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

Filip Obradović*

Link to Most Recent Version

July 8, 2025

**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). I show 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 and bring *no identifying power* in their absence. Plausible inference thus hinges on justifiable restrictions on the temporal link functions. Motivated by this, I introduce two *treatment response* assumptions that may be defensible based on economic theory or intuition. To utilize them, I develop a novel two-step identification framework that computationally produces sharp bounds on the LTE for a general class of temporal link function restrictions and allows for imperfect experimental compliance – also extending existing approaches. I apply the method to estimate the long-term effects of Head Start participation. The findings indicate that the effects on educational attainment, employment, and crime involvement are lasting but smaller in magnitude than those established by sibling comparisons.

*Northwestern University, Department of Economics. Email: obradovicfilip@u.northwestern.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, Eric Auerbach, Denis Chetverikov, Piotr Dworczak, Danil Fedchenko, Joel Horowitz, Matías Martinez, Jana Obradović, Alexander Torgovitsky, and Gabriel Ziegler for valuable suggestions. I thank ICPSR for providing the data and the participants at the 2024 Junior Econometrics Conference at Notre Dame, and seminars at Bocconi, Booth, Duke, Emory, Mannheim, Michigan, Northwestern, UPenn, UCLA, and Yale for comments. Financial support from the Robert Eisner Memorial Fellowship is gratefully acknowledged.

# 1 Introduction

Identification of the long-term average treatment effect (henceforth LTE) is an important goal in economics and various other fields of science. For example, one may be interested in the effects of childhood intervention on earnings in adulthood; the impact of conditional cash transfers early in life on employment prospects; or the adverse/protective effects of vaccination years after administration. Gupta et al. (2019) explain that identifying the LTE is also recognized as an important challenge by researchers in the private sector.

Point identification of the LTE is commonly done using observational data (for examples, see Currie and Almond (2011), Hoynes and Schanzenbach (2018)). However, observational studies critically rely on identifying assumptions that may often be deemed implausible. While randomized controlled trials (RCTs) can eliminate the need for such assumptions, long-term experiments may be prohibitively costly or infeasible.[1] Short-term RCTs may be more feasible, but they do not reveal the long-term outcomes and hence the LTE. Nevertheless, short-term RCTs may complement observational data.

Motivated by this, a large body of recent work following Athey, Chetty, and Imbens (2020) and Athey et al. (2024) aims to provide alternatives to conventional observational studies that rely on a combination of: 1) a long-term observational dataset with non-randomized treatment assignment; 2) a short-term experimental dataset with unobserved long-term outcomes.[2] Pursuing point identification, this literature commonly imposes assumptions on the selection mechanism in the observational data, mirroring conventional observational studies. Ghassami et al. (2022), Van Goffrier, Maystre, and Gilligan-Lee (2023) and Imbens et al. (2024) argue that frequently used assumptions may not hold in contexts of economic interest; Park and Sasaki (2024a) show that they are incompatible with common selection models, including the Roy model. It is acknowledged that selection assumptions may broadly be challenging to justify based on economic theory (Manski (1997)).

This paper makes three main contributions. First, it uncovers the roles of the experimental data in this context. For identification of the LTE, the experimental data serve only to amplify the identifying power of restrictions on *temporal link functions* – means of long-term potential outcomes conditional on short-term potential outcomes – henceforth referred to as modeling assumptions. Therefore, plausible and informative inference hinges on imposing justifiable modeling assumptions. Motivated by this, the paper introduces two *treatment response* modeling

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.

2. Structural modeling with experimental data predates this work (Todd and Wolpin (2006), Attanasio, Meghir, and Santiago (2012), García et al. (2020), Todd and Wolpin (2023)). The focus here is on "reduced form" methods.

assumptions that may be defensible based on economic theory or intuition, as the second contribution. Third, it develops a novel two-step identification framework that computationally produces the smallest possible, or *sharp*, bounds on the LTE under a general class of restrictions on temporal link functions. This enables the utilization of the two assumptions and facilitates the implementation and development of new restrictions. Additionally, the framework allows for imperfect experimental compliance and subsumes relevant restrictions stemming from previously proposed identifying assumptions. This extends existing approaches by allowing them to account for imperfect compliance.

The first contribution of the paper is that it uncovers the roles of the experimental data and modeling assumptions, which relates to ongoing discussions (see Remark 5 and Park and Sasaki (2024b)). I show that experimental data provide no identifying power, per se; the identified sets for the LTE obtained from combined and solely observational data are equal in the absence of restrictions on temporal link functions. Modeling assumptions are thus *necessary* to leverage the experimental data and to identify the sign of the LTE. When these assumptions are imposed, the identified set based on combined data is a subset of the one that uses only observational data. It need not be a strict subset. Hence, the experimental data serve to potentially, but not necessarily, *amplify* the identifying power of the modeling assumptions. These observations reveal the *auxiliary* role of experimental data. Modeling assumptions are *central* under data combination, mirroring the prominence of identifying assumptions in observational studies.

I illustrate this point via a selection assumption that is often used in the literature relying on data combination – latent unconfoundedness (LUC), introduced by Athey, Chetty, and Imbens (2020). I show that LUC may commonly have identifying power for the LTE using observational data alone. When experimental data are added, LUC can have more identifying power, point identifying the LTE. The experimental data may thus amplify the identifying power of LUC. In extreme cases, observational data point identify the LTE under LUC even without experimental data. Then, the experimental data do not provide additional identifying power. Therefore, the experimental data potentially, but not necessarily, amplify the identifying power of LUC.

The amplifying role of the experimental data highlights the importance of justifiable modeling assumptions. Under misspecified assumptions, the identified set obtained using combined data can never be closer to the true LTE than the set obtained using only observational data. If the imposed assumptions are challenging to justify based on economic substantives, it may be preferable to discard the experimental data. It may be possible to obtain informative bounds on the LTE using observational data only, and these bounds may *only* be closer to the true value of the parameter than if combined data are used. This finding is an application of a more general lemma. The lemma states that whenever two misspecified identified sets are nested, the smaller one must be at least as far from the truth as the larger one. This result is strikingly simple, but

it appears that it was not formalized before.

The central role of modeling assumptions motivates the main identification results. To summarize them, let $Y(d) \in \mathcal{Y}$ and $S(d) \in \mathcal{S}$ denote long- and short-term potential outcomes under treatment $d \in \{0, 1\}$ and let:

$$m_d(s) := E[Y(d)|S(d) = s]$$
$$\gamma_d := P(S(d)).$$

I refer to $m_d(s)$ as the *temporal link functions*, while $\gamma_d$ are the distributions of short-term potential outcomes $S(d)$. We can then write the LTE using the identity:

$$LTE := E[Y(1) - Y(0)] = \underbrace{\int_{\mathcal{S}} m_1(s) d\gamma_1(s)}_{E[Y(1)]} - \underbrace{\int_{\mathcal{S}} m_0(s) d\gamma_0(s)}_{E[Y(0)]}, \tag{1}$$

For the second contribution, I introduce two assumptions on temporal link functions $(m_0, m_1)$ which are defensible based on economic theory or intuition – *latent monotone instrumental variables (LIV)* and *treatment invariance (TI)*. LIV asserts that functions $m_d$ are non-decreasing in each component of the short-term potential outcomes for any $d$. That is, the means of long-term potential outcomes are non-decreasing in the short-term potential outcomes. It is related to the *monotone instrumental variable* assumption of Manski and Pepper (2000). LIV may be interpreted as maintaining that the latent potential outcomes $S(d)$ are themselves monotone instrumental variables. TI posits that the temporal link functions are invariant to the treatment – $m_1 = m_0$. In other words, TI states that the relationship between the short-term potential outcomes and the mean long-term potential outcome is unaffected by the treatment. While LIV may be justifiable based on intuition, TI is implied by a model. For example, TI would hold under the model proposed by García et al. (2020) in the context of early childhood intervention. LIV and TI do not impose any restrictions on the selection mechanism and represent assumptions on treatment response. In contrast, existing work primarily imposes restrictions on the selection mechanism.

The third contribution is a novel two-step identification framework that operationalizes a broad class of modeling assumptions, including LIV and TI. In the first step, I find all $(m_0, m_1, \gamma_0, \gamma_1)$ compatible with the data and assumptions. In the second step, I collect all values for the LTE that possible $(m_0, m_1, \gamma_0, \gamma_1)$ produce via (1), which yields the identified set for the LTE. The LTE is point identified whenever this set is a singleton. If it is not a singleton, the LTE may be partially identified. Either is permitted, and which occurs depends on the imposed assumptions and the observed data distributions. It should be emphasized that in either

case, all produced bounds are the smallest possible under the maintained assumptions, or sharp.

In the first step, I find the set of all possible $(m_0, m_1, \gamma_0, \gamma_1)$ under a generic restriction on $(m_0, m_1)$, which embodies modeling assumptions maintained by the researcher. This is done via appropriately defined random sets, extending the arguments in Beresteanu, Molchanov, and Molinari (2012). To operationalize the result, I combine two concepts from random set theory – Artstein's inequalities and the conditional Aumann expectation. This characterizes the restrictions on $(m_0, m_1, \gamma_0, \gamma_1)$ via a collection of moment conditions. The two concepts are commonly used in isolation, or combined when the conditioning variable in the Aumann expectation is observed (Chesher and Rosen (2017), Chesher and Rosen (2020)). A distinguishing feature of the setting here is that the conditioning variable is latent and thus itself a measurable selection of a random set.

In the second step, I collect all values that possible $(m_0, m_1, \gamma_0, \gamma_1)$ produce via (1), which yields the identified set for the LTE. I prove that this set can be characterized as an interval bounded by solutions to two *generalized bilinear programs* under the proposed and existing assumptions (Al-Khayyal (1992)). This characterization leads to tractable estimators based on results in Shi and Shum (2015) and Russell (2021). Bilinear programs are non-convex and solvable to provable optimality by modern spatial branch-and-bound algorithms (Gurobi Optimization (2024)). However, they are computationally more demanding than linear programs that are commonly used to characterize identified sets. To alleviate the computational burden, I reduce the number of constraints in the optimization problems via the concept of *core determining classes* (Galichon and Henry (2011)). I further demonstrate that programs may be decomposed into more tractable separable subprograms or bilevel (nested) programs where the inner optimization programs have closed-form solutions.

