau_0\in\mathcal{H}(\tau)}||\tau_0-\hat{\tau}||\right\}\xrightarrow{\text{P}}0.$$

Remark 1. Since τ is a subvector of (m,γ), inference on τ may be recast as a subvector inference problem based on an appropriate criterion function. This would enable the use of general methods for subvector inference, such as those in Bugni, Canay, and Shi (2017), to construct confidence sets for τ and test hypotheses about it.

S.1.1 Reducing Computational Complexity of Bilinear Programming

Exploiting the structure of the identified set $\mathcal{H}(m,\gamma)$ and the objective T may further reduce the computational burden of non-convex optimization.

First, the optimization problems may be separable into lower-dimensional subproblems. Solving the individual subproblems may be less resource-intensive than solving the original problem jointly (Nocedal and Wright 1999). This is feasible when the modeling assumption yields a rectangular set $\mathcal{M}^A = \mathcal{M}_0^A \times \mathcal{M}_1^A$ for some $\mathcal{M}_0^A$ and $\mathcal{M}_1^A$. Since the remaining constraints on (m, γ) are also separable in d, it follows immediately that the identified set $\mathcal{H}(m, \gamma)$ is rectangular as well. Letting $\mathcal{T}(m_d, \gamma_d) := \int_{\mathcal{S}} m_d(s) d\gamma_d(s)$, we have:

$$\begin{aligned}
\min_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}) &= \min_{(\tilde{m}_1,\tilde{\gamma}_1)\in\mathcal{H}(m_1,\gamma_1)} \mathcal{T}(\tilde{m}_1,\tilde{\gamma}_1) - \max_{(\tilde{m}_0,\tilde{\gamma}_0)\in\mathcal{H}(m_0,\gamma_0)} \mathcal{T}(\tilde{m}_0,\tilde{\gamma}_0) \\
\max_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}) &= \max_{(\tilde{m}_1,\tilde{\gamma}_1)\in\mathcal{H}(m_1,\gamma_1)} \mathcal{T}(\tilde{m}_1,\tilde{\gamma}_1) - \min_{(\tilde{m}_0,\tilde{\gamma}_0)\in\mathcal{H}(m_0,\gamma_0)} \mathcal{T}(\tilde{m}_0,\tilde{\gamma}_0)
\end{aligned} \tag{7}$$

where $\mathcal{H}(m_d, \gamma_d)$ collects all constraints on (m_d, γ_d) with $m_d \in \mathcal{M}_d^A$. For example, $\mathcal{M}^A$ is rectangular whenever the modeling assumption does not relate values of m_1 and m_0, such as with Assumptions LIV and LUC.

Second, the optimization problems become linear in certain cases. They are therefore convex, which substantially simplifies computation. As suggested by examples in the previous section, this occurs in two cases. First, if $\mathcal{H}(m, \gamma) = \{m\} \times \mathcal{H}(\gamma)$, the problems reduce to linear programs over γ. For example, assumptions that point-identify m independently of γ, such as LUC, yield $\mathcal{H}(m, \gamma)$ of this form. Second, if $\mathcal{H}(m, \gamma) = \mathcal{H}(m) \times \{\gamma\}$ and $\mathcal{M}^A$ is representable by linear constraints, the problems reduce to linear programs over m. This arises under Assumptions LIV and TI when the right-hand sides of the constraints for each γ_d sum to one. Section 6 exploits both separability and linearization to simplify computation in the application.

Third, for each feasible γ, the inner optimization over m may admit a closed-form solution. In such cases, fixing m appropriately can reduce the dimension of the parameter space explored by branch-and-bound algorithms. To formalize this idea, rewrite the optimization problems as bilevel programs. Decompose $\mathcal{H}(m, \gamma)$ into its projection $\mathcal{H}(\gamma) := \{\gamma' : \exists m' \text{ s.t. } (m', \gamma') \in \mathcal{H}(m, \gamma)\}$ and corresponding fibers $\mathcal{H}(m|\gamma') := \{m' : (m', \gamma') \in \mathcal{H}(m, \gamma)\}$ at each $\gamma' \in \mathcal{H}(\gamma)$. These fibers form a correspondence $\mathcal{H}(m|\cdot) : \mathcal{H}(\gamma) \rightrightarrows \mathcal{M}^A$. The identified set can then be written as:

$$\mathcal{H}(\tau) = \left[\min_{\tilde{\gamma}\in\mathcal{H}(\gamma)} \min_{\tilde{m}\in\mathcal{H}(m|\tilde{\gamma})} T(\tilde{m},\tilde{\gamma}), \max_{\tilde{\gamma}\in\mathcal{H}(\gamma)} \max_{\tilde{m}\in\mathcal{H}(m|\tilde{\gamma})} T(\tilde{m},\tilde{\gamma}) \right]. \tag{8}$$

The inner optimization problems may have closed-form solutions given by selectors of the correspondence $\mathcal{H}(m|\cdot)$. This is formalized by the following definition.

Definition 1 (Minimal and maximal selectors)**.** Let $\mathcal{H}(m|\cdot) : \mathcal{H}(\gamma) \rightrightarrows \mathcal{M}^A$ be the correspondence defined by the fibers of $\mathcal{H}(m, \gamma)$ over its projection $\mathcal{H}(\gamma)$. A selector $L : \mathcal{H}(\gamma) \to \mathcal{M}^A$ is *minimal* *with respect to* T if, for every $\gamma \in \mathcal{H}(\gamma)$, $T(L(\gamma),\gamma) \leq T(m,\gamma)$ for all $m \in \mathcal{H}(m|\gamma)$. A selector $U : \mathcal{H}(\gamma) \to \mathcal{M}^A$ is *maximal with respect to* T if, for every $\gamma \in \mathcal{H}(\gamma)$, $T(U(\gamma),\gamma) \geq T(m,\gamma)$ for all $m \in \mathcal{H}(m|\gamma)$.

If minimal and maximal selectors exist, using them will yield the identified set $\mathcal{H}(\tau)$ since:

$$\left[\min_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}), \max_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma})\right] = \left[\min_{\tilde{\gamma}\in\mathcal{H}(\gamma)} T(L_{\tilde{\gamma}},\tilde{\gamma}), \max_{\tilde{\gamma}\in\mathcal{H}(\gamma)} T(U_{\tilde{\gamma}},\tilde{\gamma})\right].$$

To operationalize the result, consider the lower bound $\min_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma})$. In the minimization problem, one can replace the constraints $m \in \mathcal{M}^A$ in Corollary 1 with $m(s) = L_\gamma(s)$ for each $s \in \mathcal{S}$, thereby obtaining the lower bound without separately optimizing over m for each γ. Similarly, in the maximization problem, one can replace $m \in \mathcal{M}^A$ with $m(s) = U_\gamma(s)$ to obtain the upper bound. Unlike the previous two cases, this simplification may be useful for estimation only when $\bar{Q}_n = 0$, or equivalently, when the plug-in estimator is non-empty.[1] Lemma S.9 derives minimal and maximal selectors under Assumptions LIV and TI.

S.2 Proofs and Auxiliary Lemmas

Proof of Theorem S.1. The proof proceeds in three steps. Step 1 reduces the problem to Hausdorff consistency of $\mathcal{H}_n(\theta)$. Step 2 reformulates the criterion using slack variables and establishes Hausdorff consistency of the resulting argmin set. Step 3 shows that this is sufficient via Step 1.

Step 1. *Sufficiency of* $d_H(\mathcal{H}_n(\theta),\mathcal{H}(\theta)) \xrightarrow{\text{p}} 0$.

By Assumption E *ii)*, $|\mathcal{S}| < \infty$. Thus, (m,γ) is a finite-dimensional vector and T is jointly continuous in (m,γ). Since $\mathcal{H}(m,\gamma)$ is defined by finitely many weak inequalities and equalities, it is closed. Moreover, $\mathcal{H}(m,\gamma) \subseteq \mathcal{Y}^{2k} \times (\Delta(k))^2$, and the latter set is bounded because $\mathcal{Y}$ is bounded by Assumption E *ii)*. Therefore, $\mathcal{H}(m,\gamma)$ is compact. Hence $\mathcal{H}(\tau) = \left[\min_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}), \max_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma})\right]$ is a closed interval. The same holds for $\mathcal{H}_n(\tau)$ by construction. Since the Hausdorff distance between closed intervals $[a,b]$ and $[c,d]$ equals $\max\{|a-c|,|b-d|\}$, it suffices to show that the endpoints converge in probability. I treat the upper bound; the lower bound is symmetric.

By boundedness of $\mathcal{Y}$, each component of μ_d lies in a compact interval and each component of η_d lies in $[0,1]$, by definition. By the definition of Θ and boundedness of $\mathcal{Y}$, Θ is a product of finitely many compact sets, and hence compact. By Assumption E *iii)*, $\mathfrak{B}$ is compact. Therefore

1. The plug-in and criterion estimators are numerically equivalent when $\bar{Q}_n = 0$. The plug-in is also Hausdorff consistent for $\mathcal{H}(\theta)$ under Assumption E since $\bar{Q}_n = 0$ w.p.a. 1.

$\Theta \times \mathfrak{B}$, and hence $\Theta \times \text{proj}_\vartheta(\mathfrak{B})$, are compact. Since $T(\theta, \vartheta)$ is jointly continuous on $\Theta \times \text{proj}_\vartheta(\mathfrak{B})$, the Heine-Cantor theorem implies that T is uniformly continuous on this domain.

Fix any $\varepsilon > 0$. $\tau_n^{UB} = \max_{\theta \in \mathcal{H}_n(\theta)} T(\theta, \vartheta_n)$ and $\tau^{UB} = \max_{\theta \in \mathcal{H}(\theta)} T(\theta, \vartheta)$. By the triangle inequality:

$$|\tau_n^{UB} - \tau^{UB}| \le \underbrace{\max_{\theta \in \Theta} |T(\theta, \vartheta_n) - T(\theta, \vartheta)|}_{=:\rho_n} + \underbrace{\left| \max_{\theta \in \mathcal{H}_n(\theta)} T(\theta, \vartheta) - \max_{\theta \in \mathcal{H}(\theta)} T(\theta, \vartheta) \right|}_{=:\Delta_n}. \tag{9}$$

Since T is uniformly continuous on $\Theta \times \text{proj}_\vartheta(\mathfrak{B})$, for any $\varepsilon > 0$ there exists $\delta > 0$ such that for any $\vartheta' \in \text{proj}_\vartheta(\mathfrak{B})$: $\|\vartheta' - \vartheta\| < \delta \Rightarrow \max_{\theta \in \Theta} |T(\theta, \vartheta') - T(\theta, \vartheta)| < \varepsilon$. On the event $\{\|\vartheta_n - \vartheta\| < \delta\}$ this implies $\rho_n < \varepsilon$. Therefore: $P(\rho_n > \varepsilon) \le P(\|\vartheta_n - \vartheta\| \ge \delta) \to 0$, since $\vartheta_n \xrightarrow{p} \vartheta$. Hence $\rho_n \xrightarrow{p} 0$.

For Δ_n, fix any $\varepsilon' > 0$ and choose $\delta > 0$ such that $\|\theta - \theta'\| < \delta \Rightarrow |T(\theta, \vartheta) - T(\theta', \vartheta)| < \varepsilon'$ for all $\theta, \theta' \in \Theta$. Since $\mathcal{H}_n(\theta)$ and $\mathcal{H}(\theta)$ are compact and $T(\cdot, \vartheta)$ is continuous, the maxima are attained. If $d_H(\mathcal{H}_n(\theta), \mathcal{H}(\theta)) < \delta$, then for any $\theta \in \mathcal{H}_n(\theta)$ there exists $\theta' \in \mathcal{H}(\theta)$ with $\|\theta - \theta'\| < \delta$, hence $T(\theta, \vartheta) \le T(\theta', \vartheta) + \varepsilon' \le \max_{\theta^* \in \mathcal{H}(\theta)} T(\theta^*, \vartheta) + \varepsilon'$. Taking the maximum over $\theta \in \mathcal{H}_n(\theta)$:

$$\max_{\theta \in \mathcal{H}_n(\theta)} T(\theta, \vartheta) \le \max_{\theta \in \mathcal{H}(\theta)} T(\theta, \vartheta) + \varepsilon'.$$

The reverse inequality follows by the same argument interchanging $\mathcal{H}_n(\theta)$ and $\mathcal{H}(\theta)$. Therefore:

$$\Delta_n = \left| \max_{\theta \in \mathcal{H}_n(\theta)} T(\theta, \vartheta) - \max_{\theta \in \mathcal{H}(\theta)} T(\theta, \vartheta) \right| \le \varepsilon'$$

whenever $d_H(\mathcal{H}_n(\theta), \mathcal{H}(\theta)) < \delta$. Taking $\varepsilon' = \varepsilon/2$ implies:

$$P(\Delta_n > \tfrac{\varepsilon}{2}) \le P(d_H(\mathcal{H}_n(\theta), \mathcal{H}(\theta)) \ge \delta). \tag{10}$$

Combining (9) and (10):

$$P(|\tau_n^{UB} - \tau^{UB}| > \varepsilon) \le P(\rho_n > \tfrac{\varepsilon}{2}) + P(d_H(\mathcal{H}_n(\theta), \mathcal{H}(\theta)) \ge \delta).$$

Thus, for $d_H(\mathcal{H}_n(\tau), \mathcal{H}(\tau)) \xrightarrow{p} 0$, it is sufficient to show $d_H(\mathcal{H}_n(\theta), \mathcal{H}(\theta)) \xrightarrow{p} 0$.