The two-step identification framework has additional appealing features, beyond enabling the use of LIV and TI. It can computationally produce sharp bounds under a broad class of modeling assumptions representable as constraints on $(m_0, m_1)$ in the optimization problems. The identification results thus provide a tool for researchers to characterize identified sets under new assumptions tailored to their empirical setting, without requiring proofs of sharpness. For similar results in different settings, see Mogstad, Santos, and Torgovitsky (2018), Torgovitsky (2019), Russell (2021), Kamat (2024) and references therein. The approach developed here may also be of independent interest in other settings as it facilitates tractable utilization of assumptions restricting latent conditional means.

One additional important advantage of the framework is that it can accommodate imperfect compliance in the experimental data by allowing partial identification of the subvector $(\gamma_0, \gamma_1)$. Accommodating imperfect compliance is of great practical relevance. Compliance issues are

prevalent in RCTs, and especially in the experiments previously used in this context.[3] Moreover, often considered alternative parameters under non-compliance, such as the intent-to-treat effect (ITT) and local average treatment effect (LATE) Imbens and Angrist (1994) pertaining to experimental treatment offer are unidentified in this setting because the long-term outcomes are never observed in the experimental data.[4] However, despite its practical relevance and dearth of identified alternative target parameters, related literature did not explicitly account for imperfect experimental compliance, to the extent of my knowledge. Since corresponding modeling assumptions stemming from previously proposed identifying assumptions may be represented as restrictions on $(m_0, m_1)$, the framework also extends existing approaches by allowing them to account for imperfect compliance.

I apply the method to estimate the long-term effects of participation in Head Start, the largest federally funded early childhood education program in the United States. To do so, I combine the 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 positive 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 2.4%, and decrease the probability of repeating a grade by 1.2% to 5.3%. The program is also estimated to lower the probability of idleness (neither working nor in school) by 2.8% to 4.2% and criminal involvement by 1.3% to 3.9%. The findings suggest that the effects of Head Start are lasting but smaller in magnitude than those reported by sibling comparison studies (Deming (2009)).

Section 2 introduces the setting, summarizes the roles of the experimental data and modeling assumptions, and introduces LIV and TI. Section 3 characterizes the identified set and a consistent estimator. Section 4 details the roles of experimental data and modeling assumptions. Section 5 provides the empirical illustration. Section 6 concludes. Appendix A contains the extensions of the findings, and Appendix B collects the proofs.

## 2 Setting and 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

3. Athey, Chetty, and Imbens (2020) and Park and Sasaki (2024a) use the Project STAR and Aizer et al. (2024) the Job Corps RCT. Both had significant treatment reassignment/non-compliance (e.g. see Chen, Flores, and Flores-Lagunes (2018) and Russell (2021)). In the empirical illustration, 16.2% of individuals fail to comply.

4. Even when identified, ITT or LATE may or may not be of interest, depending on the research question. For more details see discussions in Deaton (2009), Heckman and Urzua (2010) and Imbens (2010).

treatment $d \in \{0,1\}$, respectively.[5] Denote the realized treatment by $D \in \{0,1\}$. The observed outcomes are:

$$\begin{aligned} Y &= DY(1) + (1-D)Y(0) \\ S &= DS(1) + (1-D)S(0). \end{aligned} \tag{2}$$

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) $\tau(x)$:

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

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. Throughout the paper, I assume $E[|Y(d)|] < \infty$ for $d \in \{0,1\}$, which ensures that the parameters are well defined.

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

## 2.1 Observed Data

As in Athey, Chetty, and Imbens (2020), I maintain the existence of a population divided into two subpopulations from which the two datasets are randomly drawn: a short-term experimental and a long-term observational dataset. Let $G \in \{O, E\}$ be the indicator for the subpopulation, where $G = O$ generates the observational and $G = E$ the experimental dataset.[6] Let $Z \in \mathcal{Z}$ be an exogenous (i.e. randomly assigned) instrument in the experiment, inducing individuals into treatment. In the experimental dataset, the researcher observes $(S, D, X, Z)$, but not $Y$. In the observational dataset, $(Y, S, D, X)$ are observed, but $Z$ is absent as there is no instrument in the observational data.

Usually, $Z \in \{0,1\}$, representing random assignment to the treatment or the control group. The identification analysis can accommodate bounded $\mathcal{Z}$ with multiple or even a continuum of points $\mathcal{Z} = [0,1]$, as in Heckman and Vytlacil (1999). For expositional simplicity, I refer to experiments with $P(D = Z|G = E) < 1$ as having imperfect compliance, as opposed to

5. Supports are invariant to the treatment. This can be relaxed at the expense of more complicated notation.

6. This setting has become common. See also García et al. (2020), Athey, Chetty, and Imbens (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), Aizer et al. (2024) and Imbens et al. (2024).

perfect compliance when $P(D = Z|G = E) = 1$. I thus also refer to $Z$ as treatment assignment regardless of its support, keeping in mind that the $\mathcal{Z}$ may contain points beyond $\{0, 1\}$.

The main purpose of $Z$ is to allow for imperfect experimental compliance, which is practically relevant. This represents a critical distinction between the setting of this paper and related existing work. Researchers often obviate experimental compliance issues by focusing on parameters such as the $ITT_z^{z'} := E[Y|Z = z', G = E] - E[Y|Z = z, G = E]$ and $LATE_z^{z'} = \frac{ITT_z^{z'}}{E[D|Z=z',G=E]-E[D|Z=z,G=E]}$ pertaining to experimental treatment assignments $z', z \in \mathcal{Z}$. Identification of these parameters requires jointly observing $Z$ and the long-term outcomes $Y$. Since $Z$ is never jointly observed with $Y$ in this setting, both parameters are unidentified.

**Example 1 (continued).** The observational dataset is the Child and Young Adult Supplement to the National Longitudinal Survey of Youth, and the experimental dataset is the Head Start Impact Study (HSIS). $D$ is the indicator for true participation. In the HSIS, $Z = 1$ 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 $D \neq Z$.

I maintain the following assumptions throughout the paper.

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

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

Assumption RA holds if $Z$ in the experimental data is randomly assigned. It is a standard assumption in the program evaluation literature. $D \not\perp\!\!\!\perp (Y(1), Y(0), S(1), S(0))|X, G = g$ is permitted for any $g \in \{O, E\}$. This is expected in the observational dataset, and in the experimental data under imperfect compliance. When compliance is perfect, Assumption RA implies $D \perp\!\!\!\perp (Y(1), Y(0), S(1), S(0))|X, G = E$. I do not assume that $P(D = 1|G = g) \in (0, 1)$ for any $g \in \{0, 1\}$. Instead, $P(D = 1|G = g) \in [0, 1]$ which may be relevant for $g = O$ when a certain treatment is only available in the experiment. This is the case with some "model" early childhood intervention programs or novel vaccines.

Assumption EV is a standard assumption in the data combination literature, linking the two datasets. It states that the subpopulations generating them do not differ in terms of counterfactual distributions (conditional on $X$). It holds when participants are randomly recruited into the datasets from the same population (conditional on $X$).

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:

$$\tau = E[Y(1) - Y(0)] \tag{4}$$

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

**Notation:** I denote laws of random elements using subscripts when the element needs to be specified (e.g. $P_{S(d)}$ is the law of $S(d)$). If the random element is clear from the context, 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 $E[\cdot|\mathcal{E}, G = g] = E_g[\cdot|\mathcal{E}]$ and $V[\cdot|\mathcal{E}, G = g] = V_g[\cdot|\mathcal{E}]$.

## 2.2 Identification Preliminaries

This paper proposes a novel identification approach. To introduce it, recall that for $s \in \mathcal{S}$ and $d \in \{0, 1\}$:

$$m_d(s) := E[Y(d)|S(d) = s] \tag{5}$$

$$\gamma_d := P_{S(d)}. \tag{6}$$

I refer to $m_d(s)$ as *temporal link functions*, since they "link" the short-term and long-term potential outcomes in a way that is meaningful for identification of $\tau$. We can write the parameter of interest as:

$$\tau = E[Y(1) - Y(0)] = \int_{\mathcal{S}} m_1(s) d\gamma_1(s) - \int_{\mathcal{S}} m_0(s) d\gamma_0(s). \tag{7}$$

Denote the pair of temporal link functions $m := (m_0, m_1)$, and the pair of short-term potential outcome distribution functions by $\gamma := (\gamma_0, \gamma_1) = (P_{S(0)}, P_{S(1)})$. Observe that $\gamma$ consists of the marginal distributions $P_{S(d)}$, and is not the joint-distribution function $P(S(0), S(1))$. Given functions $(m, \gamma)$, the corresponding value of $\tau$ follows by (7). Relying on this, the approach identifies $(m, \gamma)$ as an intermediate step towards identifying $\tau$.

As mentioned in the introduction, this approach two benefits. First, it will computationally produce sharp bounds for a broad class of modeling assumptions, removing the need for proving sharpness for each assumption. Second, it allows one to account for imperfect compliance in the experiment by permitting partial identification of $\gamma$, which is of great practical relevance. To formalize the class of modeling assumptions, let $\mathcal{M}$ be the set of all temporal link functions, i.e. measurable functions mapping $\mathcal{S} \times \mathcal{S} \to \mathcal{Y} \times \mathcal{Y}$.[7] I assume that the researcher knows or can identify the subset $\mathcal{M}^A \subseteq \mathcal{M}$ to which $m$ belongs, which represents a generic modeling assumption.

7. More precisely, $\mathcal{M}$ 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 $\varsigma : \Omega \to \mathcal{S} \times \mathcal{S}$.

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

Modeling assumptions may be classified as: *selection assumptions*, restricting the relationship between $(Y(1), Y(0), S(1), S(0))$ and $D$; and *treatment response assumptions*, restricting how $(Y(1), Y(0), S(1), S(0))$ are related to each other.

Assumption MA can accommodate both treatment response and selection assumptions. I will introduce two treatment response assumptions in Section 2.3. Remark 1 explains that Assumption MA nests relevant restrictions from existing selection assumptions and approaches. Thus, the identification framework will directly extend previously proposed approaches by allowing for imperfect compliance under the corresponding modeling assumptions.

Let $\mathcal{H}(\cdot)$ be the identified set for a specified parameter. Finding all $(m, \gamma)$ consistent with the data and maintained assumptions, including any restriction in the form of Assumption MA, yields $\mathcal{H}(m, \gamma)$. In turn, by the identity (7), $\mathcal{H}(\tau)$ follows directly. To this end, define the functional $T : \mathcal{M} \times \mathcal{P}^{\mathcal{S}} \times \mathcal{P}^{\mathcal{S}} \to \bar{\mathbb{R}}$, where $\mathcal{P}^{\mathcal{S}}$ collects distribution functions supported on $\mathcal{S}$:

$$T(m, \gamma) = \int_{\mathcal{S}} m_1(s) d\gamma_1(s) - \int_{\mathcal{S}} m_0(s) d\gamma_0(s). \tag{8}$$

By definition, the identified set $\mathcal{H}(\tau)$ is then equivalent to the set of values $T$ can produce over the identified set $\mathcal{H}(m, \gamma)$:

$$\mathcal{H}(\tau) := \{T(m, \gamma) : (m, \gamma) \in \mathcal{H}(m, \gamma)\}. \tag{9}$$

Section 3 constructs $\mathcal{H}(m, \gamma)$, and develops a tractable characterization and estimators of $\mathcal{H}(\tau)$.

## 2.3 Modeling Assumptions

Section 4 provides a detailed discussion on the roles of experimental data and modeling assumptions. It reveals that restrictions in the form of Assumption MA are *central* for identification of $\tau$ under data combination, mirroring the prominence of identifying assumptions in conventional observational studies. Experimental data bring *no identifying power* in the absence of restrictions on $m$, and serve only to *amplify* their identifying power. Hence, if a modeling assumption fails, bounds on $\tau$ obtained using just observational data may only be closer to the truth than the bounds obtained using combined data. Therefore, plausible inference hinges on plausible modeling assumptions, despite the use of experimental data.

As mentioned previously, Ghassami et al. (2022), Van Goffrier, Maystre, and Gilligan-Lee (2023) and Imbens et al. (2024) argue that frequently used assumptions may fail in contexts of

economic interest; Park and Sasaki (2024a) indicate that they are incompatible with standard models of selection. I thus propose treatment response assumptions that may be defensible based on economic theory or intuition. These assumptions rely on the identification approach for implementability. Recall that $S(d)$ is a vector of dimension $d_s \geq 1$.

**Assumption OLIV.** *(One-dimensional Latent Monotone Instrumental Variable) Let* $d_s = 1$. *For any* $m \in \mathcal{M}^A$ *and* $s, s' \in \mathcal{S}$ *such that* $s < s'$ *it holds that* $m_d(s) \leq m_d(s')$ *for* $d \in \{0, 1\}$.

Assumption OLIV states that the mean of the long-term *potential outcome* $Y(d)$ is non-decreasing conditional on the scalar short-term *potential outcome* $S(d)$. This is a restriction on the underlying counterfactuals that may have an intuitive economic interpretation.

**Example 2.** *(LIV and Head Start)* Suppose that there is a single childhood cognitive test score. Assumption OLIV states that people with higher *potential* test score $S(d)$, on average, also have weakly higher *potential* earnings in adulthood $Y(d)$. In other words, under an exogenously fixed treatment, people with higher test scores would have weakly higher average adulthood earnings.

More generally, $S(d)$ may have dimension higher than one, so that $d_s > 1$. To generalize the assumption, denote by $m_d(s_j, s_{-j}) := E[Y(d)|S_j(d) = s_j, S_{-j}(d) = s_{-j}]$ where $S_{-j}(d) \in \mathcal{S}_{-j}$ is a subvector of $S(d) \in \mathcal{S}$ where the $j-th$ component $S_j(d) \in \mathcal{S}_j$ is omitted.

**Assumption LIV.** *(Latent Monotone Instrumental Variables) For any* $m \in \mathcal{M}^A$, $j \in \{1, \ldots, d_s\}$, $s_{-j} \in \mathcal{S}_{-j}$ *and* $s_j, s'_j \in \mathcal{S}_j$ *such that* $s_j < s'_j$ *it holds that* $m_d(s_j, s_{-j}) \leq m_d(s'_j, s_{-j})$ *for* $d \in \{0, 1\}$.

Assumption LIV states that the conditional mean of $Y(d)$ is non-decreasing in any individual short-term potential outcome $S_j(d)$. It is immediate that Assumptions LIV and OLIV are equivalent when $d_s = 1$.

**Example 2 (continued).** Suppose there are multiple childhood cognitive test scores. LIV states that people with higher *potential* childhood test scores $S_j(d)$, on average, also have weakly higher *potential* earnings in adulthood $Y(d)$, holding remaining scores $S_{-j}(d)$ fixed for $j \in \{1, \ldots, d_s\}$.

One can symetrically assume that $E[Y(d)|S_j(d) = s_j, S_{-j}(d) = s_{-j}]$ is non-increasing in $s_j$ for any sub-collection of elements of $s$. Results follow directly by defining $\tilde{S}_j(d) = -S_j(d)$ and observing that $E[Y(d)|\tilde{S}_j(d) = s_j, S_{-j}(d) = s_{-j}]$ satisfies LIV.

LIV is related to the monotone instrumental variable (MIV) assumption of Manski and Pepper (2000) (see also Manski and Pepper (2009)). MIV maintains that there exists a variable $V \in \mathcal{V}$ such that $E[Y(d)|V = v]$ is non-decreasing in $v \in \mathcal{V}$, which is observed for *all* individuals. The critical distinction is that the conditioning variables in Assumption LIV are latent counterfactuals. This introduces further complexity, which will be addressed by the identification approach.

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

The assumption intuitively states that the relationship between the *potential* outcomes $S(d)$ and mean long-term *potential* outcomes $Y(d)$ does not vary with the underlying treatment $d$.

**Example 4.** *(TI and Head Start)* TI follows from previously used models in the context of early childhood intervention. Consider the following separable model of potential earnings:

$$Y(d) = \phi_d(S(d)) + \varepsilon_d = \phi(S(d)) + \varepsilon_d, \;\; \varepsilon_1 \sim \varepsilon_0, \;\; \varepsilon_{d'} \perp\!\!\!\perp S(d), \forall d, d' \in \{0,1\}. \tag{10}$$

If $S(d)$ in this model is a vector of short-term potential outcomes including test scores and measures of non-cognitive skills. $S(d)$ represents inputs in the production function $\phi_d$ for $Y(d)$. The production function $\phi_d$ and the distributions of unobservables $\varepsilon_d$ 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 utilize TI whenever they find the model from Example 4 to be plausible. For example, based on mediation results in Heckman, Pinto, and Savelyev (2013) and extensive falsification testing, García et al. (2020) argue the plausibility of a similar model when the treatment is an early childhood intervention. They then identify $\tau$ by combining observational and experimental data in the special case where $P_O(D = 0) = 1$ and compliance is perfect, i.e. when there is no selection in either dataset. This paper provides an extension of their approach by demonstrating that one may use implications of the same model to bound $\tau$ when there is selection in either dataset.

In the special case of perfect compliance, TI is implied by the statistical surrogacy assumption of Prentice (1989) – $Y \perp\!\!\!\perp S|D, G = E$. Appendix A.1.4 explains the differences. However, researchers may still wish to assign the informal interpretation of the surrogacy assumption to Assumption TI. Intuitively, one may choose to say that the treatment affects the mean long-term outcome only through the short-term outcomes.

To connect the findings of this paper with previous results, I will refer to an established selection assumption introduced by Athey, Chetty, and Imbens (2020).

**Assumption LUC.** *(Latent Unconfoundedness) For all* $d \in \{0,1\} : Y(d) \perp\!\!\!\perp D|S(d), G = O$.

According to Chen and Ritzwoller (2023): "Informally, LUC states that all unobserved confounding in the observational sample is mediated through the short-term outcomes". Park and Sasaki (2024a) describe it as a "statistical assumption" and indicate that it is difficult to interpret economically outside of restricted non-parametric selection models.

Previous work notes that Assumption LUC may be untenable in economic contexts such as early childhood interventions and job-training programs. In the former case, parental interference and the child's inherent ability may be confounding factors for $(D, S(d), Y(d))$, invalidating the assumption. In the latter, the confounding factors may be worker's innate motivation and resourcefulness. For more details see Ghassami et al. (2022) and Imbens et al. (2024). For examples of its use, see Hu, Zhou, and Wu (2022), Park and Sasaki (2024b), Aizer et al. (2024).

**Remark 1.** Existing approaches are subsumed under Assumption MA. For example, relevant restrictions for identification of $\tau$ under Assumption LUC can be restated as $\mathcal{M}^{LUC} = \{m \in \mathcal{M} : m_d(s) = E_O[Y|S = s, D = d], \forall s \in \mathcal{S}\}$. One can do the same for the outcome bridge function approach of Imbens et al. (2024, Theorem 1). Let $S_t$ for $t \in \{1, 2, 3\}$ be subvectors of $S$. Then under the corresponding assumptions: $\mathcal{M}^{Bridge} = \{m \in \mathcal{M} : m_d(s_3, s_2) = h(s_3, s_2, d),\ h \text{ solves } E_O[Y|S_2, S_1, D] = E_O[h(S_3, S_2, D)|S_2, S_1, D]\}$.

# 3 Main Results

Section 3.1 summarizes the main identification results and the underlying intuition behind the two-step identification approach; technical discussions follow. Section 3.2 characterizes $\mathcal{H}(m, \gamma)$. Section 3.3 provides a tractable implementation of $\mathcal{H}(\tau)$ based on this characterization. Section 3.4 proposes a consistent estimator for $\mathcal{H}(\tau)$.

## 3.1 Identification Intuition

The identification approach aims to operationalize the proposed modeling assumptions. However, it also presents a novel challenge; it necessitates finding $\mathcal{H}(m, \gamma)$ as an intermediate step. Both $m_d(s) = E[Y(d)|S(d) = s]$ and $\gamma_d = P_{S(d)}$ are features of *latent* random variables $Y(d)$ and $S(d)$, and thus $(m, \gamma)$ are not directly revealed by the data. The identification results exploit this apparent complexity to construct $\mathcal{H}(m, \gamma)$. By characterizing the feasible potential outcomes, the corresponding $(m, \gamma)$ follow by definition.