Step 2. *Reformulation and Hausdorff consistency.*

To apply Lemma A.1 of Shi and Shum (2015), I reformulate the problem using slack variables. Convert all weak inequality constraints $\tilde{h}(\theta, \beta) \ge 0$ to equalities by introducing $\lambda \in \mathbb{R}_+^q$ so that $\tilde{h}(\theta, \beta) - \lambda = 0$. Denote by $\tilde{\theta} := (\theta, \lambda) \in \mathfrak{T}$ the augmented parameter vector and define $g(\tilde{\theta}, \beta) := \begin{pmatrix} \tilde{h}(\theta, \beta) - \lambda \\ \tilde{g}(\theta, \beta) \end{pmatrix}$. When Θ is defined using simplices, as when θ includes γ, replace each

$\Delta(k)$ with $[0,1]^k$ in Θ and adjoin the simplex equalities $\sum_s \gamma_d(s) - 1 = 0$ for $d \in \{0,1\}$ to $\tilde{g}$ as additional rows. This does not alter the feasible set in θ-space, since $\gamma \in (\Delta(k))^2$ if and only if $\gamma \in [0,1]^{2k}$ and the simplex equalities hold. When Θ is already a product of compact intervals, no reformulation is needed.

Next, absorb the constraints defining Θ into $\tilde{h}$ by appending $\theta_i - a_i \geq 0$ and $b_i - \theta_i \geq 0$ for each coordinate i of θ, where $[a_i, b_i]$ denotes the corresponding factor of Θ. The slack variables for these additional constraints are handled identically. Then enlarge each factor of Θ to $[a_i - 1, b_i + 1]$. Since the original box constraints are now enforced through $\tilde{h}$, this reformulation does not alter the feasible set in θ-space and ensures $\mathcal{H}(\theta) \subset \operatorname{int}(\Theta)$.

By Assumption E *iii)*, any restrictions implied by $\mathcal{M}^A$ can be written as finitely many linear equalities and weak inequalities in θ whose coefficients may depend on β. Together with $|\mathcal{S}| < \infty$ and $|\mathcal{Z}| < \infty$ from Assumption E *ii)*, this implies that each component of $\tilde{h}(\theta, \beta)$ is obtained from finitely many linear or bilinear expressions in θ and finitely many coordinates of β. Therefore $(\theta, \beta) \mapsto \tilde{h}(\theta, \beta)$ is continuous on $\Theta \times \mathfrak{B}$. Since $\Theta \times \mathfrak{B}$ is compact, each component of $\tilde{h}$ is bounded above on $\Theta \times \mathfrak{B}$. Henceforth, let q denote the total number of inequality constraints, including the appended box constraints. Define $\bar{h} \in \mathbb{R}^q$ with components $\bar{h}_t := \sup_{(\theta,\beta)\in\Theta\times\mathfrak{B}} \tilde{h}_t(\theta, \beta) < \infty$ and set $\Lambda := \prod_{t=1}^q [0, \bar{h}_t + 1]$. Then any feasible (θ, λ) satisfying $\tilde{h}(\theta, \beta) - \lambda = 0$ for some $\beta \in \mathfrak{B}$ necessarily has $\lambda_t = \tilde{h}_t(\theta, \beta) \geq 0$ for each t, since feasibility requires $\tilde{h}_t(\theta, \beta) \geq 0$. Moreover, $\lambda_t = \tilde{h}_t(\theta, \beta) \leq \bar{h}_t < \bar{h}_t + 1$ for each t. Therefore $\lambda \in \Lambda$, and imposing $\lambda \in \Lambda$ is without loss of generality.

The augmented parameter space $\mathfrak{T} := \Theta \times \Lambda$ is a product of compact intervals, and thus compact, with $\operatorname{cl}(\operatorname{int}(\mathfrak{T})) = \mathfrak{T}$. For each $\theta \in \Theta$ and $\beta' \in \mathfrak{B}$, minimizing over $\lambda \in \Lambda$ yields:

$$\min_{\lambda\in\Lambda} \|g((\theta, \lambda), \beta')\|_1 = \|\tilde{g}(\theta, \beta')\|_1 + \sum_{t=1}^{q} \max(-\tilde{h}_t(\theta, \beta'), 0) = Q(\theta, \beta'),$$

because $\min_{\lambda_t\in[0,\bar{h}_t+1]} |\tilde{h}_t(\theta, \beta') - \lambda_t| = \max(-\tilde{h}_t(\theta, \beta'), 0)$ for each t. Let $\bar{Q}_n := \min_{\tilde{\theta}\in\mathfrak{T}} \|g(\tilde{\theta}, \beta_n)\|_1$. The minimum $\bar{Q}_n$ exists by continuity of $\|g(\cdot, \beta_n)\|_1$ on the compact set $\mathfrak{T}$, so $\tilde{\Theta}_n$ is nonempty and compact. Define:

$$\tilde{\Theta} := \{\tilde{\theta} \in \mathfrak{T} : g(\tilde{\theta}, \beta) = 0\}, \quad \tilde{\Theta}_n := \operatorname*{argmin}_{\tilde{\theta}\in\mathfrak{T}} \|g(\tilde{\theta}, \beta_n)\|_1 = \{\tilde{\theta} \in \mathfrak{T} : \|g(\tilde{\theta}, \beta_n)\|_1 = \bar{Q}_n\}.$$

It follows that $\mathcal{H}_n(\theta)$ and $\mathcal{H}(\theta)$ are the projections of the augmented sets onto θ-coordinates $\mathcal{H}(\theta) = \left\{\theta \in \Theta : \exists\lambda \in \Lambda \text{ s.t. } (\theta, \lambda) \in \tilde{\Theta}\right\}$, and $\mathcal{H}_n(\theta) = \left\{\theta \in \Theta : \exists\lambda \in \Lambda \text{ s.t. } (\theta, \lambda) \in \tilde{\Theta}_n\right\}$. I then establish $d_H(\tilde{\Theta}_n, \tilde{\Theta}) \xrightarrow{p} 0$ in two parts.

Step 2a. $\sup_{\tilde{\theta}\in\tilde{\Theta}_n} \inf_{\tilde{\theta}_0\in\tilde{\Theta}} \|\tilde{\theta} - \tilde{\theta}_0\| \xrightarrow{p} 0.$

By the reverse triangle inequality:

$$\sup_{\tilde{\theta}\in\mathfrak{T}}\left|\|g(\tilde{\theta},\beta_n)\|_1-\|g(\tilde{\theta},\beta)\|_1\right|\leq\sup_{\tilde{\theta}\in\mathfrak{T}}\|g(\tilde{\theta},\beta_n)-g(\tilde{\theta},\beta)\|_1=:r_n. \tag{11}$$

The function g is polynomial in $(\tilde{\theta},\beta)$ and hence jointly continuous. Since it is jointly continuous on the compact set $\mathfrak{T}\times\mathfrak{B}$, it is uniformly continuous. By Assumption E *i)* to *iii)*, $\beta_n\xrightarrow{\text{p}}\beta$. Hence $r_n\xrightarrow{\text{p}}0$. Fix any $\tilde{\theta}_0\in\tilde{\Theta}$. Then:

$$0\leq\bar{Q}_n\leq\|g(\tilde{\theta}_0,\beta_n)\|_1\leq\|g(\tilde{\theta}_0,\beta)\|_1+r_n=r_n\xrightarrow{\text{p}}0,$$

where the first inequality is by non-negativity of Q, the second is by definition of $\bar{Q}_n$, and the last equality uses (11) and $g(\tilde{\theta}_0,\beta)=0$ since $\tilde{\theta}_0\in\tilde{\Theta}$. Thus $\bar{Q}_n\xrightarrow{\text{p}}0$.

Next, fix any $\varepsilon>0$ and define:

$$A_\varepsilon:=\left\{\tilde{\theta}\in\mathfrak{T}:\inf_{\tilde{\theta}_0\in\tilde{\Theta}}\|\tilde{\theta}-\tilde{\theta}_0\|\geq\varepsilon\right\}.$$

Note that $\delta_\varepsilon:=\inf_{\tilde{\theta}\in A_\varepsilon}\|g(\tilde{\theta},\beta)\|_1>0$. Since A_ε is by definition a closed subset of the compact set $\mathfrak{T}$, it is compact, and the continuous function $\|g(\cdot,\beta)\|_1$ attains its infimum on A_ε. If $\delta_\varepsilon=0$, the minimizer $\bar{\tilde{\theta}}\in A_\varepsilon$ satisfies $g(\bar{\tilde{\theta}},\beta)=0$, so $\bar{\tilde{\theta}}\in\tilde{\Theta}$, contradicting $\bar{\tilde{\theta}}\in A_\varepsilon$. Hence $\delta_\varepsilon>0$.

Now consider the event $\{\tilde{\Theta}_n\cap A_\varepsilon\neq\emptyset\}$. On this event, pick any $\tilde{\theta}\in\tilde{\Theta}_n\cap A_\varepsilon$. By definition of δ_ε, $\|g(\tilde{\theta},\beta)\|_1\geq\delta_\varepsilon$. Since $\tilde{\theta}\in\tilde{\Theta}_n$, $\|g(\tilde{\theta},\beta_n)\|_1=\bar{Q}_n$. By (11), $\|g(\tilde{\theta},\beta)\|_1\leq\bar{Q}_n+r_n$. Therefore, $\delta_\varepsilon\leq\bar{Q}_n+r_n$ on the event, and hence $\{\tilde{\Theta}_n\cap A_\varepsilon\neq\emptyset\}\subseteq\{\bar{Q}_n+r_n\geq\delta_\varepsilon\}$. Since $\bar{Q}_n\xrightarrow{\text{p}}0$ and $r_n\xrightarrow{\text{p}}0$, it follows that $\bar{Q}_n+r_n\xrightarrow{\text{p}}0$. Since $\delta_\varepsilon>0$ is fixed $P(\tilde{\Theta}_n\cap A_\varepsilon\neq\emptyset)\leq P(\bar{Q}_n+r_n\geq\delta_\varepsilon)\to0$. Therefore:

$$\begin{aligned}P\left(\sup_{\tilde{\theta}\in\tilde{\Theta}_n}\inf_{\tilde{\theta}_0\in\tilde{\Theta}}\|\tilde{\theta}-\tilde{\theta}_0\|\geq\varepsilon\right)&=P(\tilde{\Theta}_n\cap A_\varepsilon\neq\emptyset)\\&\leq P(\bar{Q}_n+r_n\geq\delta_\varepsilon)\to0.\end{aligned}$$

Step 2b. $\sup_{\tilde{\theta}_0\in\tilde{\Theta}}\inf_{\tilde{\theta}\in\tilde{\Theta}_n}\|\tilde{\theta}-\tilde{\theta}_0\|\xrightarrow{\text{p}}0.$