There will exist a set of random vectors $(S(0), S(1), Y(0), Y(1))$ that are consistent with the data and maintained assumptions because the potential outcomes are latent. Concretely, let $\mathcal{Q}$ be the set of all $(S(0), S(1), Y(0), Y(1))$ that are consistent with the data, and Assumptions RA and EV. The researcher can determine $\mathcal{Q}$. To find $\mathcal{H}(m, \gamma)$, one then only needs to collect all corresponding $(m, \gamma)$ such that they additionally satisfy the modeling assumption $m \in \mathcal{M}^A$. By definition:

$$\mathcal{H}(m,\gamma) = \left\{ (m,\gamma) : \begin{array}{c} \overbrace{m \in \mathcal{M}^A,}^{\textit{Modeling assumption}} \quad \overbrace{\exists (S(0), S(1), Y(0), Y(1)) \in \mathcal{Q}}^{\textit{Data + Assumptions RA/EV}} \\ \underbrace{\forall d \in \{0,1\} : \gamma_d \overset{d}{=} S(d),\ m_d(S(d)) = E[Y(d)|S(d)] \text{ a.s.}}_{(m,\gamma) \textit{ correspond to } S(d) \textit{ and } Y(d)} \end{array} \right\}. \tag{11}$$

Then, again definitionally, $\mathcal{H}(\tau) = \{T(m,\gamma) : (m,\gamma) \in \mathcal{H}(m,\gamma)\}$. These expressions demonstrate that one may use the information on the potential outcomes to relate $(m,\gamma)$ and $\tau$ to observed data and assumptions. However, while intuitive, the definitions are intractable.

The first main identification result utilizes (11) to provide a general equivalent characterization of $\mathcal{H}(m,\gamma)$ in terms of moment restrictions. When $\mathcal{S}$ and $\mathcal{Z}$ are finite with $|\mathcal{S}| = k$, the general characterization simplifies to:[8]

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(|\mathcal{S}|))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(\max_{z \in \mathcal{Z}} P_E(S=s, D=d|Z=z), P_O(S=s, D=d)\right), \\ (m_d(s) + \inf \mathcal{Y})\,\gamma_d(s) \geq (E_O[Y|S=s, D=d] + \inf \mathcal{Y})\,P_O(S=s, D=d)\ , \\ (\sup \mathcal{Y} - m_d(s))\,\gamma_d(s) \geq (\sup \mathcal{Y} - E_O[Y|S=s, D=d])\,P_O(S=s, D=d) \end{array} \right\} \tag{12}$$

where $\Delta(k)$ denotes the $k$-dimensional simplex. In turn, this yields the second main identification result – characterization of $\mathcal{H}(\tau)$ using optimization problems:

$$\mathcal{H}(\tau) = \left[ \min_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}), \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}) \right]. \tag{13}$$

where the moment conditions definining $\mathcal{H}(m,\gamma)$ take the role of the constraint set.

Therefore, the researcher may determine $\mathcal{H}(\tau)$ by solving two constrained optimization problems. Beyond producing sharp bounds under the previously introduced assumptions, (13) also provides a tool for researchers to computationally obtain sharp bounds on $\tau$ under tailor-made modeling assumptions. To do so, it is sufficient to solve the optimization problems with appropriately defined constraints based on $\mathcal{M}^A$.

Computational characterizations of identified sets have been exploited previously to obviate the need to prove sharpness for each set of assumptions; for recent examples in different settings, see Mogstad, Santos, and Torgovitsky (2018), Torgovitsky (2019), Russell (2021), Kamat (2024) and references therein. Commonly, the corresponding optimization problems are linear or have linear equivalents. Here, both the objective $T$ and the constraints imposed by $\mathcal{H}(m,\gamma)$ are

8. Conceptually, the identification results do not require $\mathcal{S}$ or $\mathcal{Z}$ to be finite. The former allows one to computationally characterize the set in full. Both will be assumed to construct the consistent estimator.

bilinear in $(m, \gamma)$ under existing and proposed assumptions. Hence, the optimization problems in (13) are *generalized bilinear programs* (see Al-Khayyal (1992)). Bilinear problems also appear in Dutz et al. (2021) and Shea (2022).

## 3.2 Identification of $(m, \gamma)$

This section represents $\mathcal{H}(m, \gamma)$ in terms of moment restrictions. This set is conducive to tractable implementation of $\mathcal{H}(\tau)$. To present the result, I introduce the necessary basic definitions from random set theory specialized to finite-dimensional Euclidean spaces. Appendix B.1.1 contains a more complete but brief overview of the results used in the proofs. I henceforth maintain that all random elements are defined on a common non-atomic probability space $(\Omega, \mathcal{F}, P)$.[9]

**Notation**: $A$, $B$ and $K$ represent sets. $\mathcal{K}(A)$, $\mathcal{C}(A)$, and $\mathcal{B}(A)$ are the families of all compact, closed, and Borel subsets of the set $A$, respectively. $co(A)$ is the closed convex hull of the set $A$. I write random sets using boldface letters (e.g. $\mathbf{Y}$), and $\mathbf{Y} \times \mathbf{X}$ as $(\mathbf{Y}, \mathbf{X})$.

**Definition 1.** A measurable map $\mathbf{R} : \Omega \to \mathcal{C}(\mathbb{R}^d)$ is called a *random (closed) set.*[10]

**Definition 2.** A random vector $R : \Omega \to \mathbb{R}^d$ 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.

**Definition 3.** If the random variable $||\mathbf{R}|| = sup\{||R|| : R \in Sel(\mathbf{R})\}$ is integrable $E[||\mathbf{R}||] < \infty$, then the random set $\mathbf{R}$ is said to be *integrably bounded.*

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

$$\mathbf{Y}(d) := \begin{cases} \{Y\}, \text{ if } (D, G) = (d, O) \\ \mathcal{Y}, \text{ otherwise} \end{cases}, \quad \mathbf{S}(d) := \begin{cases} \{S\}, \text{ if } (D, G) \in \{(d, E), (d, O)\} \\ \mathcal{S}, \text{ otherwise} \end{cases}. \tag{14}$$

$\mathbf{Y}(d)$ and $\mathbf{S}(d)$ serve to summarize information on the counterfactuals $(S(0), S(1), Y(0), Y(1))$, and thus $(m, \gamma)$, contained in the data and assumptions. Their properties lead to the following result.

**Theorem 1.** *Let Assumptions RA, EV and MA hold. If* $\mathbf{Y}(d)$ *is integrably bounded, the identified*

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

*set for* $(m, \gamma)$ *is:*

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \forall B \in \mathcal{C}(\mathcal{S}), \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ \forall u \in \{-1,1\}\text{: } um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s))\ \gamma_d\text{−}a.e. \end{array} \right\} \quad (15)$$

*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* $\forall B \in \mathfrak{C}$.

Theorem 1 equivalently characterizes $\mathcal{H}(m, \gamma)$ for any modeling restriction in the form $m \in \mathcal{M}^A$ via moment restrictions that are identified by the data. This includes, but is not limited to, assumptions and approaches in Section 2.3 and Remark 1. It also offers the possibility of computational simplifications via the concept of *core-determining classes* – sub-families of $\mathcal{C}(\mathcal{S})$ which are sufficient to completely characterize $\gamma_d$ (Galichon and Henry (2011)). Informally, a core determining class allows one to remove redundant restrictions on each $\gamma_d$, without any loss of information, which will be beneficial for tractability. Theorem 1 will be used to provide a tractable implementation of $\mathcal{H}(\tau)$ in Section 3.3.

The technical contribution of the theorem lies in jointly identifying conditional means and corresponding distributions of latent random variables via moment restrictions. This necessitates novel arguments that may be of independent interest. Namely, the proof combines *Artstein's theorem* (Artstein (1983, Theorem 2.1)) and the *conditional Aumann expectation* when the relevant conditioning $\sigma$-algebra is generated by a selection of a random set, i.e. a latent random vector. Section 3.2.1 thus sketches how Theorem 1 is obtained. The main results in Section 3.3 do not require these discussions.

### 3.2.1 Mechanics behind Theorem 1

Recalling the intuition, to find all feasible $(m, \gamma)$, one may first summarize the information about the counterfactuals $Y(d)$ and $S(d)$. By definition, random sets $\mathbf{Y}(d)$ and $\mathbf{S}(d)$ express all information on $Y(d)$ and $S(d)$ contained in *the data*, respectively. As Beresteanu, Molchanov, and Molinari (2012) explain, *all* information in the data about $S(d)$ and $Y(d)$ can be expressed as $(S(0), S(1), Y(0), Y(1)) \in Sel((\mathbf{S}(0), \mathbf{S}(1), \mathbf{Y}(0), \mathbf{Y}(1)))$. Intuitively, we can think about random sets $\mathbf{S}(d)$ and $\mathbf{Y}(d)$ as bundles of random vectors and variables, respectively. The data only reveal that $S(d)$ and $Y(d)$ are elements of these bundles, but not which ones exactly.

Assumptions RA, EV and $E[|Y(d)|] < \infty$ further restrict which elements the potential outcomes may be. The counterfactuals are consistent with the three assumptions if and only if $(S(0), S(1), Y(0), Y(1)) \in \mathcal{I}$, where $\mathcal{I}$ is the set of random elements $(E_1, E_2, E_3, E_4)$ such that

$(E_1, E_2, E_3, E_4) \perp\!\!\!\perp Z|G = E$, $(E_1, E_2, E_3, E_4) \perp\!\!\!\perp G$ and $E[|E_3|], E[|E_4|] < \infty$. Therefore, all information about the counterfactuals in the *data* and the *three assumptions* can be expressed by:

$$(S(0), S(1), Y(0), Y(1)) \in Sel((\mathbf{S}(0), \mathbf{S}(1), \mathbf{Y}(0), \mathbf{Y}(1))) \cap \mathcal{I} := \mathcal{Q}.$$

The identified set for $\mathcal{H}(m, \gamma)$ follows by definition as all corresponding $(m, \gamma)$ that additionally satisfy the modeling assumption:

$$\mathcal{H}(m,\gamma) := \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \exists (S(0), S(1), Y(0), Y(1)) \in \mathcal{Q}, \\ \forall d \in \{0,1\},\ \ \gamma_d \overset{d}{=} S(d),\ \ m_d(S(d)) = E[Y(d)|S(d)] \text{ a.s.} \end{array} \right\}. \tag{16}$$

The definition imposes redundant restrictions on $(m, \gamma)$, which preclude the use of appropriate tools needed to obtain moment conditions. The following lemma disposes of such restrictions and is important for explaining how Theorem 1 is obtained.

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

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \ \exists S(d) \in Sel(\mathbf{S}(d)) \cap \bar{I}, \\ \exists Y(d) \in Sel^1(\mathbf{Y}(d)),\ \ \gamma_d \overset{d}{=} S(d),\ \ m_d(S(d)) = E_O[Y(d)|S(d)]\ \ a.s. \end{array} \right\}. \tag{17}$$

*where $\bar{I}$ 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 indicates that, for identification of $(m, \gamma)$ it is sufficient to: 1) consider restrictions on $m_d$ imposed by $Y(d)$ and $S(d)$ only conditional on $G = O$, reflected by $m_d(S(d)) = E_O[Y(d)|S(d)]$; 2) only impose marginal independence conditions $S(d) \perp\!\!\!\perp G$ and $S(d) \perp\!\!\!\perp Z|G = E$, instead of the full joint independence as in Assumption RA and EV.