If $\tilde{\Theta}$ is a singleton, the claim follows immediately from Step 2a. Suppose that $\tilde{\Theta}$ is non-singleton. Define the zero-set correspondence $\tilde{\Theta}^0(\beta'):=\{\tilde{\theta}\in\mathfrak{T}:g(\tilde{\theta},\beta')=0\}$. I verify the conditions of Lemma A.1 of Shi and Shum (2015) for this correspondence at $\beta'=\beta$. By construction, $\mathfrak{T}$ is a product of compact intervals, so $\mathfrak{T}$ is compact and $\text{cl}(\text{int}(\mathfrak{T}))=\mathfrak{T}$. Next, the function g is polynomial in $(\tilde{\theta},\beta)$ and hence continuously differentiable in $\tilde{\theta}$ for each fixed β' in an open set containing $\mathfrak{B}$. Now note that the Jacobian of g with respect to $\tilde{\theta}=(\theta,\lambda)$ has the block structure:

$$\frac{\partial g(\tilde{\theta},\beta)}{\partial \tilde{\theta}'} = \begin{pmatrix} \partial \tilde{h}(\theta,\beta)/\partial\theta' & -I_q \\ \partial \tilde{g}(\theta,\beta)/\partial\theta' & 0 \end{pmatrix},$$

where I_q is the $q \times q$ identity matrix. The $-I_q$ block arises because $\partial(\tilde{h}_t - \lambda_t)/\partial\lambda_t = -1$ and $\partial(\tilde{h}_t - \lambda_t)/\partial\lambda_s = 0$ for $s \neq t$, while the remaining equalities do not involve λ. Since the slackness rows have a $-I_q$ block in the λ-columns while all other rows have zeros there, the slackness rows are linearly independent of each other and of all other rows. Therefore the full Jacobian has full row rank if and only if $\partial\tilde{g}(\theta,\beta)/\partial\theta'$ has full row rank, which holds by Assumption E *iii)*. Finally, take any $(\theta,\lambda) \in \tilde{\Theta}$. By Assumption E *iv)*, there exists a sequence $\{\theta^{(j)}\} \subseteq \text{int}(\mathcal{H}^{ie}(\theta)) \cap \mathcal{H}(\theta)$ with $\theta^{(j)} \to \theta$. Since $\theta^{(j)} \in \text{int}(\mathcal{H}^{ie}(\theta))$, all inequalities hold strictly, so $\tilde{h}(\theta^{(j)},\beta) > 0$ componentwise. Setting $\lambda^{(j)} = \tilde{h}(\theta^{(j)},\beta)$, we have $\lambda^{(j)} \in \text{int}(\Lambda)$ because $0 < \lambda_t^{(j)} = \tilde{h}_t(\theta^{(j)},\beta) \leq \bar{h}_t < \bar{h}_t+1$ for each t. By the enlargement of Θ, $\mathcal{H}(\theta) \subset \text{int}(\Theta)$, so $\theta^{(j)} \in \text{int}(\Theta)$. Therefore $(\theta^{(j)},\lambda^{(j)}) \in \tilde{\Theta} \cap \text{int}(\mathfrak{T})$, and $(\theta^{(j)},\lambda^{(j)}) \to (\theta,\lambda)$ by continuity of $\tilde{h}$. Hence, $\text{cl}(\tilde{\Theta} \cap \text{int}(\mathfrak{T})) = \tilde{\Theta}$.

Lemma A.1 of Shi and Shum (2015) therefore implies that $\tilde{\Theta}^0(\cdot)$ is continuous at β. Fix any $\varepsilon > 0$. By continuity, there exists $\delta > 0$ such that $\|\beta' - \beta\| < \delta$ implies $d_H(\tilde{\Theta}^0(\beta'),\tilde{\Theta}) < \varepsilon$ and $\tilde{\Theta}^0(\beta') \neq \emptyset$. Then:

$$P\left(d_H(\tilde{\Theta}^0(\beta_n),\tilde{\Theta}) \geq \varepsilon\right) + P\left(\tilde{\Theta}^0(\beta_n) = \emptyset\right) \leq 2\,P\left(\|\beta_n - \beta\| \geq \delta\right) \to 0, \tag{12}$$

where the inequality holds because both events on the left are subsets of $\{\|\beta_n - \beta\| \geq \delta\}$ by continuity, and the convergence follows from $\beta_n \xrightarrow{p} \beta$. On the event $\{\tilde{\Theta}^0(\beta_n) \neq \emptyset\}$, $\bar{Q}_n \leq \|g(\tilde{\theta},\beta_n)\|_1 = 0$ for any $\tilde{\theta} \in \tilde{\Theta}^0(\beta_n)$, so $\bar{Q}_n = 0$ and $\tilde{\Theta}_n = \tilde{\Theta}^0(\beta_n)$. Therefore,

$$P\left(d_H(\tilde{\Theta}_n,\tilde{\Theta}) \geq \varepsilon\right) \leq P\left(d_H(\tilde{\Theta}^0(\beta_n),\tilde{\Theta}) \geq \varepsilon\right) + P\left(\tilde{\Theta}^0(\beta_n) = \emptyset\right) \to 0,$$

where the inequality follows because on the event $\{\tilde{\Theta}^0(\beta_n) \neq \emptyset\}$, $\tilde{\Theta}_n = \tilde{\Theta}^0(\beta_n)$ and hence $d_H(\tilde{\Theta}_n,\tilde{\Theta}) = d_H(\tilde{\Theta}^0(\beta_n),\tilde{\Theta})$, so the event $\{d_H(\tilde{\Theta}_n,\tilde{\Theta}) \geq \varepsilon\}$ is contained in $\{d_H(\tilde{\Theta}^0(\beta_n),\tilde{\Theta}) \geq \varepsilon\} \cup \{\tilde{\Theta}^0(\beta_n) = \emptyset\}$, and the convergence from (12). Hence $d_H(\tilde{\Theta}_n,\tilde{\Theta}) \xrightarrow{p} 0$ when $\tilde{\Theta}$ is non-singleton. Steps 2a and 2b together yield $d_H(\tilde{\Theta}_n,\tilde{\Theta}) \xrightarrow{p} 0$.

Step 3. *Hausdorff consistency of $\tilde{\Theta}_n$ implies Hausdorff consistency of $\mathcal{H}_n(\theta)$.*

Let $\pi : \mathfrak{T} \to \Theta$ denote the coordinate projection $\pi(\tilde{\theta}) = \theta$ for $\tilde{\theta} = (\theta,\lambda)$. For any $\tilde{\theta}_1 = (\theta_1,\lambda_1)$ and $\tilde{\theta}_2 = (\theta_2,\lambda_2)$ in $\mathfrak{T}$, $\|\pi(\tilde{\theta}_1) - \pi(\tilde{\theta}_2)\| = \|\theta_1 - \theta_2\| \leq \|(\theta_1,\lambda_1) - (\theta_2,\lambda_2)\| = \|\tilde{\theta}_1 - \tilde{\theta}_2\|$. Thus Hausdorff distance cannot increase under π. Therefore, by Step 2:

$$d_H(\mathcal{H}_n(\theta),\mathcal{H}(\theta)) = d_H(\pi(\tilde{\Theta}_n),\pi(\tilde{\Theta})) \leq d_H(\tilde{\Theta}_n,\tilde{\Theta}) \xrightarrow{p} 0.$$

where the equality follows since, by arguments in Step 2, $\pi(\tilde{\Theta}) = \mathcal{H}(\theta)$ and $\pi(\tilde{\Theta}_n) = \mathcal{H}_n(\theta)$. Step 1 then yields $d_H(\mathcal{H}_n(\tau), \mathcal{H}(\tau)) \xrightarrow{p} 0$.

□

Lemma S.1. *Let $\mathbf{R}_0$ and $\mathbf{R}_1$ be two random closed sets. Then $Sel(\mathbf{R}_0 \times \mathbf{R}_1) = Sel(\mathbf{R}_0) \times Sel(\mathbf{R}_1)$.*

Proof. Fix an arbitrary selection $(\rho_0, \rho_1) \in Sel(\mathbf{R}_0 \times \mathbf{R}_1)$. Then:

$$1 = P\Big((\rho_0, \rho_1) \in \mathbf{R}_0 \times \mathbf{R}_1\Big) = P\Big(\rho_0 \in \mathbf{R}_0, \rho_1 \in \mathbf{R}_1\Big) \leq P\Big(\rho_0 \in \mathbf{R}_0\Big) \tag{13}$$

where the first equality follows by $(\rho_0, \rho_1) \in Sel(\mathbf{R}_0 \times \mathbf{R}_1)$, the second equality and the inequality are by observation. Hence $P(\rho_0 \in \mathbf{R}_0) = 1$. By a similar argument, $P(\rho_1 \in \mathbf{R}_1) = 1$. Therefore $(\rho_0, \rho_1) \in Sel(\mathbf{R}_0) \times Sel(\mathbf{R}_1)$.

Next, fix an arbitrary $(\rho_0, \rho_1) \in Sel(\mathbf{R}_0) \times Sel(\mathbf{R}_1)$. Then:

$$\begin{aligned} 1 = P\Big(\rho_0 \in \mathbf{R}_0\Big) &= P\Big(\rho_0 \in \mathbf{R}_0, \rho_1 \in \mathbf{R}_1\Big) + P\Big(\rho_0 \in \mathbf{R}_0, \rho_1 \notin \mathbf{R}_1\Big) \\ &= P\Big(\rho_0 \in \mathbf{R}_0, \rho_1 \in \mathbf{R}_1\Big) = P\Big((\rho_0, \rho_1) \in \mathbf{R}_0 \times \mathbf{R}_1\Big) \end{aligned} \tag{14}$$

where the first equality is by $\rho_0 \in Sel(\mathbf{R}_0)$, the second is by observation, the third is since $P\Big(\rho_0 \in \mathbf{R}_0, \rho_1 \notin \mathbf{R}_1\Big) \leq P\Big(\rho_1 \notin \mathbf{R}_1\Big) = 0$ given that $\rho_1 \in Sel(\mathbf{R}_1)$, and the fourth is by observation. Thus $(\rho_0, \rho_1) \in Sel(\mathbf{R}_0 \times \mathbf{R}_1)$. □

Lemma S.2. *Suppose the probability space $(\Omega, \mathcal{F}, P)$ is non-atomic and that $\mathcal{F}_0 \subseteq \mathcal{F}$ is a sub-σ-algebra. After possibly enlarging the probability space, P is atomless over $(\Omega, \mathcal{F}_0)$. That is, for all $A \in \mathcal{F}$ with $P(A) > 0$ there exists $B \in \mathcal{F}$ with $B \subseteq A$ such that $0 < P(B|\mathcal{F}_0) < P(A|\mathcal{F}_0)$ with positive probability.*

Proof. Fix $A \in \mathcal{F}$ with $P(A) > 0$. There exists a random variable $U : \Omega \to [0, 1]$ which is uniformly distributed on $[0, 1]$ and independent of the σ–algebra $\sigma(\mathcal{F}_0, \mathbb{1}_A)$.[2]

Define:

$$B := A \cap \{U \leq \tfrac{1}{2}\}. \tag{15}$$

It is immediate that $B \in \mathcal{F}$ and $B \subset A$. By independence of U from $\sigma(\mathcal{F}_0, \mathbb{1}_A)$:

$$P(B|\mathcal{F}_0) = E\left[\mathbb{1}_A \mathbb{1}_{\{U \leq 1/2\}} | \mathcal{F}_0\right] = E\left[\mathbb{1}_A | \mathcal{F}_0\right] E\left[\mathbb{1}_{\{U \leq 1/2\}}\right] = \frac{1}{2} P(A|\mathcal{F}_0) \quad P\text{–a.s.}$$

2. If necessary, replace $(\Omega, \mathcal{F}, P)$ by $(\Omega \times [0, 1], \mathcal{F} \otimes \mathcal{B}([0, 1]), P \otimes \lambda)$ and let $U(\omega, u) = u$.

Because $E[P(A|\mathcal{F}_0)] = P(A) > 0$:

$$P\big(P(A|\mathcal{F}_0) > 0\big) > 0,$$

On the event $\{P(A|\mathcal{F}_0) > 0\}$ it follows $0 < P(B|\mathcal{F}_0) = \frac{1}{2}P(A|\mathcal{F}_0) < P(A|\mathcal{F}_0)$, and thus:

$$\big\{0 < P(B|\mathcal{F}_0) < P(A|\mathcal{F}_0)\big\} \supseteq \big\{P(A|\mathcal{F}_0) > 0\big\}. \tag{16}$$

Therefore $\big\{0 < P(B|\mathcal{F}_0) < P(A|\mathcal{F}_0)\big\}$ has strictly positive probability. □

Lemma S.3. *Suppose that Assumptions RA, EV hold. Then the identified set $\mathcal{H}(P_{Y(0),S(0)}, P_{Y(1),S(1)})$ for marginal distribution functions $P_{Y(d),S(d)}$ is:*

$$\mathcal{H}(P_{Y(0),S(0)}, P_{Y(1),S(1)}) = \mathcal{H}(P_{Y(0),S(0)}) \times \mathcal{H}(P_{Y(1),S(1)}) \neq \emptyset. \tag{17}$$

Let $C(B) = \{s : \mathcal{Y} \times \{s\} \subseteq B\}$. $\mathcal{H}(P_{Y(d),S(d)})$ when combined data are used is:

$$\left\{\delta \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}} : \quad \begin{array}{l} \forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S}), \\ \delta(B) \geq \max\big\{P_O\big((Y,S) \in B, D = d\big), ess\sup_Z P_E\big(S \in C(B), D = d|Z\big)\big\} \end{array} \right\}. \tag{18}$$

If the experimental data are not observed, then $\mathcal{H}(P_{Y(d),S(d)})$ is equivalent to:

$$\{\delta \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}} : \forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S}), \;\; \delta(B) \geq P_O((Y,S) \in B, D = d)\} \tag{19}$$

Proof. The proof proceeds by extending arguments of Beresteanu, Molchanov, and Molinari (2012, Proposition 2.3). Recall that $\tilde{Z} = \mathbb{1}[G = E]Z + \mathbb{1}[G = O](sup\mathcal{Z} + 1) \in \tilde{\mathcal{Z}}$. Note that $\tilde{Z} = Z$ when $G = E$ and $\tilde{Z}$ equals a distinct constant when $G = O$. Therefore, Assumptions RA and EV hold if and only if $\tilde{Z} \perp\!\!\!\perp (Y(d), S(d))$ for all $d \in \{0,1\}$. Let $\tilde{I}$ be the set of random elements (E_1, E_2, E_3) such that $(E_1, E_2, E_3) \in \mathcal{Y} \times \mathcal{S} \times \tilde{\mathcal{Z}}$ and $(E_1, E_2) \perp\!\!\!\perp E_3$, i.e. that satisfy the two assumptions. Define the random set:

$$(\mathbf{Y}_0, \mathbf{S}_0, \mathbf{Y}_1, \mathbf{S}_1) := \begin{cases} \{(Y,S)\} \times \mathcal{Y} \times \mathcal{S}, \text{ if } (D,G) = (0,O) \\ \mathcal{Y} \times \mathcal{S} \times \{(Y,S)\}, \text{ if } (D,G) = (1,O) \\ \mathcal{Y} \times \{S\} \times \mathcal{Y} \times \mathcal{S}, \text{ if } (D,G) = (0,E) \\ \mathcal{Y} \times \mathcal{S} \times \mathcal{Y} \times \{S\}, \text{ if } (D,G) = (1,E) \\ \mathcal{Y} \times \mathcal{S} \times \mathcal{Y} \times \mathcal{S}, \text{ otherwise} \end{cases} . \tag{20}$$

By definition of $(\mathbf{Y}_0, \mathbf{S}_0, \mathbf{Y}_1, \mathbf{S}_1)$, all information on $(Y(0), S(0), Y(1), S(1))$ in the combined data can be summarized by $(Y(0), S(0), Y(1), S(1)) \in Sel((\mathbf{Y}_0, \mathbf{S}_0, \mathbf{Y}_1, \mathbf{S}_1))$. Also, by definition of the random set, $(\mathbf{Y}_0, \mathbf{S}_0, \mathbf{Y}_1, \mathbf{S}_1) = (\mathbf{Y}_0, \mathbf{S}_0) \times (\mathbf{Y}_1, \mathbf{S}_1)$.

By Lemma S.1, $Sel((\mathbf{Y}_0, \mathbf{S}_0, \mathbf{Y}_1, \mathbf{S}_1)) = Sel((\mathbf{Y}_0, \mathbf{S}_0)) \times Sel((\mathbf{Y}_1, \mathbf{S}_1))$. All information in the data and assumptions about $(Y(0), S(0), Y(1), S(1))$ can thus equivalently be expressed as $(Y(d), S(d), \tilde{Z}) \in Sel((\mathbf{Y}_d, \mathbf{S}_d), \tilde{Z}) \cap \tilde{I}$ for $d \in \{0, 1\}$. If Assumptions RA and EV hold, $(Y(d), S(d), \tilde{Z}) \in Sel((\mathbf{Y}_d, \mathbf{S}_d), \tilde{Z}) \cap \tilde{I} \neq \emptyset$ for $d \in \{0, 1\}$. Then:

$$\begin{aligned}
&\mathcal{H}(P_{Y(0),S(0)}, P_{Y(1),S(1)}) \\
&= \left\{(\delta_0, \delta_1) \in \left(\mathcal{P}^{\mathcal{Y}\times\mathcal{S}}\right)^2 : \forall d \in \{0,1\},\ \exists(\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel((\mathbf{Y}_d, \mathbf{S}_d), \tilde{Z}) \cap \tilde{I} \text{ s.t. } \delta_d \overset{d}{=} (\upsilon_d, \varsigma_d)\right\} \\
&= \bigtimes_{d\in\{0,1\}} \left\{\delta_d \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}} : \exists(\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel((\mathbf{Y}_d, \mathbf{S}_d), \tilde{Z}) \cap \tilde{I} \text{ s.t. } \delta_d \overset{d}{=} (\upsilon_d, \varsigma_d)\right\} \\
&= \mathcal{H}(P_{Y(0),S(0)}) \times \mathcal{H}(P_{Y(1),S(1)}) \neq \emptyset
\end{aligned} \tag{21}$$

where the first equality follows by definition, the second by observation, the third is by definition of $\mathcal{H}(P_{Y(d),S(d)})$, and the final since $(Y(d), S(d), \tilde{Z}) \in Sel((\mathbf{Y}_d, \mathbf{S}_d), \tilde{Z}) \cap \tilde{I} \neq \emptyset$ for $d \in \{0, 1\}$, showing (17).

By Artstein (1983, Theorem 2.1), the distribution function $P((Y(d), S(d), \tilde{Z})) \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}\times\tilde{\mathcal{Z}}}$ characterizes a selection in $Sel((\mathbf{Y}_d, \mathbf{S}_d), \tilde{Z})$ if and only if:

$$\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S} \times \tilde{\mathcal{Z}}): \quad P((Y(d), S(d), \tilde{Z}) \in B) \geq P(((\mathbf{Y}_d, \mathbf{S}_d), \tilde{Z}) \subseteq B) \tag{22}$$

By Molchanov and Molinari (2018, Theorem 2.33), (22) is equivalent to:

$$\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S}): \quad P((Y(d), S(d)) \in B|\tilde{Z}) \geq P((\mathbf{Y}_d, \mathbf{S}_d) \subseteq B|\tilde{Z})\ P\text{−a.s.} \tag{23}$$

Possible forms that B can take are: 1) $B = \mathcal{Y} \times \mathcal{S}$;[3] 2) $B \subsetneq \mathcal{Y} \times \mathcal{S}$. For $B = \mathcal{Y} \times \mathcal{S}$, $P((\mathbf{Y}_d, \mathbf{S}_d) \subseteq B|\tilde{Z}) = 1$ P−a.s. Note that the constraint imposed by the containment functional in this case is redundant when $P((Y(d), S(d), \tilde{Z})) \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}\times\tilde{\mathcal{Z}}}$, given that $P((Y(d), S(d)) \in \mathcal{Y} \times \mathcal{S}|\tilde{Z}) = 1$ P−a.s. by definition.

Consider the containment functional for $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ when $B \subsetneq \mathcal{Y} \times \mathcal{S}$. If $D \neq d$, $(\mathbf{Y}_d, \mathbf{S}_d) = \mathcal{Y} \times \mathcal{S}$, so the random set can be a subset of B only if $D = d$. Events $\{G = O, D = d\}$ and $\{G = E, D = d\}$ are mutually exclusive. Hence, for any closed $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ with $B \subsetneq \mathcal{Y} \times \mathcal{S}$,

3. The support of a random vector X is the smallest closed set $\mathcal{X}$ such that $P(X \in \mathcal{X}) = 1$. Hence $\mathcal{Y} \times \mathcal{S} \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$.

$P-$a.s:

$$\begin{aligned} P\big((\mathbf{Y}_d, \mathbf{S}_d) \subseteq B|\tilde{Z}\big) &= P\big(G = O, D = d, (\mathbf{Y}_d, \mathbf{S}_d) \subseteq B|\tilde{Z}\big) + P\big(G = E, D = d, (\mathbf{Y}_d, \mathbf{S}_d) \subseteq B|\tilde{Z}\big) \\ &= P\big(G = O, D = d, (Y, S) \in B|\tilde{Z}\big) + P\big(G = E, D = d, \mathcal{Y} \times \{S\} \subseteq B|\tilde{Z}\big) \\ &= P\big(G = O, D = d, (Y, S) \in B|\tilde{Z}\big) + P\big(G = E, D = d, S \in C(B)|\tilde{Z}\big). \end{aligned}$$

where the first line follows by the definition of the random set and the fact that events $\{G = O, D = d\}$ and $\{G = E, D = d\}$ are mutually exclusive, the second follows by definition of the random set, and the third by definition of $C(B)$.

Now, use the non-redundant case to characterize the constraints imposed by (23). The distribution function $P((Y(d), S(d), \tilde{Z})) \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}\times\tilde{\mathcal{Z}}}$ characterizes a selection in $Sel((\mathbf{Y}_d, \mathbf{S}_d), \tilde{Z})$ if and only if $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ such that $B \subsetneq \mathcal{Y} \times \mathcal{S}$ $P-$a.s.:

$$P\big((Y(d), S(d)) \in B\big|\tilde{Z}\big) \geq P\big(G = O, D = d, (Y, S) \in B|\tilde{Z}\big) + P\big(G = E, D = d, S \in C(B)|\tilde{Z}\big). \tag{24}$$

To incorporate the fact that $\tilde{Z} \perp\!\!\!\perp (Y(d), S(d))$, intersect $Sel((\mathbf{Y}_d, \mathbf{S}_d), \tilde{Z}) \cap \tilde{I}$ which yields $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ such that $B \subsetneq \mathcal{Y} \times \mathcal{S}$:

$$\begin{aligned} P\big((Y(d), S(d)) \in B\big) &\geq ess\sup_{\tilde{Z}} \begin{bmatrix} P\big(G = O, D = d, (Y, S) \in B|\tilde{Z}\big) + \\ P\big(G = E, D = d, S \in C(B)|\tilde{Z}\big) \end{bmatrix} \\ &= \max\left\{P_O\big((Y, S) \in B, D = d\big), ess\sup_{Z} P_E\big(S \in C(B), D = d|Z\big)\right\}, \end{aligned} \tag{25}$$

where the equality follows by the definition of $\tilde{Z}$. If $\tilde{Z} \notin \mathcal{Z}$ (i.e. $G = O$), $P\big(G = E, D = d, S \in C(B)|\tilde{Z}\big) = 0$ and the expression reduces to $P_O\big((Y, S) \in B, D = d\big)$. If $\tilde{Z} \in \mathcal{Z}$ (i.e. $G = E$), $P\big(G = O, D = d, (Y, S) \in B|\tilde{Z}\big) = 0$ and the expression reduces to $ess\sup_Z P_E\big(S \in C(B), D = d|Z\big)$.

Finally, note that when experimental data are not observed, $P(G = O) = 1$ and thus $P\big(G = E, D = d, S \in C(B)|\tilde{Z}\big) = 0$. Thus, (25) simplifies to:

$$P((Y(d), S(d)) \in B) \geq P_O\big((Y, S) \in B, D = d\big) \tag{26}$$

Thus, (19) follows by (23) and (26).

Sharpness follows by construction. For any $(P_{Y(0),S(0)}, P_{Y(1),S(1)}) \in \mathcal{H}(P_{Y(0),S(0)}, P_{Y(1),S(1)})$ there exist $(Y(0), S(0), Y(1), S(1))$that are consistent with the data and assumptions such that

$(Y(d), S(d)) \stackrel{d}{=} P_{Y(d),S(d)}$ for $d \in \{0,1\}$. $\square$

Lemma S.4. *Let $\mathcal{H}(P_{Y(0),S(0)}, P_{Y(1),S(1)})$, and $\mathcal{H}^O(P_{Y(0),S(0)}, P_{Y(1),S(1)})$ be the identified sets for pairs of distributions $P_{Y(d),S(d)}$ for $d \in \{0,1\}$ when the experimental data are observed and unobserved, respectively. Let $\mathcal{H}(P_{Y(d)})$, $\mathcal{H}(P_{S(d)})$, $\mathcal{H}^O(P_{Y(d)})$, and $\mathcal{H}^O(P_{S(d)})$ be the identified sets for the corresponding marginals. Suppose only that Assumptions RA, EV hold. Then:*

i) $\mathcal{H}^O(P_{Y(0)}, P_{Y(1)}) = \mathcal{H}(P_{Y(0)}, P_{Y(1)})$.

ii) $\mathcal{H}(P_{Y(0)}, P_{S(0)}, P_{Y(1)}, P_{S(1)}) = \bigtimes_{d\in\{0,1\}} \Big(\mathcal{H}(P_{Y(d)}) \times \mathcal{H}(P_{S(d)})\Big)$;

iii) $\mathcal{H}^O(P_{Y(0)}, P_{S(0)}, P_{Y(1)}, P_{S(1)}) = \bigtimes_{d\in\{0,1\}} \Big(\mathcal{H}^O(P_{Y(d)}) \times \mathcal{H}^O(P_{S(d)})\Big)$;

Proof. i)

Let $\pi_k\delta$ denote the projections of a distribution function δ onto the k-th marginal. Observe that:

$$\begin{aligned}\mathcal{H}(P_{Y(0)}, P_{Y(1)}) &= \big\{(\pi_1\delta_0, \pi_1\delta_1) \in (\mathcal{P}^{\mathcal{Y}})^2 : (\delta_0, \delta_1) \in \mathcal{H}(P_{Y(0),S(0)}, P_{Y(1),S(1)})\big\} \\ &= \bigtimes_{d\in\{0,1\}} \big\{\pi_1\delta_d \in \mathcal{P}^{\mathcal{Y}} : \delta_d \in \mathcal{H}(P_{Y(d),S(d)})\big\} = \bigtimes_{d\in\{0,1\}} \mathcal{H}(P_{Y(d)})\end{aligned} \tag{27}$$

where the first equality is by definition of $\mathcal{H}(P_{Y(0)}, P_{Y(1)})$, the second follows since by Lemma S.3 $\mathcal{H}(P_{Y(0),S(0)}, P_{Y(1),S(1)}) = \mathcal{H}(P_{Y(0),S(0)}) \times \mathcal{H}(P_{Y(1),S(1)})$, and the third is by definition of $\mathcal{H}(P_{Y(d)})$. By an analogous argument:

$$\mathcal{H}^O(P_{Y(0)}, P_{Y(1)}) = \bigtimes_{d\in\{0,1\}} \mathcal{H}^O(P_{Y(d)}) \tag{28}$$

Hence, it suffices to show that $\mathcal{H}(P_{Y(d)}) = \mathcal{H}^O(P_{Y(d)})$ for $d \in \{0,1\}$.