With Lemma 1, the characterization in Theorem 1 can be constructed. First, for any $S(d) \in Sel(\mathbf{S}(d)) \cap \bar{I}$, collect all conditional expectations $E_O[Y(d)|S(d)]$ over $Y(d) \in Sel^1(\mathbf{Y}(d))$ into a set. This yields the *random* set $\{E_O[Y(d)|S(d)] : \ Y(d) \in Sel^1(\mathbf{Y}(d))\}$. When $\mathbf{Y}(d)$ is integrably bounded, Li and Ogura (1998, Theorem 1) show that this random set is equivalent to the *conditional Aumann expectation* denoted by $\mathbb{E}_O[\mathbf{Y}(d)|S(d)]$.[11] Then, it is easy to see that for a given $S(d)$:

$$\exists Y(d) \in Sel^1(\mathbf{Y}(d)) : \ \ m_d(S(d)) = E_O[Y(d)|S(d)] \text{ a.s.} \Leftrightarrow m_d(S(d)) \in \mathbb{E}_O[\mathbf{Y}(d)|S(d)] \text{ a.s.} \tag{18}$$

$\mathbb{E}_O[\mathbf{Y}(d)|S(d)]$ is convex on non-atomic probability spaces. Therefore, it can be represented using its support function, which equivalently characterizes the condition $m_d(S(d)) \in$

11. The conditional Aumann expectation is defined with respect to any conditioning sub-$\sigma$-algebra $\mathcal{F}_0 \subsetneq \mathcal{F}$. Here, this is the $\sigma$-algebra generated by events $\{\{S(d) \in B\} \cap \{G = O\} : B \in \mathcal{B}(\mathcal{S})\}$, which I keep implicit for ease of notation. See Section B.1.1 for a formal definition.

$\mathbb{E}_O[\mathbf{Y}(d)|S(d)]$ as a set of moment restrictions. For a given $S(d)$ with a distribution $\gamma_d$, (18) holds if and only if:

$$\forall u \in \{-1,1\}\text{: } um_d(s) \leq uE_O[Y|S=s,D=d]\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1-\pi_{\gamma_d}(s)) \ \gamma_d\text{−a.e.} \tag{19}$$

recalling that $h_{co(\mathcal{Y})}(u) = \sup_{y \in co(\mathcal{Y})} uy$. All restrictions on $m_d$ depend on the selection $S(d)$ *only* up to its distribution $\gamma_d$, which will be essential in the next step. Observe that all elements on the right-hand side of (19) are either known or identified from the data, given $\gamma_d$. Notably, $\pi_{\gamma_d}$ is identified given $\gamma_d$, which is evident when $\gamma_d$ is discretely supported. Then, $\pi_{\gamma_d}(s) = \frac{P_O(S=s,D=d)}{\gamma_d(s)}$ for $s \in \mathcal{S}$ (which implies that $\gamma_d(s) > 0$). This intuition extends to the case when $\mathcal{S}$ is not discrete. Note that $P_O(D = d|S'(d) = s) = \pi_{\gamma_d}(s)$, so $\pi_{\gamma_d}$ can be interpreted as the *latent* propensity score, conditioning on a latent vector $S'(d) \overset{d}{=} \gamma_d$ (for other uses see Masten and Poirier (2023)).

The result in (19) removes the need to search over $Y(d) \in Sel^1(\mathbf{Y}(d))$, but the need to search over $S(d) \in Sel(\mathbf{S}(d)) \cap \bar{I}$ remains. However, by Artstein's theorem:

$$\begin{aligned} &\exists S(d) \in Sel(\mathbf{S}(d)) \cap \bar{I} \text{ such that } \gamma_d \overset{d}{=} S(d) \\ \Leftrightarrow &\forall B \in \mathcal{C}(\mathcal{S}) : \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right). \end{aligned} \tag{20}$$

This characterizes the set of distributions $\gamma_d$ such that they are "rationalized" by a selection $S(d)$ satisfying conditions of Lemma 1; (19) characterizes the set of link functions $m_d$ such that they "rationalized" by a selection $Y(d)$ satisfying conditions of Lemma 1, *given* a distribution $\gamma_d$. By putting the two results together, Theorem 1 follows. Hence, $(m, \gamma)$ can be characterized using only moment conditions.

The theorem provides an additional important simplification that reduces the number of conditions for $\gamma_d$ imposed by (20). This is done via a *core-determining class* – a subfamily $\mathfrak{C} \subseteq \mathcal{C}(\mathcal{S})$ sufficient to summarize all restrictions on $\gamma_d$ in (20). More precisely, $\mathfrak{C}$ is a core determining class when:

$$\begin{aligned} &\forall B \in \mathfrak{C} : \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right) \\ \Leftrightarrow &\forall B \in \mathcal{C}(\mathcal{S}) : \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right). \end{aligned} \tag{21}$$

If a core-determining class exists, using it can substantially reduce the number of constraints on $\gamma_d$. This is instrumental in reducing computational burden when determining $\mathcal{H}(\tau)$ in the next step.

### 3.3 Tractable Characterization of $\mathcal{H}(\tau)$

Recall that $\mathcal{H}(m,\gamma)$ yields the identified set $\mathcal{H}(\tau) = \{T(m,\gamma) : (m,\gamma) \in \mathcal{H}(m,\gamma)\}$, where $T(m,\gamma) = \int_{\mathcal{S}} m_1(s)d\gamma_1(s) - \int_{\mathcal{S}} m_0(s)d\gamma_0(s)$. This section operationalizes $\mathcal{H}(\tau)$ using Theorem 1.

To verify if a candidate $(m,\gamma)$ is in the identified set, one must establish that each $\gamma_d$ satisfies an inequality condition for each closed subset $B \in \mathcal{C}(\mathcal{S})$. If $\mathcal{S}$ is infinite, then so is $\mathcal{C}(\mathcal{S})$. Some of these restrictions may be redundant. However, even the smallest set of non-redundant restrictions on $\gamma_d$, i.e. the smallest core-determining class $\mathfrak{C}$, will contain infinitely many sets (Ponomarev (2024, Theorem 1)). It is thus generally computationally infeasible to fully characterize $\mathcal{H}(m,\gamma)$ and $\mathcal{H}(\tau)$ when the relevant outcome space is infinite. This is a well-known issue with identified sets that follow from Artstein's theorem. A common way of addressing it is to discretize the relevant variables or focus on settings where they are finitely supported (Galichon and Henry (2011), Russell (2021), Ponomarev (2024)). Here, I do the same. A computationally tractable characterization of $\mathcal{H}(\tau)$ that minimizes the loss of information with infinitely supported $S$ is an interesting avenue for future research.

Henceforth, I maintain that $S(d) \in \mathcal{S}$ is a finite set with $|\mathcal{S}| = k$, either by definition or following discretization performed by the researcher. Without loss, let $S_j(d) \in \mathcal{S}_j = \{1,\dots,k_j\}$ and $k = \sum_{j=1}^{d_s} k_j$. Subtleties related to the interpretation of results under discretization are discussed in Appendix A.2. I do not require the long-term outcome support $\mathcal{Y}$ to be a finite or discrete set, but I maintain $\mathcal{Y} \subseteq [Y_L, Y_U]$ for some known finite $Y_L, Y_U$, normalized to $[0,1]$ without loss of generality. This is commonly required for informative inference under nonparametric treatment response assumptions. The support restriction may be natural for various $Y(d)$ such as binary indicators, or discrete and continuous variables that are logically bounded. For some $Y(d)$, it may be restrictive. When, $|\mathcal{S}| < \infty$, one can represent $\gamma_d$ as an element of a $k-$dimensional simplex $\Delta(k)$, and $\gamma \in \Delta(k) \times \Delta(k)$. Similarly, $m \in \mathcal{M} = \mathcal{Y}^k \times \mathcal{Y}^k$, and the modeling assumption can be represented as $\mathcal{M}^A \subseteq \mathcal{Y}^k \times \mathcal{Y}^k$. Let $\gamma_d(s)$ and $m_d(s)$ denote the $s-$th element of the corresponding vectors. This leads to the following characterization result.

**Theorem 2.** *Let Assumptions RA, EV, and MA hold. Suppose $\mathcal{S}$ is a finite set and that $\mathcal{M}^A$ is closed and convex. Then:*

$$\mathcal{H}(\tau) = \left[ \min_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}), \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}) \right] \tag{22}$$

*where:*

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s, D=d|Z), P_O(S=s, D=d)\right), \\ m_d(s)\gamma_d(s) \geq E_O[Y|S=s, D=d]P_O(S=s, D=d)\ , \\ (1-m_d(s))\,\gamma_d(s) \geq E_O[1-Y|S=s, D=d]P_O(S=s, D=d) \end{array} \right\} \tag{23}$$

By the theorem, $\mathcal{H}(\tau)$ can be equivalently represented as an interval bounded by solutions to two optimization problems where $\mathcal{H}(m,\gamma)$ represents the constraint set. The characterization follows under easily verifiable high-level conditions on $\mathcal{M}^A$. Remark 2 explains that these conditions are satisfied by the proposed and existing assumptions.

Using optimization problems to characterize identified sets has become common in partial identification analyses. Such representations usually follow directly from the convexity of the constraint set and linearity of the objective function. Theorem 2 requires a different argument since $T$ is a difference of two Riemann-Stieltjes integrals, thus bilinear and therefore separately continuous in $m$ and $\gamma$. The proof shows that $T$ is jointly continuous, and that Theorem 1 yields $\mathcal{H}(m,\gamma)$ which is compact and convex under the assumptions of the theorem. Then $\mathcal{H}(\tau)$ is a continuous image of a compact and convex set, hence a compact connected set, i.e., a closed interval.

**Remark 2.** The assumptions considered here are representable via linear equality and inequality restrictions on $m$. Therefore, the resulting $\mathcal{M}^A$ are polytopes when $|\mathcal{S}| < \infty$, and thus closed and convex. Assumption LIV states that vectors $m_d \in \mathcal{Y}^k$ have non-decreasing components $m_d(s_j, s_{-j}) \leq m_d(s_j+1, s_{-j})$ for any $j \in \{1,\ldots,d_s\}$, $d \in \{0,1\}$ and $s_{-j} \in \mathcal{S}_{-j}$; Assumption TI maintains that $m_1(s) = m_0(s)$. Moreover, whenever $m$ is identified by the data, such as under LUC, $\mathcal{M}^A$ is a singleton and hence closed and convex.

Constraints imposed by $\mathcal{H}(m,\gamma)$ are linear or bilinear in $(m,\gamma)$ under previously considered assumptions. Coupling this with the fact that $T$ is bilinear, the optimization problems in Theorem 2 represent *generalized bilinear programs* (see Al-Khayyal (1992)). While such programs are generally non-convex, modern general-purpose optimizers can solve them to provable global optimality using spatial branch-and-bound algorithms (e.g. Gurobi Optimization (2024)). Regardless, depending on the complexity of the constraint set, finding the solution may be computationally demanding. To reduce the complexity, Theorem 2 utilizes the fact that the family of sets $\{\{s\} : s \in \mathcal{S}\}$ represents a *core-determining class* (CDC henceforth) for $\mathbf{S}(d)$ when $|\mathcal{S}| < \infty$. The CDC removes redundant constraints on $\mathcal{H}(m,\gamma)$ in the optimization problems without any loss of information.

The reduction in the number of constraints due to the CDC depends on the size of $|\mathcal{S}|$, but it is sizeable even for relatively few support points. Without the CDC, there would be $2^{k-1}$ inequality conditions for each for each $\gamma_d$, one for each nontrivial proper subset in $\mathcal{C}(\mathcal{S})$. With the CDC, the number of constraints for each $\gamma_d$ is reduced to $k-1$. Table 1 depicts the magnitude of this reduction, showing the total number of constraints on $\gamma$ with and without the CDC in a single optimization problem with respect to $|\mathcal{S}|$.[12] If $S(d)$ represents percentiles, then not using the CDC results in a prohibitively complex constraint set. The number of constraints may potentially be further reduced by adapting methods in Luo and Wang (2018) and Ponomarev (2024), as $\{\{s\} : s \in \mathcal{S}\}$ is not necessarily the smallest CDC. Appendix A.3 discusses additional simplifications exploiting the structure of the identified sets and the programs that may further alleviate computational burden.

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

| | $\lvert\mathcal{S}\rvert$ | | | | |
|---|---|---|---|---|---|
| Constraint # for $\gamma$ | 2 | 5 | 10 | 20 | 100 |
| Without CDC | 4 | 32 | 1024 | 1,048,576 | $> 10^{30}$ |
| With CDC $\{\{s\} : s \in \mathcal{S}\}$ | 2 | 8 | 18 | 38 | 198 |

**Remark 3.** Optimization problems $\max / \min_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m}, \tilde{\gamma})$ become linear and simpler to solve in some cases. This happens whenever either $\mathcal{H}(m, \gamma) = \{m\} \times \mathcal{H}(\gamma)$; or $\mathcal{H}(m, \gamma) = \mathcal{H}(m) \times \{\gamma\}$ and $\mathcal{M}^A$ can be expressed using linear constraints. Assumptions that point identify $m$ independently of $\gamma$, such as Assumption LUC, yield $\mathcal{H}(m, \gamma) = \{m\} \times \mathcal{H}(\gamma)$. $\mathcal{H}(m, \gamma) = \mathcal{H}(m) \times \{\gamma\}$ occurs for Assumptions LIV and TI if the right-hand sides of constraints for each $\gamma_d$ sum to 1, such as under perfect compliance. Note that then Lemma 12 could even yield a closed-form expression for $\mathcal{H}(\tau)$ for certain modeling assumptions using simplifications in Appendix A.3.

## 3.4 Estimation

The analysis thus far has focused on identification. I now propose a consistent estimation procedure for $\mathcal{H}(\tau)$. For this, 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. Let $n := \min\{n_O, n_E\}$, and let $P_{E,n}(S \in A, D = d, Z = z) := \frac{1}{n_E}\sum_{j=1}^{n_E} \mathbb{1}\{S_j \in A, D_j = d, Z_j = z\}$ and $P_{O,n}(S \in A, D = d) := \frac{1}{n_O}\sum_{i=1}^{n_O} \mathbb{1}\{S_i \in A, D_i = d\}$ be standard empirical measures. Denote by $E_{E,n}$ and $E_{O,n}$ the

12. The number of constraints on $m$ imposed by the data given $\gamma$ is $4k$; the total number depends on the modeling assumptions.

corresponding empirical expectations. Note that their population counterparts, along with $\mathcal{M}^A$, fully characterize $\mathcal{H}(m,\gamma)$ and thus $\mathcal{H}(\tau)$. Define the empirical analog of $\mathcal{H}(m,\gamma)$:

$$\mathcal{H}_n(m,\gamma) := \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}_n^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(\max_{z\in\mathcal{Z}} P_{E,n}(S=s,D=d|Z=z), P_{O,n}(S=s,D=d)\right), \\ m_d(s)\gamma_d(s) \geq E_{O,n}[Y|S=s,D=d]P_{O,n}(S=s,D=d)\ , \\ (1-m_d(s))\,\gamma_d(s) \geq E_{O,n}[1-Y|S=s,D=d]P_{O,n}(S=s,D=d) \end{array} \right\}. \quad (24)$$

where $\mathcal{M}_n^A$ highlights that restrictions imposed by the modeling assumptions may depend on estimated population parameters. For example, this would happen with Assumption LUC which imposes that $m_d(s) = E_O[Y|S=s,D=d]$, but not with Assumptions LIV and TI since they only restrict the parameter space for $m$. $\mathcal{H}(\tau)$ can be estimated using:

$$\mathcal{H}_n(\tau) := \left[\min_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}_n(m,\gamma)} T(\tilde{m},\tilde{\gamma}), \max_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}_n(m,\gamma)} T(\tilde{m},\tilde{\gamma})\right]. \quad (25)$$

To establish consistency in the Hausdorff distance, I introduce additional notation. Let $\mathcal{H}^{ie}(m,\gamma)$ be the set of $(m,\gamma)$ satisfying inequality constraints imposed on the parameter space. This is the set of $(m,\gamma) \in \mathcal{Y}^{2k} \times [0,1]^{2k}$, which satisfy all imposed modeling assumptions that do not involve population parameters. For example, it would contain $(m,\gamma)$ satisfying Assumption LIV, if imposed. Assumption LUC would not affect $\mathcal{H}^{ie}(m,\gamma)$ since it imposes equality constraints involving population parameters $E_O[Y|D=d,S=s]$. 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)* $|\mathcal{S}|, |\mathcal{Z}| < \infty$*;*

*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* $\tilde{\beta} \in \tilde{\mathfrak{B}}$ *where* $\tilde{\mathfrak{B}}$ *is compact. The Jacobian of linear equality constraints, if imposed, has full row rank.*

*iv)* $cl(int(\mathcal{H}^{ie}(m,\gamma)) \cap \mathcal{H}(m,\gamma)) = \mathcal{H}(m,\gamma)$ *or* $\mathcal{H}(m,\gamma)$ *is a singleton.*

Assumption E *i)* is standard under random sampling, *ii)* maintains that short-term potential outcomes and the instrument $Z$ are finitely supported. Assumption E *iii)* defines the class of modeling assumptions that are compatible with the estimation procedure. If necessary, it may be further weakened to allow for continuously differentiable restrictions on $m$, but it is

sufficiently general to encompass all previously stated modeling assumptions. Assumption LIV can be represented only using linear inequality constraints on the parameter space $m_d(s_j, s_{-j}) \leq m_d(s_j+1, s_{-j})$ for $j \in \{1, \ldots, d_s\}$, and $s_{-j} \in \mathcal{S}_{-j}$ and $d \in \{0, 1\}$ so *iii)* holds directly. Assumption TI involves only linear constraints on the parameter space $m_1(s) = m_0(s)$ for $s \in \mathcal{S}$. Since each constraint restricts a different $s$, it is immediate that the constraints will be linearly independent, and the Jacobian matrix will have full row rank. Similar arguments apply to assumptions that use equality restrictions involving consistently estimable population parameters, such as Assumption LUC for which $m_d(s) = E_O[Y|D = d, S = s]$.

Condition E *iv)* is a mild condition in Shi and Shum (2015, Theorem 2.1) which leads to a consistent estimator without requiring a tuning parameter. For example, it holds when $int(\mathcal{H}(m, \gamma)) \neq \emptyset$, i.e. when components of $(m, \gamma)$ are partially identified, or when $\mathcal{H}(m, \gamma)$ is in the interior of $\mathcal{H}^{ie}(m, \gamma)$. The former is typically not restrictive whenever treatment response assumptions are maintained and there is imperfect experimental compliance. The latter is typically not restrictive when $E_O[Y|D = d, S = s]$ does not take values on the boundary of the support of $Y$ under Assumption LUC. The condition may be relaxed at the expense of introducing tuning parameters, as explained by Shi and Shum (2015, Section 2).

**Theorem 3.** *Let Assumptions RA, EV, MA, and E hold. Then as* $n \to \infty$*:*

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

The proof relies on the fact that $T$ is a continuous functional in finite-dimensional spaces which implies that it is sufficient to show that $d_H(\mathcal{H}_n(m, \gamma), \mathcal{H}(m, \gamma)) \xrightarrow{p} 0$ to ensure $d_H(\mathcal{H}_n(\tau), \mathcal{H}(\tau)) \xrightarrow{p} 0$. I do so by applying arguments of Russell (2021, Theorem 2) to verify the conditions of Shi and Shum (2015, Theorem 2.1) which yields a consistent criterion-based estimator of $\mathcal{H}(m, \gamma)$. $\mathcal{H}_n(m, \gamma)$ is numerically equivalent to the criterion-based estimator whenever $\mathcal{H}_n(m, \gamma) \neq \emptyset$. This happens with probability approaching 1 for large $n$, yielding consistency of $\mathcal{H}_n(m, \gamma)$ and thus the plug-in procedure.

**Remark 4.** $\mathcal{H}_n(m, \gamma)$ and hence $\mathcal{H}_n(\tau)$ may be empty in finite samples even when $\mathcal{H}(m, \gamma)$ and $\mathcal{H}(\tau)$ are not. The proof of Theorem 3 shows that in that case, one may consistently estimate $\mathcal{H}(m, \gamma)$ using the estimator of Shi and Shum (2015), which will always be nonempty. In turn, this will yield a nonempty estimate of $\mathcal{H}(\tau)$. However, doing so may increase the computational burden, as it involves an additional minimization of a criterion function. The plug-in procedure is more computationally parsimonious and numerically equivalent when it produces a non-empty set. Hence, researchers may prefer to first attempt plug-in estimation and resort to the criterion

approach should the plug-in yield an empty set. Appendix A.4 discusses the criterion-based estimator.