To that end, fix any $d \in \{0,1\}$. I show that $\mathcal{H}(P_{Y(d)}) \subseteq \mathcal{H}^O(P_{Y(d)})$ and $\mathcal{H}^O(P_{Y(d)}) \subseteq \tilde{\mathcal{H}} \subseteq \mathcal{H}(P_{Y(d)})$, where $\tilde{\mathcal{H}} := \big\{\delta \in \mathcal{P}^{\mathcal{Y}} : \forall B \in \mathcal{C}(\mathcal{Y}),\ \ \delta(B) \geq P_O(Y \in B, D = d)\big\}$.

$\underline{\mathcal{H}(P_{Y(d)}) \subseteq \mathcal{H}^O(P_{Y(d)})}$

Observe that:

$$\max\left\{P_O\big((Y,S) \in B, D = d\big), ess\sup_Z P_E\big(S \in C(B), D = d|Z\big)\right\} \geq P_O((Y,S) \in B, D = d).$$

By Lemma S.3 then $\mathcal{H}(P_{Y(d),S(d)}) \subseteq \mathcal{H}^O(P_{Y(d),S(d)})$, and therefore $\mathcal{H}(P_{Y(d)}) \subseteq \mathcal{H}^O(P_{Y(d)})$.

$\underline{\mathcal{H}^O(P_{Y(d)}) \subseteq \tilde{\mathcal{H}}}$

Fix an arbitrary $\delta \in \mathcal{H}^O(P_{Y(d)})$. Let $\bar{\delta} \in \mathcal{H}^O(P_{Y(d),S(d)})$ be such that $\delta = \pi_1\bar{\delta}$. By Lemma S.3, for any $B \in \mathcal{C}(\mathcal{Y})$:

$$\delta(B) = \bar{\delta}(B \times \mathcal{S}) \geq P_O((Y, S) \in B \times \mathcal{S}, D = d) = P_O(Y \in B, D = d).$$

Hence, $\mathcal{H}^O(P_{Y(d)}) \subseteq \tilde{\mathcal{H}}$.

$\underline{\tilde{\mathcal{H}} \subseteq \mathcal{H}(P_{Y(d)})}$.

Fix an arbitrary $\delta_{Y(d)} \in \tilde{\mathcal{H}}$. I construct a distribution $\delta_{Y(d),S(d)} \in \mathcal{H}(P_{Y(d),S(d)})$ whose $Y(d)$–marginal is $\delta_{Y(d)}$. For $d \in \{0, 1\}$ and $B_S \in \mathcal{C}(\mathcal{S})$, define:

$$L_d(B_S) := \max \left\{ P_O(S \in B_S, D = d), ess \sup_Z P_E(S \in B_S, D = d|Z) \right\}. \tag{29}$$

If Assumptions RA and EV hold, by Lemma S.3 there exists at least one joint distribution $\bar{\delta} \in \mathcal{H}\big(P_{Y(d),S(d)}\big) \subseteq \mathcal{P}^{\mathcal{Y} \times \mathcal{S}}$. By the lemma, for any $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$:

$$\bar{\delta}(B) \geq \max \left\{ P_O\big((Y, S) \in B, D = d\big), ess \sup_Z P_E\big(S \in C(B), D = d|Z\big) \right\} \tag{30}$$

Observe that $\pi_2 \bar{\delta}$ denotes the $S(d)$–marginal of $\bar{\delta}$. Fix any $B_S \in \mathcal{C}(\mathcal{S})$. By (30):

$$\pi_2 \bar{\delta}(B_S) = \bar{\delta}(\mathcal{Y} \times B_S) \geq \max \left\{ P_O(S \in B_S, D = d), ess \sup_Z P_E(S \in B_S, D = d|Z) \right\} = L_d(B_S).$$

Therefore, there always exists a distribution $\delta_{S(d)}$ such that $\delta_{S(d)}(B_S) \geq L_d(B_S)$ for any $B_S \in \mathcal{C}(\mathcal{S})$. Fix any such $\delta_{S(d)} = \pi_2 \bar{\delta}$. For $B_Y \subseteq \mathcal{Y}$, and $B_S \subseteq \mathcal{S}$, define finite signed measures:

$$\nu_{d,Y}(B_Y) := \delta_{Y(d)}(B_Y) - P_O(Y \in B_Y, D = d), \qquad \nu_{d,S}(B_S) := \delta_{S(d)}(B_S) - P_O(S \in B_S, D = d).$$

Since $\delta_{Y(d)} \in \tilde{\mathcal{H}}$ and $\delta_{S(d)}(B_S) \geq L_d(B_S) \geq P_O(S \in B_S, D = d)$, $\nu_{d,Y}$ and $\nu_{d,S}$ are nonnegative finite measures. Moreover, $\nu_{d,Y}(\mathcal{Y}) = \nu_{d,S}(\mathcal{S}) = 1 - P_O(D = d)$, so they have the same total mass. Since $\nu_{d,Y}$ and $\nu_{d,S}$ are non-negative finite measures on Polish spaces with the same total mass, there exists a coupling ν_d on $\mathcal{Y} \times \mathcal{S}$ with marginals $\nu_{d,Y}$ and $\nu_{d,S}$. Define for any $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$:

$$\delta_{Y(d),S(d)}(B) := P_O((Y, S) \in B, D = d) + \nu_d(B). \tag{31}$$

By construction, $\delta_{Y(d),S(d)}$ has marginals $\delta_{Y(d)}$ and $\delta_{S(d)}$:

$$\pi_1 \delta_{Y(d),S(d)} = P_O(Y \in \cdot, D = d) + \nu_{d,Y} = \delta_{Y(d)},$$
$$\pi_2 \delta_{Y(d),S(d)} = P_O(S \in \cdot, D = d) + \nu_{d,S} = \delta_{S(d)}.$$

Again, by construction, for any $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$:

$$\delta_{Y(d),S(d)}(B) \geq P_O((Y,S) \in B, D = d) \tag{32}$$

Moreover, by definition $\mathcal{Y} \times C(B) \subseteq B$ for any $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$. Therefore:

$$\begin{aligned} \delta_{Y(d),S(d)}(B) &\geq \delta_{Y(d),S(d)}\big(\mathcal{Y} \times C(B)\big) = \delta_{S(d)}\big(C(B)\big) \\ &\geq \max\left\{P_O(S \in C(B), D = d), ess\sup_Z P_E(S \in C(B), D = d|Z)\right\} \\ &\geq ess\sup_Z P_E(S \in C(B), D = d|Z) \end{aligned} \tag{33}$$

where the first inequality is by $\mathcal{Y} \times C(B) \subseteq B$, the equality is by $\pi_2\delta_{Y(d),S(d)} = \delta_{S(d)}$, the second inequality by $\delta_{S(d)}(B_S) \geq L_d(B_S)$ for any $B_S \in \mathcal{C}(\mathcal{S})$, and the final inequality is by observation. Combining (32) and (33), for any $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$:

$$\delta_{Y(d),S(d)}(B) \geq \max\left\{P_O\big((Y,S) \in B, D = d\big), ess\sup_Z P_E\big(S \in C(B), D = d|Z\big)\right\} \tag{34}$$

By Lemma S.3, $\delta_{Y(d),S(d)} \in \mathcal{H}(P_{Y(d),S(d)})$. Therefore, $\delta_{Y(d)} \in \mathcal{H}(P_{Y(d)})$ so $\tilde{\mathcal{H}} \subseteq \mathcal{H}(P_{Y(d)})$.

Combining all three statements, $\mathcal{H}(P_{Y(d)}) \subseteq \mathcal{H}^O(P_{Y(d)})$ and $\mathcal{H}^O(P_{Y(d)}) \subseteq \tilde{\mathcal{H}} \subseteq \mathcal{H}(P_{Y(d)})$. This implies $\mathcal{H}(P_{Y(d)}) = \tilde{\mathcal{H}} = \mathcal{H}^O(P_{Y(d)})$. By (27) and (28) then:

$$\mathcal{H}(P_{Y(0)}, P_{Y(1)}) = \bigtimes_{d\in\{0,1\}} \mathcal{H}(P_{Y(d)}) = \bigtimes_{d\in\{0,1\}} \mathcal{H}^O(P_{Y(d)}) = \mathcal{H}^O(P_{Y(0)}, P_{Y(1)}). \tag{35}$$

ii)

By definition:

$$\begin{aligned} &\mathcal{H}(P_{Y(0)}, P_{S(0)}, P_{Y(1)}, P_{S(1)}) \\ &= \big\{(\pi_1\delta_0, \pi_2\delta_0, \pi_1\delta_1, \pi_2\delta_1) \in (\mathcal{P}^{\mathcal{Y}})^2 \times (\mathcal{P}^{\mathcal{S}})^2 : (\delta_0, \delta_1) \in \mathcal{H}(P_{Y(0),S(0)}, P_{Y(1),S(1)})\big\} \\ &= \bigtimes_{d\in\{0,1\}} \big\{(\pi_1\delta_d, \pi_2\delta_d) \in \mathcal{P}^{\mathcal{Y}} \times \mathcal{P}^{\mathcal{S}} : \delta_d \in \mathcal{H}(P_{Y(d),S(d)})\big\} = \bigtimes_{d\in\{0,1\}} \mathcal{H}(P_{Y(d)}, P_{S(d)}) \end{aligned} \tag{36}$$

where the second equality follows since by Lemma S.3 $\mathcal{H}(P_{Y(0),S(0)}, P_{Y(1),S(1)}) = \mathcal{H}(P_{Y(0),S(0)}) \times \mathcal{H}(P_{Y(1),S(1)})$ and the third is by definition.