# 4 Results on the Roles of Assumptions and Data

Since only $S$ is observed in both datasets, the sole benefit of experimental data lies in providing *additional* information on $\gamma$. This section examines how this can be beneficial.

Let $\mathcal{H}^O(m,\gamma)$ be the identified set for $(m,\gamma)$ if only observational data are used. Continue denoting by $\mathcal{H}(m,\gamma)$ the identified set for $(m,\gamma)$ when both datasets are used. If $(m,\gamma)$ are consistent with both datasets, they must be consistent with just one dataset under the same assumptions. Thus, $\mathcal{H}(m,\gamma) \subseteq \mathcal{H}^O(m,\gamma)$. Usually $\mathcal{H}(m,\gamma) \subsetneq \mathcal{H}^O(m,\gamma)$ with or without modeling assumptions in the observational data because experimental data provide additional information on $\gamma$. By definition (9), the corresponding identified sets for $\tau$ are:

$$\begin{aligned} \mathcal{H}^O(\tau) &= \left\{T(m,\gamma): \ \ (m,\gamma) \in \mathcal{H}^O(m,\gamma)\right\} \\ \mathcal{H}(\tau) &= \left\{T(m,\gamma): \ \ (m,\gamma) \in \mathcal{H}(m,\gamma)\right\} \end{aligned} \tag{26}$$

recalling that $T(m,\gamma) = \int_{\mathcal{S}} m_1(s)d\gamma_1(s) - \int_{\mathcal{S}} m_0(s)d\gamma_0(s)$. By definition $\mathcal{H}(\tau) \subseteq \mathcal{H}^O(\tau)$. Similarly, let $\mathcal{H}^O(P_{Y(0),Y(1)})$ and $\mathcal{H}(P_{Y(0),Y(1)})$ be the corresponding identified sets for the distribution function $P_{Y(0),Y(1)}$ and observe that $\mathcal{H}(P_{Y(0),Y(1)}) \subseteq \mathcal{H}^O(P_{Y(0),Y(1)})$

**Central Role of Modeling Assumptions**

I first ask whether it is possible to have $\mathcal{H}(\tau) \subsetneq \mathcal{H}^O(\tau)$ if no modeling assumptions are imposed. This would be desirable as then the additional identifying power would solely be the result of random assignment in an appropriate experiment. However, this is not the case.

**Proposition 1.** *Suppose Assumptions RA and EV hold. Then:*

*i)* $\mathcal{H}^O(\tau) = \mathcal{H}(\tau)$;

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

*On their own*, the experimental data bring *no identifying power for* $\tau$ or any functional of $P_{Y(0),Y(1)}$. Modeling assumptions are *central* in the identification argument for $\tau$, mirroring the importance of identifying assumptions in conventional observational studies. They are *necessary* to benefit from the existence of the short-term experiment in terms of identification. Corollary 1 in Appendix A.1 further proves that without such assumptions: 1) $\tau$ is unindentified so $\mathcal{H}(\tau) = \mathbb{R}$;

2) $\mathcal{H}(\tau)$ is equivalent to the bounds of Manski (1990) when the support of $Y(d)$ is bounded. Modeling assumptions are thus necessary to identify at least the sign of $\tau$.

The intuition behind the result is simple. Since $S(d)$ is revealed whenever $D = d$, experimental data only provide more information on the distribution of $S(d)$ for individuals who choose $D \neq d$. However, for them, no data restrict the relationship between $Y(d)$ and $S(d)$. If this relationship is left unrestricted, then additional information on $S(d)$ does not yield more information on $Y(d)$.

**Remark 5.** The roles of datasets and modeling assumptions are a topic of ongoing discussion. A seemingly similar analysis can be found in Park and Sasaki (2024b); however, the conclusion is fundamentally different. They find that observational data alone yield worst-case bounds on the treatment effects on treated survivors (ATETS) from Vikström, Ridder, and Weidner (2018), and demonstrate that combined data may be more informative under Assumption LUC. They thus do not uncover the central role of modeling assumptions.

Athey, Chetty, and Imbens (2020, Lemma 2) show that the addition of experimental data is not sufficient to point identify $\tau$ in the absence of modeling assumptions, but note that it has identifying power. I further clarify that it may have identifying power for functionals of distributions pertaining to short-term potential outcomes $S(d)$. However, the addition of experimental data provides no identifying power for any functional of $P_{Y(0),Y(1)}$, in the absence of assumptions beyond RA and EV.

**Auxiliary Amplifying Role of Experimental Data**

Since modeling assumptions are central, experimental data have an *auxiliary role*. To make the role precise, continue to denote by $\mathcal{H}^O(\tau)$ the identified set for $\tau$ when only observational data are used without modeling assumptions, and let $\mathcal{H}^{O/A}(\tau)$ be the identified set when a modeling assumption is added. Finally, denote by $\mathcal{H}(\tau)$ the identified set from combined data under the modeling assumption. It is easy to see that by definition $\mathcal{H}(\tau) \subseteq \mathcal{H}^{O/A}(\tau) \subseteq \mathcal{H}^O(\tau)$.

By Proposition 1, more information on $\gamma$ does not result in tighter bounds on $\tau$ alone. Any assumption that only restricts $\gamma$ thus cannot provide more information on $\tau$. Therefore, any set of assumptions that has identifying power for $\tau$ must also restrict $m$, so $\mathcal{M}^A \subsetneq \mathcal{M}$. This yields the following observations.

First, modeling assumptions restricting $m$ may be informative of $\tau$ even in the absence of experimental data, since some information on $\gamma$ is available in both datasets. It is possible that $\mathcal{H}^{O/A}(\tau) \subsetneq \mathcal{H}^O(\tau)$. Second, more information on $\gamma$ may make assumptions restricting $m$ more informative. Experimental data may thus *amplify* the identifying power of such modeling assumptions so $\mathcal{H}(\tau) \subsetneq \mathcal{H}^{O/A}(\tau)$. Third, $\mathcal{H}(\tau) = \mathcal{H}^{O/A}(\tau)$ is possible. So experimental data do not *necessarily* amplify the identifying power of modeling assumptions. The following remark

illustrates these three points using Assumption LUC. Similar results can be derived for other modeling assumptions.

**Remark 6.** Proposition 2 in Appendix A.1 demonstrates that: 1) LUC provides identifying power for $\tau$ without experimental data for common data distributions, so $\mathcal{H}^{O/A}(\tau) \subsetneq \mathcal{H}^{O}(\tau)$ is possible; 2) Since LUC point identifies $\tau$ with combined data, usually $\mathcal{H}(\tau) \subsetneq \mathcal{H}^{O/A}$; 3) there exist data distributions for which LUC point identifies $\tau$ without experimental data, so $\mathcal{H}(\tau) = \mathcal{H}^{O/A}(\tau)$ is possible.

**Importance of Plausible Modeling Assumptions**

In terms of the importance of modeling assumptions, approaches that rely on data combination effectively conduct observational studies. The amplifying role of the experimental data emphasizes the importance of plausible modeling assumptions. If the assumptions fail, adding experimental data may be *detrimental*. To see this, suppose a modeling assumption fails and let $\tilde{\mathcal{H}}$ be the misspecified identified set for $\tau$ following from combined data. Similarly, let $\tilde{\mathcal{H}}^{O/A}$ be the misspecified set that follows from observational data under the same assumptions. Any value consistent with both datasets must be consistent with just one dataset, so $\tilde{\mathcal{H}} \subseteq \tilde{\mathcal{H}}^{O/A}$.

**Lemma 2.** *(Nested Misspecification) Let $\tilde{\mathcal{H}} \subseteq \tilde{\mathcal{H}}^{O/A}$ be misspecified identified sets for some parameter $\tau$. Let $d$ be the point-to-set distance defined as $d(A,t) := \inf\{\|t-a\| : a \in A\}$ for $A \subseteq \mathbb{R}$ and $t \in \mathbb{R}$. Then:*

$$d(\tilde{\mathcal{H}}^{O/A}, \tau) \leq d(\tilde{\mathcal{H}}, \tau)$$

Lemma 2 states that further reducing the size of any misspecified identified set *necessarily* produces results that are weakly farther away from the truth. Thus, adding experimental data can only move the resulting identified set farther away from the true $\tau$ when a modeling assumption fails. In that case, the researcher may only obtain results closer to the ground truth by discarding the available experimental data, and these results may be informative.

Example 5 in Appendix A.1 shows that $\tilde{\mathcal{H}}^{O/A}$ can be informative of the sign of $\tau$ and strictly, not only weakly, closer to $\tau$ than $\tilde{\mathcal{H}}$ when the modeling assumption fails. It relies on a non-pathological data-generating process and standard assumptions. It also demonstrates that adding experimental data may lead the researcher to incorrectly dismiss the true value of $\tau$. We may have $\tau \notin \tilde{\mathcal{H}}$ and $\tau \in \tilde{\mathcal{H}}^{O/A}$, but never the converse. If the modeling assumption holds, $\tau$ is in both identified sets, and adding experimental data cannot produce results farther away from the truth.

**Remark 7.** Lemma 2 is a general misspecification result. It implies that reducing the size of the identified set can never result in the set being closer to the truth. Here, the reduction may happen through the addition of data. More commonly, it is a result of layering additional assumptions.

# 5 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.[13] It was introduced in 1965 as part of the "War on Poverty" and aimed to help close the gap between disadvantaged and non-disadvantaged children at the national level.

The long-term treatment effects of the program have been studied by a large body of research, primarily relying on observational studies. One common approach is to make within-family comparisons between siblings who did and did not participate in Head Start, identifying the effects on individuals with such siblings (Currie and Thomas (1995), Garces, Thomas, and Currie (2002), Deming (2009), Bauer and Schanzenbach (2016)). Another is to leverage variation in program funding, income-related eligibility, or program rollout timing to identify the relevant local average treatment effects (Ludwig and Miller (2007), Carneiro and Ginja (2014), Bailey, Sun, and Timpe (2021)). Finally, Kline and Walters (2016) estimate the effect on adulthood earnings for compliers by monetizing the corresponding LATE of Head Start on test scores in a short-term experiment. They do so using estimates of the relationship between test scores and earnings based on follow-up of Tennessee Project STAR experiment participants in administrative data (Chetty et al. (2011)).