It is thus sufficient to show that $\mathcal{H}(P_{Y(d)}, P_{S(d)}) = \mathcal{H}(P_{Y(d)}) \times \mathcal{H}(P_{S(d)})$ for $d \in \{0,1\}$. Fix an arbitrary $d \in \{0,1\}$. Since a Cartesian product of projections of a set is always a superset of the set itself, it is immediate that $\mathcal{H}(P_{Y(d)}, P_{S(d)}) \subseteq \mathcal{H}(P_{Y(d)}) \times \mathcal{H}(P_{S(d)})$. It remains to show that

$\mathcal{H}(P_{Y(d)}) \times \mathcal{H}(P_{S(d)}) \subseteq \mathcal{H}(P_{Y(d)}, P_{S(d)})$. Let $\tilde{\tilde{H}} := \{\delta \in \mathcal{P}^{\mathcal{S}} : \forall B \in \mathcal{C}(\mathcal{S}),\ \ \delta(B) \geq L_d(B)\}$. Recall from the proof of *i)* that $\tilde{\mathcal{H}} = \{\delta \in \mathcal{P}^{\mathcal{Y}} : \forall B \in \mathcal{C}(\mathcal{Y}),\ \ \delta(B) \geq P_O(Y \in B, D = d)\}$. The proof also shows: 1) $\tilde{\tilde{\mathcal{H}}} \neq \emptyset$ when Assumptions RA and EV hold; and 2) for any $\delta_{Y(d)} \in \tilde{\mathcal{H}}$ and any $\delta_{S(d)} \in \tilde{\tilde{\mathcal{H}}}$ there exists a joint distribution $\delta_{Y(d),S(d)} \in \mathcal{H}(P_{Y(d),S(d)})$ such that $\pi_1 \delta_{Y(d),S(d)} = \delta_{Y(d)}$ and $\pi_2 \delta_{Y(d),S(d)} = \delta_{S(d)}$ on $\mathcal{C}(\mathcal{Y} \times \mathcal{S})$. Hence:

$$\tilde{\mathcal{H}} \times \tilde{\tilde{\mathcal{H}}} \subseteq \{(\pi_1\delta, \pi_2\delta) \in \mathcal{P}^{\mathcal{Y}} \times \mathcal{P}^{\mathcal{S}} : \delta \in \mathcal{H}(P_{Y(d),S(d)})\} = \mathcal{H}(P_{Y(d)}, P_{S(d)}) \tag{37}$$

where the equality is by definition. Recall also that $\mathcal{H}(P_{Y(d)}) = \tilde{\mathcal{H}}$, by the proof of *i)*. It thus remains to show that $\tilde{\tilde{\mathcal{H}}} = \mathcal{H}(P_{S(d)})$. Since for any $\delta_{S(d)} \in \tilde{\tilde{\mathcal{H}}}$ there exists $\delta_{Y(d),S(d)} \in \mathcal{H}(P_{Y(d),S(d)})$ such that $\pi_2 \delta_{Y(d),S(d)} = \delta_{S(d)}$, then $\tilde{\tilde{\mathcal{H}}} \subseteq \mathcal{H}(P_{S(d)})$. For the converse, note that for any $\delta_{S(d)} \in \mathcal{H}(P_{S(d)})$ there exists $\delta_{Y(d),S(d)} \in \mathcal{H}(P_{Y(d),S(d)})$ such that $\pi_2 \delta_{Y(d),S(d)} = \delta_{S(d)}$. By Lemma S.3, for that $\delta_{Y(d),S(d)}$ then $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$:

$$\delta_{Y(d),S(d)}(B) \geq \max\left\{P_O\big((Y,S) \in B, D = d\big), ess\sup_Z P_E\big(S \in C(B), D = d|Z\big)\right\}. \tag{38}$$

Fix any $B_S \in \mathcal{C}(\mathcal{S})$. Then by definition of $C(B_S)$:

$$\begin{aligned}\delta_{Y(d),S(d)}(\mathcal{Y} \times B_S) &= \delta_{S(d)}(B_S) \\ &\geq \max\left\{ess\sup_Z P_E(S \in B_S, D = d|Z), P_O(S \in B_S, D = d)\right\} = L_d(B_S).\end{aligned} \tag{39}$$

Therefore, $\delta_{S(d)} \in \tilde{\tilde{\mathcal{H}}}$ and thus $\mathcal{H}(P_{S(d)}) \subseteq \tilde{\tilde{\mathcal{H}}}$.

It then follows that $\mathcal{H}(P_{S(d)}) = \tilde{\tilde{\mathcal{H}}}$. Since $\mathcal{H}(P_{Y(d)}) = \tilde{\mathcal{H}}$, by (37) then $\mathcal{H}(P_{Y(d)}) \times \mathcal{H}(P_{S(d)}) \subseteq \mathcal{H}(P_{Y(d)}, P_{S(d)})$ and therefore $\mathcal{H}(P_{Y(d)}) \times \mathcal{H}(P_{S(d)}) = \mathcal{H}(P_{Y(d)}, P_{S(d)})$. The result is then by (36).

iii)

The result follows from the steps in the proof of *ii)* by redefining: $L_d(B_S) := P_O(S \in B_S, D = d)$. □

Lemma S.5. *Let $\tilde{Z} = \mathbb{1}[G = E]Z + \mathbb{1}[G = O](sup\mathcal{Z} + 1)$ and $\bar{I}$ be the set of random elements (E_1, E_2) such that $(E_1, E_2) \in \mathcal{S} \times \tilde{\mathcal{Z}}$ and $E_1 \perp\!\!\!\perp E_2$. Define random sets $\mathbf{Y}_d$ and $\mathbf{S}_d$ for $d \in \{0, 1\}$:*

$$\mathbf{Y}_d := \begin{cases}\{Y\}, & \text{if } (D,G) = (d,O) \\ \mathcal{Y}, & \text{otherwise}\end{cases}, \quad \mathbf{S}_d := \begin{cases}\{S\}, & \text{if } (D,G) \in \{(d,E),(d,O)\} \\ \mathcal{S}, & \text{otherwise}\end{cases}. \tag{40}$$

Then $\mathcal{H}^{EV/RA}(m,\gamma) = \tilde{\mathcal{H}}^{EV/RA}(m,\gamma)$ *for:*

$$\mathcal{H}^{EV/RA}(m,\gamma) := \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M} \times \left(\mathcal{P}^{\mathcal{S}}\right)^2 : \forall d \in \{0,1\}, \;\; \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}, \\ \exists \upsilon_d \in Sel(\mathbf{Y}_d), \;\; (\upsilon_d, \varsigma_d) \perp\!\!\!\perp \tilde{Z}, \;\; \gamma_d \overset{d}{=} \varsigma_d, \;\; m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d] \;\; a.s. \end{array} \right\}$$

$$\tilde{\mathcal{H}}^{EV/RA}(m,\gamma) := \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M} \times \left(\mathcal{P}^{\mathcal{S}}\right)^2 : \forall d \in \{0,1\}, \;\; \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}, \\ \exists \upsilon_d \in Sel(\mathbf{Y}_d), \;\; \gamma_d \overset{d}{=} \varsigma_d, \;\; m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d] \;\; a.s. \end{array} \right\}$$

Proof. $\mathcal{H}^{EV/RA}(m,\gamma) \subseteq \tilde{\mathcal{H}}^{EV/RA}(m,\gamma)$ since the former imposes a strict superset of conditions on (m,γ). For the converse, fix an arbitrary $(m,\gamma) \in \tilde{\mathcal{H}}^{EV/RA}(m,\gamma)$ and $d \in \{0,1\}$, and let $(\upsilon_d, \varsigma_d)$ be the corresponding selections in $Sel(\mathbf{Y}_d) \times \left[Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}\right]$ that generate m_d and γ_d. To prove that $(m,\gamma) \in \mathcal{H}^{EV/RA}(m,\gamma)$, I show that there exist $(\upsilon'_d, \varsigma'_d)$ such that: 1) $(\varsigma'_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}$ and $\upsilon'_d \in Sel(\mathbf{Y}_d)$; 2) $m_d(\varsigma'_d) = E_O[\upsilon'_d|\varsigma'_d]$ and $\varsigma'_d \overset{d}{=} \gamma_d$; and 3) $(\upsilon'_d, \varsigma'_d) \perp\!\!\!\perp \tilde{Z}$.

Let $P_{\upsilon'_d,\varsigma'_d}$ be a distribution such that $\forall z \in \tilde{\mathcal{Z}}$, $P_{\upsilon'_d,\varsigma'_d}(\cdot|\varsigma'_d = s, \tilde{Z} = z) = P_{\upsilon_d,\varsigma_d}(\cdot|\varsigma_d = s, G = O)$ $\forall s \in \mathcal{S}$, and $P_{\varsigma'_d}(\cdot|\tilde{Z} = z) = P_{\varsigma_d}(\cdot)$. Note that these conditions fully specify $P_{\upsilon'_d,\varsigma'_d}$. I first show that there exist $(\varsigma'_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}$ and $\upsilon'_d \in Sel(\mathbf{Y}_d)$ such that $(\upsilon'_d, \varsigma'_d) \overset{d}{=} P_{\upsilon'_d,\varsigma'_d}$. I then show that $(\upsilon'_d, \varsigma'_d)$ fulfill conditions 2) $m_d(\varsigma'_d) = E_O[\upsilon'_d|\varsigma'_d]$ and $\varsigma'_d \overset{d}{=} \gamma_d$; and 3) $(\upsilon'_d, \varsigma'_d) \perp\!\!\!\perp \tilde{Z}$. Recall that, as in the proof of Lemma S.3, by Artstein (1983, Theorem 2.1) and Molchanov and Molinari (2018, Theorem 2.33), $(\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel(\mathbf{Y}_d) \times Sel(\mathbf{S}_d) \times \{\tilde{Z}\}$ if and only if $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ $P-$a.s.:

$$P_{\upsilon_d,\varsigma_d}(B|\tilde{Z}) \geq P\big(G = O, D = d, (Y,S) \in B|\tilde{Z}\big) + P\big(G = E, D = d, S \in C(B)|\tilde{Z}\big) \tag{41}$$

where $C(B) = \{s : \mathcal{Y} \times \{s\} \subseteq B\}$. Since $(\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel(\mathbf{Y}_d) \times \left[Sel(\mathbf{S}_d) \times \{\tilde{Z}\} \cap \bar{I}\right]$, it must be that $P_{\varsigma_d}(\cdot|\tilde{Z}) = P_{\varsigma_d}(\cdot)$ $P-$a.s. Then for any $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ $P-$a.s.:

$$P_{\upsilon'_d,\varsigma'_d}(B) = P_{\upsilon'_d,\varsigma'_d}(B|\tilde{Z}) = P_{\upsilon_d,\varsigma_d}(B|G = O) \geq P_O((Y,S) \in B, D = d) \tag{42}$$

where the first equality is by $P_{\upsilon'_d,\varsigma'_d}(\cdot|\tilde{Z}) = P_{\upsilon'_d,\varsigma'_d}(\cdot)$ which follows by $P_{\varsigma'_d}(\cdot|\tilde{Z} = z) = P_{\varsigma_d}(\cdot)$ and $P_{\upsilon'_d,\varsigma'_d}(\cdot|\varsigma'_d = s, \tilde{Z} = z) = P_{\upsilon_d,\varsigma_d}(\cdot|\varsigma_d = s, G = O)$ $\forall s \in \mathcal{S}$ and $\forall z \in \tilde{\mathcal{Z}}$. The second is by $P_{\upsilon'_d,\varsigma'_d}(\cdot|\varsigma'_d = s, \tilde{Z} = z) = P_{\upsilon_d,\varsigma_d}(\cdot|\varsigma_d = s, G = O)$, $P_{\varsigma'_d}(\cdot|\tilde{Z} = z) = P_{\varsigma_d}(\cdot)$ and $P_{\varsigma_d}(\cdot) = P_{\varsigma_d}(\cdot|\tilde{Z})$ $P-$a.s. The inequality is by (41) fixing $\tilde{Z}$ to the value corresponding to $G = O$. For any $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ also:

$$\begin{aligned} P_{\upsilon'_d,\varsigma'_d}(B) &\geq P_{\upsilon'_d,\varsigma'_d}(\mathcal{Y} \times C(B)) = P_{\varsigma'_d}(C(B)) \\ &= P_{\varsigma_d}(C(B)) = P_{\varsigma_d}(C(B)|\tilde{Z}) = P_{\upsilon_d,\varsigma_d}(\mathcal{Y} \times C(B)|\tilde{Z}) \\ &\geq P\big(G = O, D = d, (Y,S) \in \mathcal{Y} \times C(B)|\tilde{Z}\big) + P\big(G = E, D = d, S \in C(B)|\tilde{Z}\big) \end{aligned} \tag{43}$$

where the first inequality is because $\mathcal{Y} \times C(B) \subseteq B$, the first equality is by definition of a marginal distribution, the third is by $P_{\varsigma'_d}(\cdot|\tilde{Z}) = P_{\varsigma_d}(\cdot)$, the fourth is by $P_{\varsigma_d}(\cdot|\tilde{Z}) = P_{\varsigma_d}(\cdot)$ $P-$a.s, the fifth is by definition of a marginal distribution, and the inequality is by (41). Since (43) holds for almost any $\tilde{Z} = \tilde{z}$ and hence for almost any $Z = z$ when $G = E$:

$$P_{\upsilon'_d,\varsigma'_d}(B) \geq ess \sup_{Z} P_E(S \in C(B), D = d|Z). \tag{44}$$

By (42) and (44) then $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$:

$$P_{\upsilon'_d,\varsigma'_d}(B) \geq \max\left\{P_O\big((Y,S) \in B, D = d\big), ess \sup_{Z} P_E\big(S \in C(B), D = d|Z\big)\right\} \tag{45}$$

Then recall that $\tilde{I}$ is the set of random elements (E_1, E_2, E_3) such that $(E_1, E_2, E_3) \in \mathcal{Y} \times \mathcal{S} \times \tilde{\mathcal{Z}}$ and $(E_1, E_2) \perp\!\!\!\perp E_3$. By Lemma S.3, $\exists(\upsilon'_d, \varsigma'_d, \tilde{Z}) \in Sel(\mathbf{Y}_d) \times Sel(\mathbf{S}_d) \times \{\tilde{Z}\} \cap \tilde{I}$ such that $(\upsilon'_d, \varsigma'_d) \stackrel{d}{=} P_{\upsilon'_d,\varsigma'_d}$ if and only if $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ (45) holds. Since $Sel(\mathbf{Y}_d) \times Sel(\mathbf{S}_d) \times \{\tilde{Z}\} \cap \tilde{I} \subseteq Sel(\mathbf{Y}_d) \times \left[Sel(\mathbf{S}_d) \times \{\tilde{Z}\} \cap \bar{I}\right]$, there also exist $(\varsigma'_d, \tilde{Z}) \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}$ and $\upsilon'_d \in Sel(\mathbf{Y}_d)$ such that $(\upsilon'_d, \varsigma'_d) \stackrel{d}{=} P_{\upsilon'_d,\varsigma'_d}$.

Since $P_{\upsilon'_d,\varsigma'_d}(\cdot|\varsigma'_d = s) = P_{\upsilon_d,\varsigma_d}(\cdot|\varsigma_d = s, G = O)$, then $m_d(\varsigma'_d) = E_O[\upsilon'_d|\varsigma'_d]$ a.s. Because $P_{\varsigma'_d}(\cdot|\tilde{Z}) = P_{\varsigma_d}(\cdot) = P_{\varsigma'_d}(\cdot)$, $\varsigma'_d \stackrel{d}{=} \varsigma_d \stackrel{d}{=} \gamma_d$. Finally, because $(\upsilon'_d, \varsigma'_d, \tilde{Z}) \in Sel(\mathbf{Y}_d) \times Sel(\mathbf{S}_d) \times \{\tilde{Z}\} \cap \tilde{I}$, $(\upsilon'_d, \varsigma'_d) \perp\!\!\!\perp \tilde{Z}$. Therefore, if $(m, \gamma) \in \tilde{\mathcal{H}}^{EV/RA}(m, \gamma)$, then $(m, \gamma) \in \mathcal{H}^{EV/RA}(m, \gamma)$. □

Lemma S.6. *Let $\tilde{Z}$, $\bar{I}$, and $\mathbf{S}_d$ be defined as in Lemma S.5. For any γ_d that is a distribution of a selection in $Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}$, there exists a γ_d-integrable function π_{γ_d} such that for any measurable set $B \in \mathcal{B}(\mathcal{S})$:*

$$P_O(S \in B, D = d) = \int_B \pi_{\gamma_d} d\gamma_d. \tag{46}$$

Then for the propensity score functional $\pi_{\gamma_d} := \frac{dP_O(S,D=d)}{d\gamma_d}$ and any $\varsigma'_d \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}$ with $\varsigma'_d \stackrel{d}{=} \gamma'_d$:

$$P_O(D = d|\varsigma'_d) = \pi_{\gamma'_d}(\varsigma'_d) \ a.s. \tag{47}$$

Proof. Fix any γ_d such that $\exists \varsigma_d \in Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}$ and $\gamma_d \stackrel{d}{=} \varsigma_d$. Then for any $B \in \mathcal{B}(\mathcal{S})$:

$$P_O(\varsigma_d \in B, D = d) \leq P_O(\varsigma_d \in B) = P(\varsigma_d \in B) = \gamma_d(B)$$

where the inequality is by observation. For the first equality, recall that I is a set of random elements $(E_1, E_2) \in \mathcal{S} \times \tilde{\mathcal{Z}}$, and observe that $(\varsigma_d, \tilde{Z}) \in I$. Therefore, $\varsigma_d \perp\!\!\!\perp G$, by definition of $\tilde{Z}$.

For the second equality note that $\varsigma_d \overset{d}{=} \gamma_d$.

Next, note that $P_O(\varsigma_d \in B, D = d) = P_O(S \in B, D = d)$ for any measurable set $B \in \mathcal{B}(\mathcal{S})$ because $\varsigma_d \in Sel(\mathbf{S}_d)$ and $P_O(\mathbf{S}_d = \{S\}, D = d) = 1$. Therefore, $P_O(S \in B, D = d) \leq \gamma_d(B)$ for any $B \in \mathcal{B}(\mathcal{S})$. Hence, $P_O(S, D = d)$ is absolutely continuous with respect to γ_d. Then, by the Radon-Nikodym theorem there exists a measurable function π_{γ_d} such that for any measurable set $B \in \mathcal{B}(\mathcal{S})$:

$$P_O(S \in B, D = d) = \int_B \pi_{\gamma_d} d\gamma_d$$

and $\pi_{\gamma_d} = dP_O(S, D = d)/d\gamma_d$.

Therefore, for any γ'_d that is a distribution of a selection in $Sel((\mathbf{S}_d, \tilde{Z})) \cap \bar{I}$, there exists $\pi_{\gamma'_d} = dP_O(S, D = d)/d\gamma'_d$ $P_O(S \in B, D = d) = \int_B \pi_{\gamma'_d} d\gamma'_d$ for any measurable set $B \in \mathcal{B}(\mathcal{S})$. Hence, also $\pi_{\gamma'_d}(s) = P_O(D = d|\varsigma'_d = s)$, γ'_d-a.e. $s \in \mathcal{S}$, which concludes the proof. □

Lemma S.7. *Let $\mathcal{Y}$ be a compact set. If there exists $d \in \{0, 1\}$ such that $V_O[Y|S, D = d] > 0$ $P-$a.s., then $E_O[Y|S, D = d] \in (\inf \mathcal{Y}, \sup \mathcal{Y})$ $P-$a.s.*

Proof. I prove that $E_O[Y|S, D = d] < \sup \mathcal{Y}$ $P-$a.s. and $E_O[Y|S, D = d] > \inf \mathcal{Y}$ $P-$a.s follows by a symmetric argument. Since $\mathcal{Y}$ is a compact set, both $\sup \mathcal{Y}$ and $\inf \mathcal{Y}$ are finite.

By contraposition suppose that $P_O(E_O[Y|S, D = d] \geq \sup \mathcal{Y}) > 0$. Then by definition of $\mathcal{Y}$, $P_O(E_O[Y|S, D = d] = \sup \mathcal{Y}) > 0$, so there exists a Borel set $B \in \mathcal{B}(\mathcal{S})$ with $P_O(S \in B|D = d) > 0$ such that $E_O[Y|S \in B, D = d] = \sup \mathcal{Y}$. Now, I show that this implies $P_O(Y = \sup \mathcal{Y}|S \in B, D = d) = 1$. Suppose not, so that $P_O(Y = \sup \mathcal{Y}|S \in B, D = d) < 1$, then:

$$\begin{aligned} E_O[Y|S \in B, D = d] =& E_O[Y|Y = \sup \mathcal{Y}, S \in B, D = d]P_O(Y = \sup \mathcal{Y}|S \in B, D = d)+ \\ & E_O[Y|Y \neq \sup \mathcal{Y}, S \in B, D = d]P_O(Y \neq \sup \mathcal{Y}|S \in B, D = d) \\ =& E_O[Y|Y = \sup \mathcal{Y}, S \in B, D = d]P_O(Y = \sup \mathcal{Y}|S \in B, D = d)+ \\ & E_O[Y|Y < \sup \mathcal{Y}, S \in B, D = d]P_O(Y < \sup \mathcal{Y}|S \in B, D = d) \\ =& \sup \mathcal{Y} P_O(Y = \sup \mathcal{Y}|S \in B, D = d)+ \\ & E_O[Y|Y < \sup \mathcal{Y}, S \in B, D = d]P_O(Y < \sup \mathcal{Y}|S \in B, D = d) \\ <& \sup \mathcal{Y} \end{aligned} \tag{48}$$

where the first equality is by LIE, second is by definition of $\mathcal{Y}$, third by observation, and the fourth by $E_O[Y|Y < \sup \mathcal{Y}, S \in B, D = d] < \sup \mathcal{Y}$ and $P_O(Y = \sup \mathcal{Y}|S \in B, D = d) < 1$. By assumption, $E_O[Y|S \in B, D = d] = \sup \mathcal{Y}$. Then (48) yields a contradiction, showing that $P_O(Y = \sup \mathcal{Y}|S \in B, D = d) = 1$. But then $V_O[Y|S \in B, D = d] = 0$ and $P_O(S \in B|D =$

$d) > 0$, so $P_O(V_O[Y|S, D = d] = 0) > 0$ which contradicts $V_O[Y|S, D = d] > 0$ $P-$a.s. Thus $V_O[Y|S, D = d] > 0$ $P-$a.s. implies $E_O[Y|S, D = d] < \sup \mathcal{Y}$ $P-$a.s.

□

Proposition S.1. *Let Assumptions RA and EV hold, and assume latent unconfoundedness (LUC): $Y(d) \perp\!\!\!\perp D|S(d), G = O$ for $d \in \{0, 1\}$. Suppose that $P_O(S(d) \in \cdot|D \neq d) \ll P_O(S(d) \in \cdot|D = d)$. Let $\mathcal{H}^{WC}(\tau)$ be the identified set for τ when latent unconfoundedness is not maintained.*

i) Suppose $\mathcal{Y}$ is a bounded set and that the observed data distribution $P_O(Y, S, D)$ is such that S is not a perfect predictor so $V_O[Y|S, D = d] > 0$ $P-$a.s. for some $d \in \{0, 1\}$. Then $\mathcal{H}^O(\tau) \subsetneq \mathcal{H}^{WC}(\tau)$.

ii) If the observed data distribution $P_O(Y, S, D)$ is such that $E_O[Y|S, D = d]$ is a trivial measurable function for all $d \in \{0, 1\}$, then τ is point-identified by the observational data, and $\mathcal{H}(\tau) = \mathcal{H}^O(\tau)$.

Proof. i) $\mathcal{Y}$ is closed by definition. Since it is bounded, it is a compact set. Then $\sup \mathcal{Y} < \infty$ and $\inf \mathcal{Y} > -\infty$. Using arguments of Manski (1990), the sharp upper bound of $\mathcal{H}^{WC}(\tau)$ is:

$$\tau \leq E_O[Y(2D - 1)] + \sup \mathcal{Y} P_O(D = 0) - \inf \mathcal{Y} P_O(D = 1) = \sup \mathcal{H}^{WC}(\tau). \tag{49}$$

Suppose that $V_O[Y|S, D = 1] > 0$ $P-$a.s. Fix $d = 1$ and the case for $d = 0$ follows by symmetric arguments. By Lemma S.7, $V_O[Y|S, D = d] > 0$ $P-$a.s. implies $E_O[Y|S, D = d] < \sup \mathcal{Y}$ $P-$a.s. If there exists $d \in \{0, 1\}$ s.t. $V_O[Y|S, D = d] > 0$ $P-$a.s., then it must be that for every Borel set $B \in \mathcal{B}(\mathcal{S})$ with $P_O(S \in B|D = d) > 0$ we have $E_O[Y|S \in B, D = d] < \sup \mathcal{Y}$. Under LUC then:

$$\begin{aligned} E_O[Y(d)|D \neq d] &= E_O[E_O[Y(d)|S(d), D \neq d]|D \neq d] \\ &= E_O[E_O[Y(d)|S(d), D = d]|D \neq d] \\ &= E_O[E_O[Y|S, D = d]|D \neq d] \\ &< \sup \mathcal{Y} \end{aligned} \tag{50}$$

where the first line is by LIE, the second by LUC, the third by definition, and the fourth because $E_O[Y \mid S = s, D = d]$ satisfies $E_O[Y \mid S = s, D = d] < \sup \mathcal{Y}$ $P_O(S \in \cdot \mid D = d)$-a.s. by Lemma S.7, and absolute continuity of $P_O(S(d) \in \cdot \mid D \neq d)$ with respect to $P_O(S \in \cdot \mid D = d)$ implies $E_O[Y \mid S(d), D = d] < \sup \mathcal{Y}$ $P_O(\cdot \mid D \neq d)$-a.s. Then under LUC:

$$\begin{aligned} E[Y(d)] &= E_O[Y \mathbb{1}[D = d]] + E[Y(d)|D \neq d] P_O(D \neq d) \\ &< E_O[Y \mathbb{1}[D = d]] + \sup \mathcal{Y} P_O(D \neq d). \end{aligned} \tag{51}$$

Therefore, under LUC:

$$
\begin{aligned}
\tau &= E[Y(1) - Y(0)] \\
&= E_O[YD] + E_O[Y(1)|D=0]P_O(D=0) - E_O[Y(1-D)] - E_O[Y(0)|D=1]P_O(D=1) \\
&< E_O[Y(2D-1)] + \sup \mathcal{Y} P_O(D=0) - \inf \mathcal{Y} P_O(D=1) = \sup \mathcal{H}^{WC}(\tau)
\end{aligned}
$$

where the inequality follows by (51). Thus $\sup \mathcal{H}^O(\tau) < \sup \mathcal{H}^{WC}(\tau)$. So there must exist a point in $\mathcal{H}^{WC}(\tau)$ which is not contained in $\mathcal{H}^O(\tau)$. Conclude that $\mathcal{H}^O(\tau) \subsetneq \mathcal{H}^{WC}(\tau)$.

ii) Suppose that for every $d \in \{0,1\}$ $E_O[Y|S, D=d]$ is a trivial measurable function. Hence there exists a $y_d \in \mathcal{Y}$ such that $E_O[Y|S, D=d] = y_d$ $P-$a.s. Then:

$$
\begin{aligned}
E_O[Y(d)|D \neq d] &= E_O[E_O[Y(d)|S(d), D \neq d]|D \neq d] \\
&= E_O[E_O[Y(d)|S(d), D = d]|D \neq d] \\
&= E_O[E_O[Y|S, D = d]|D \neq d] \\
&= y_d
\end{aligned} \tag{52}
$$

where the final line follows since $E_O[Y|S, D=d] = y_d$ $P-$a.s. Given that y_d is identified by the data, then $E_O[Y(d)]$ is identified for every $d \in \{0,1\}$, so τ is too. It is also immediate that $\mathcal{H}(\tau) = \mathcal{H}^O(\tau)$ since for every $d \in \{0,1\}$ and any $\gamma_d \in \mathcal{P}^{\mathcal{S}}$, we have that $E[Y(d)] = \int_{\mathcal{S}} y_d d\gamma_d(s) = y_d$. Since experimental data only affect the feasible γ_d, the result follows.

□

Lemma S.8. *Suppose Assumption RA holds. Assume that there is perfect experimental compliance so $Z = D|G = E$ $P-$a.s. and define conditions:*

C.1 (Surrogacy) $Y \perp\!\!\!\perp D|S, G = E$;

C.2 (Comparability) $Y \perp\!\!\!\perp G|S$.

Then:

i) C.1 *implies* $E_E[Y(1)|S(1)=s] = E_E[Y(0)|S(0)=s]$ *for all* $s \in \mathcal{S}$*;*

ii) C.1 *and* EV *imply* $E_g[Y(1)|S(1)=s] = E_{g'}[Y(0)|S(0)=s]$ *for all* $s \in \mathcal{S}$ *and* $g, g' \in \{O, E\}$*;*

iii) C.2 *implies* $E_O[Y|S=s] = E_E[Y(1)|S(1)=s]P_E(D=1|S=s) + E_E[Y(0)|S(0)=s]P_E(D=0|S=s)$ *for all* $s \in \mathcal{S}$*;*

iv) C.2 *and* EV *imply* $E_O[Y|S=s] = E_g[Y(1)|S(1)=s]P_E(D=1|S=s) + E_{g'}[Y(0)|S(0)=s]P_E(D=0|S=s)$ *for all* $s \in \mathcal{S}$ *and* $g, g' \in \{O, E\}$*;*

v) C.1 and C.2 imply $E_O[Y|S=s] = E_E[Y(d)|S(d)=s]$ *for all* $s \in \mathcal{S}$*;*

vi) C.1, C.2 and EV imply $E_O[Y|S=s] = E_g[Y(d)|S(d)=s]$ *and hence* $E_g[Y(d)|S(d)=s] = \sum_{d' \in \{0,1\}} E_O[Y(d')|S(d')=s, D=d']P_O(D=d'|S=s)$*, for all* $s \in \mathcal{S}$ *and* $g \in \{O, E\}$*.*

Proof. i) Write for any $d \in \{0,1\}$:

$$E_E[Y|S] = E_E[Y|S, D=d] = E_E[Y(d)|S(d), D=d] = E_E[Y(d)|S(d)] \tag{53}$$

where the first equality is by surrogacy, second is by definition, and third is by random assignment and perfect compliance.

ii) Under Assumption EV, $E_E[Y(d)|S(d)] = E[Y(d)|S(d)]$. The result then follows from *i)*.

iii) Write:

$$\begin{aligned} E_O[Y|S=s] &= E_E[Y|S=s] \\ &= E_E[Y(1)|S(1)=s, D=1]P_E(D=1|S=s) \\ &+ E_E[Y(0)|S(0)=s, D=0]P_E(D=0|S=s) \\ &= E_E[Y(1)|S(1)=s]P_E(D=1|S=s) + E_E[Y(0)|S(0)=s]P_E(D=0|S=s) \end{aligned} \tag{54}$$

where the first equality is by comparability, second is by LIE and definitions of Y and S, and the third is by random assignment and perfect compliance.

iv) Under Assumption EV, $E_E[Y(d)|S(d)] = E[Y(d)|S(d)]$. The result then follows from *iii)*.

v) Immediate from *i)* and *iii)*.

vi) Immediate from *v)* under Assumption EV.

□

Lemma S.9. *Let Assumptions RA and EV hold. Suppose that* $\mathcal{S}$ *is a finite set. Fix* $\gamma' \in \mathcal{H}(\gamma)$*. Define pointwise data bounds:*

$$\begin{aligned} L_d(s) &:= \mu_d(s)\pi_{\gamma'_d}(s) + \inf \mathcal{Y}\left(1 - \pi_{\gamma'_d}(s)\right), \\ U_d(s) &:= \mu_d(s)\pi_{\gamma'_d}(s) + \sup \mathcal{Y}\left(1 - \pi_{\gamma'_d}(s)\right). \end{aligned} \tag{55}$$

i) Suppose Assumption LIV holds. Define monotone envelopes:

$$m_d^{L,\gamma'}(s) := \sup_{s' \leq s} L_d(s'), \qquad m_d^{U,\gamma'}(s) := \inf_{s' \geq s} U_d(s'), \tag{56}$$

where $s' \leq s$ denotes the product order. Then:

$$m_{\gamma'}^{LB} := (m_0^{U,\gamma'}, m_1^{L,\gamma'}), \qquad m_{\gamma'}^{UB} := (m_0^{L,\gamma'}, m_1^{U,\gamma'}) \tag{57}$$

are minimal and maximal selectors of $\mathcal{H}(m|\gamma')$ with respect to T.

ii) Suppose Assumption TI holds. Define:

$$m^{L,\gamma'}(s) := \max_{d\in\{0,1\}} L_d(s), \qquad m^{U,\gamma'}(s) := \min_{d\in\{0,1\}} U_d(s), \tag{58}$$

and:

$$\begin{aligned} m^{TI,L,\gamma'}(s) &:= m^{L,\gamma'}(s)\mathbb{1}[\gamma_1'(s) \geq \gamma_0'(s)] + m^{U,\gamma'}(s)\mathbb{1}[\gamma_1'(s) < \gamma_0'(s)], \\ m^{TI,U,\gamma'}(s) &:= m^{L,\gamma'}(s)\mathbb{1}[\gamma_1'(s) < \gamma_0'(s)] + m^{U,\gamma'}(s)\mathbb{1}[\gamma_1'(s) \geq \gamma_0'(s)]. \end{aligned} \tag{59}$$

Then:

$$m_{\gamma'}^{LB} := (m^{TI,L,\gamma'}, m^{TI,L,\gamma'}), \qquad m_{\gamma'}^{UB} := (m^{TI,U,\gamma'}, m^{TI,U,\gamma'}) \tag{60}$$

are minimal and maximal selectors of $\mathcal{H}(m|\gamma')$ with respect to T.

Proof. i) Fix $\gamma' \in \mathcal{H}(\gamma)$ such that $\mathcal{H}(m|\gamma') \neq \emptyset$. By Theorem 1, for each $d \in \{0,1\}$, any $m \in \mathcal{H}(m|\gamma')$ satisfies $m_d(s) \in [L_d(s), U_d(s)]$ for all $s \in \mathcal{S}$. Under Assumption LIV, m_d is nondecreasing in the product order. Hence for any $s' \leq s$, $m_d(s) \geq m_d(s') \geq L_d(s')$, so $m_d(s) \geq \sup_{s'\leq s} L_d(s') = m_d^{L,\gamma'}(s)$. Similarly, for any $s' \geq s$, monotonicity implies $m_d(s) \leq m_d(s') \leq U_d(s')$, so $m_d(s) \leq \inf_{s'\geq s} U_d(s') = m_d^{U,\gamma'}(s)$.

To show $m_d^{L,\gamma'}$ and $m_d^{U,\gamma'}$ are feasible, note that both are nondecreasing in the product order by construction: if $s \leq s''$, then $\{s' : s' \leq s\} \subseteq \{s' : s' \leq s''\}$, so $m_d^{L,\gamma'}(s) \leq m_d^{L,\gamma'}(s'')$, and similarly $\{s' : s' \geq s\} \supseteq \{s' : s' \geq s''\}$, so $m_d^{U,\gamma'}(s) \leq m_d^{U,\gamma'}(s'')$. Since $\mathcal{H}(m|\gamma') \neq \emptyset$, pick any feasible $\tilde{m}_d$. Then $m_d^{L,\gamma'}(s) \leq \tilde{m}_d(s) \leq m_d^{U,\gamma'}(s)$ for all $s \in \mathcal{S}$, which implies $m_d^{L,\gamma'}(s) \leq m_d^{U,\gamma'}(s)$. Moreover, $L_d(s) \leq m_d^{L,\gamma'}(s)$ and $m_d^{U,\gamma'}(s) \leq U_d(s)$ for all s, so both envelopes satisfy the pointwise data restrictions. Hence $m_d^{L,\gamma'}, m_d^{U,\gamma'} \in \mathcal{M}^A$ and satisfy the constraints, so $(m_0^{U,\gamma'}, m_1^{L,\gamma'}), (m_0^{L,\gamma'}, m_1^{U,\gamma'}) \in \mathcal{H}(m|\gamma')$.

Since $T(m,\gamma') = \sum_{s\in\mathcal{S}} m_1(s)\gamma_1'(s) - \sum_{s\in\mathcal{S}} m_0(s)\gamma_0'(s)$ is nondecreasing in $m_1(s)$ and nonincreasing in $m_0(s)$ for each s, and $\gamma_d'(s) \geq 0$, $m_{\gamma'}^{LB}$ minimizes T over $\mathcal{H}(m|\gamma')$ and $m_{\gamma'}^{UB}$ maximizes it.

ii) Fix $\gamma' \in \mathcal{H}(\gamma)$ such that $\mathcal{H}(m|\gamma') \neq \emptyset$. Under Assumption TI, $m_1 = m_0 =: m$. By Theorem 1, $m(s) \in [L_d(s), U_d(s)]$ for both $d \in \{0,1\}$ and all $s \in \mathcal{S}$. Taking the intersection,

$m(s) \in [m^{L,\gamma'}(s), m^{U,\gamma'}(s)]$ for all $s \in \mathcal{S}$. Since $m^{TI,L,\gamma'}(s), m^{TI,U,\gamma'}(s) \in [m^{L,\gamma'}(s), m^{U,\gamma'}(s)]$ for all $s \in \mathcal{S}$ by construction, it follows that $m^{LB}_{\gamma'}, m^{UB}_{\gamma'} \in \mathcal{H}(m \mid \gamma')$. Under Assumption TI:

$$T(m, \gamma') = \sum_{s \in \mathcal{S}} m(s)(\gamma_1'(s) - \gamma_0'(s)). \tag{61}$$

For each s, if $\gamma_1'(s) \geq \gamma_0'(s)$, then T is nondecreasing in $m(s)$, so the minimum is attained at $m^{L,\gamma'}(s)$ and maximum at $m^{U,\gamma'}(s)$. If $\gamma_1'(s) < \gamma_0'(s)$, then T is nonincreasing in $m(s)$, so the minimum is attained at $m^{U,\gamma'}(s)$ and maximum at $m^{L,\gamma'}(s)$. This yields $m^{TI,L,\gamma'}$ as the minimizer and $m^{TI,U,\gamma'}$ as the maximizer. □

References

Artstein, Zvi. 1983. "Distributions of random sets and random selections." *Israel Journal of Mathematics* 46:313–324.

Athey, Susan, Raj Chetty, and Guido Imbens. 2025. "Using Experiments to Correct for Selection in Observational Studies." *arXiv preprint arXiv:2006.09676.*

Beresteanu, Arie, Ilya Molchanov, and Francesca Molinari. 2012. "Partial identification using random set theory." *Journal of Econometrics* 166 (1): 17–32.

Bugni, Federico A, Ivan A Canay, and Xiaoxia Shi. 2017. "Inference for subvectors and other functions of partially identified parameters in moment inequality models." *Quantitative Economics* 8 (1): 1–38.

Manski, Charles F. 1990. "Nonparametric bounds on treatment effects." *The American Economic Review* 80 (2): 319–323.

Molchanov, Ilya, and Francesca Molinari. 2018. *Random Sets in Econometrics.* Vol. 60. Cambridge University Press.

Nocedal, Jorge, and Stephen J Wright. 1999. *Numerical optimization.* Springer.

Shi, Xiaoxia, and Matthew Shum. 2015. "Simple two-stage inference for a class of partially identified models." *Econometric Theory* 31 (3): 493–520.