Despite the long history of the program and its study, this literature has yet to achieve consensus (Gibbs, Ludwig, and Miller (2011), Pages et al. (2020)). The assumptions underlying existing approaches and the generalizability of their target parameters have been the subject of extensive discussion (Ludwig and Phillips (2008), Elango et al. (2015), Gonzalez (2020), García et al. (2020), Miller, Shenhav, and Grosz (2023)). In this section, I illustrate an alternative approach to estimating the long-term average treatment effects of Head Start for eligible individuals by applying the developed method, which enables the use of assumptions that do not restrict selection into the treatment.

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

## 5.1 Data

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

HSIS was an experimental trial of Head Start mandated by the 105th US Congress for the 1998 reauthorization of the program. In the fall of 2002, a total of 4,667 children from nationally representative cohorts aged 3 and 4 were enrolled in the experiment. Across 383 randomly chosen Head Start centers, participants were randomized to a treatment group that was assigned to enroll in Head Start or a control group that was prevented from enrolling, resulting in $Z \in \{0, 1\}$. Not everyone complied with their assigned treatment, so $P_E(D \neq Z) > 0$ (Puma et al. (2010)). Since participants were followed up until the third grade, HSIS does not contain adolescence or adulthood outcomes that may be of interest. It does reveal Woodcock-Johnson III (WJ-III) cognitive assessment scores, which may be used as short-term outcomes. I follow Kline and Walters (2016) and Kamat (2024), pooling all children into a single cohort. I keep all children who had available scores, yielding the experimental sample size of $n_E = 3,540$. For compatibility with the observational data, following Griffen and Todd (2017), I create composite scores for math and reading ability by averaging the nation-level percentile scores on the corresponding components of the cognitive ability test and using them as a two-dimensional $S$ binned to a total of 400 support points.

CNLSY is a biennial longitudinal survey introduced in 1986, tracking a total of 11,545 children born to participants in the National Longitudinal Survey of Youth 1979 cohort (NLSY79), which was designed to be nationally representative, like HSIS. It reveals long-term outcomes previously considered in the literature and non-randomized Head Start participation. As corresponding measures of math and reading ability, I take the percentile scores on the Peabody Individual Achievement Math and Reading Recognition subtests, also binned to a total of 400 support points. While CNLSY directly inquires about program participation, the eligibility of non-participants must be inferred. For the analysis, I rely on participation and eligibility variables constructed by Carneiro and Ginja (2014, Section IIA).

Participation is determined by the question of whether the child has ever attended Head Start.[14] Eligibility is inferred by establishing if the child met the contemporaneous program requirements based on survey answers. Children ages three to five are eligible if their family income is below the federal poverty line, or if their family is eligible for any of the following public assistance programs: Aid to Families with Dependent Children (AFDC) or Temporary Assistance

14. 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.

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, with a sample size of $n_O = 2,535$. The individuals were eligible for the program starting from the 1980s until the early 2000s and have follow-up data up to 2020.

I consider eight long-term outcomes: grade retention, diagnosis of a learning disability, high school graduation, "idleness", criminal involvement, teenage parenthood, self-reported health status and average earnings. These individual-level outcomes were chosen based on previous studies of Head Start (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, excluding General Educational Development certification. Individuals were considered idle if they did not report any wages or being in school in their most recent year of interview. 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 indicator equal to one if it is below three out of five. Finally, all reported earnings are averaged and inflation-adjusted to 2020 dollars using the CPI index of the Bureau of Labor Statistics.

Table 2 presents descriptive statistics from the two samples. They have comparable gender compositions and ratios of white to non-white individuals. The rate of compliance with the assigned treatment is 83.8% in the experiment. While lower than in the experiment, a significant proportion of CNLSY individuals participate in the program. Compliance issues and the availability of the treatment in the observational population preclude direct application of previously used data combination methods that assume perfect compliance or where the treatment is available only in the experiment. However, the developed method can be used to produce bounds under their relevant identifying assumptions. I illustrate this in the following section by reporting bounds under assumptions following from Athey, Chetty, and Imbens (2020) and García et al. (2020).[15]

15. Athey, Chetty, and Imbens (2020) assume perfect compliance. In García et al. (2020), the intervention is available only in the experiment, which is typical for "model" early childhood intervention programs such as the

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.

## 5.2 Results

Table 3 reports estimates of the bounds under previously discussed assumptions. Worst-case bounds impose no restrictions on the temporal link functions $m$, coinciding with estimated bounds of Manski (1990) as shown by Section 4, and thus necessarily including zero. Therefore, to identify at least the sign of the effect, one must impose further assumptions. For all outcomes, any of the previously mentioned assumptions reduces the size of the estimated bounds considerably.

Assumption LIV maintains that temporal link functions are monotonic in each of the two test score components. Bound estimates pertaining to high school graduation and earnings assume a weakly increasing monotonic relationship in each potential test score. For grade repetition, learning disability diagnosis, "idleness", crime, teen pregnancy, and poor health, I impose a

Carolina Abecedarian and the Perry Preschool Projects, but not for large-scale programs such as Head Start.

Table 3: Bounds Estimates

| Outcome | $n_O$ | Modeling Assumption | | | |
|---|---|---|---|---|---|
| | | Worst-case | LIV | TI | LUC |
| Repeat Grade | 2,441 | −0.476<br>0.524 | −0.053<br>−0.012 | −0.075<br>0.050 | 0.016<br>0.016 |
| Learning Disability | 2,513 | −0.448<br>0.552 | −0.008<br>0.008 | −0.076<br>0.050 | 0.052<br>0.052 |
| HS Graduate | 2,017 | −0.525<br>0.475 | 0.024<br>0.024 | −0.079<br>0.050 | 0.034<br>0.034 |
| Idle | 2,501 | −0.474<br>0.526 | −0.042<br>−0.028 | −0.072<br>0.051 | 0.018<br>0.018 |
| Crime | 2,501 | −0.499<br>0.501 | −0.039<br>−0.013 | −0.074<br>0.048 | 0.020<br>0.020 |
| Teen Pregnancy | 2,501 | −0.460<br>0.540 | 0.004<br>0.022 | −0.072<br>0.050 | 0.043<br>0.043 |
| Poor Health | 2,501 | −0.466<br>0.534 | −0.049<br>−0.030 | −0.078<br>0.044 | 0.006<br>0.006 |
| Average earnings (in 000) | 2,382 | −107.337<br>134.766 | −0.007<br>3.183 | −19.096<br>12.036 | 12.438<br>12.438 |

*Notes*: Estimated LTE bounds, represented as $\substack{\text{Lower Bound}\\\text{Upper Bound}}$, for different long-term outcomes. Worst-case bounds impose no restrictions on $m$. Remaining bounds impose only the noted modeling assumption. All bounds use experimental data.

weakly decreasing relationship. The assumption may be particularly appealing for outcomes pertaining to education, employment, crime and earnings. Taking high school graduation as an example, it would hold if one finds it plausible that, fixing any Head Start participation, individuals with higher math or reading scores are equally or more likely to graduate from high school.

Estimates under Assumption LIV reveal the sign of the effect for all but two outcomes and indicate a nearly point-identified positive effects on high school graduation of 2.4%. This occurs despite the non-compliance due to the combined information on $\gamma$ provided by the two datasets. Indeed, since estimated lower bounds on $\gamma_d(s)$ components nearly sum up to unity, the estimated intersection bounds on the short-term potential outcome distributions almost point-identify the short-term potential outcome distributions in the data.[16] Estimates further indicate that Head Start participation reduces the probability of grade repetition by at least 1.2%, being idle by at least 2.8%, being involved in crime by at least 1.3%, and reporting poor health by at least

16. I use this fact to fix $\gamma$ in the estimation, which turns the bilinear problem into a linear one and substantially reduces the computational burden. See also Remark 3.

3%. The upper bounds on these reductions are also very 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, due to longer follow up, the dataset here reveals the earnings for a larger fraction of individuals. The sign of the effect on earnings is not identified. However, estimates do suggest that the bounds almost completely lie to the right of zero, spanning a reduction of \$7 and an increase of $\$3,183$ per year in 2020 dollars.

Next, I turn to illustrative results relying on temporal link functions restrictions that follow from previously proposed methods. The methods may not be directly applicable due to non-compliance or the availability of both treatments in CNLSY. However, as previously argued, the proposed framework may be used in conjunction with the relevant modeling assumptions, extending their applicability. Estimates under Assumption TI also lead to substantial reduction in size of the bounds, but none exclude zero. Assumption TI would hold if the model proposed by García et al. (2020) is plausible. It should be noted that HSIS does not contain all short-term outcomes the authors include in the model for the Carolina Abecederian and CARE programs, nor the medium-term experimental outcomes that they use to validate the choice of short-term outcomes. The corresponding results should thus be interpreted with caution. The final column reports estimates under Assumption LUC. Since the assumption point identifies $m$, the results are very informative and reveal the sign for every outcome. However, the effect signs are not aligned with previous findings for multiple outcomes. As highlighted 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.

# 6 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 data combination is not a replacement for tenable modeling assumptions. However, it remains appealing for the 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 assumptions and thereby may yield more informative plausible inference than observational data alone.

This paper introduces two assumptions that utilize this aspect 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. 2020. "Combining experimental and observational data to estimate treatment effects on long term outcomes." *arXiv preprint arXiv:2006.09676.*

Athey, Susan, Raj Chetty, Guido Imbens, and Hyunseung Kang. 2024. *Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index.* Technical report.

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.

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.

Chen, Xuan, Carlos A Flores, and Alfonso Flores-Lagunes. 2018. "Going beyond LATE: bounding average treatment effects of Job Corps training." *Journal of Human Resources* 53 (4): 1050–1099.

Chernozhukov, Victor, Han Hong, and Elie Tamer. 2007. "Estimation and confidence regions for parameter sets in econometric models 1." *Econometrica* 75 (5): 1243–1284.

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.

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.

Deaton, Angus S. 2009. *Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development.* Technical report. National bureau of economic research.

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.

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.

Galichon, Alfred. 2018. *Optimal transport methods in economics.* Princeton University Press.

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 Sergio Urzua. 2010. "Comparing IV with structural models: What simple IV can and cannot identify." *Journal of Econometrics* 156 (1): 27–37.

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. 2010. "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic literature* 48 (2): 399–423.

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.

Li, Shoumei, and Yukio Ogura. 1998. "Convergence of set valued sub-and supermartingales in the Kuratowski-Mosco sense." *Annals of probability,* 1384–1402.

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, and Hai Wang. 2018. "Identifying and computing the exact core-determining class." *Available at SSRN 3154285.*

Manski, Charles F. 1997. "Monotone treatment response." *Econometrica: Journal of the Econometric Society,* 1311–1334.

Manski, Charles F. 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.

Masten, Matthew A, and Alexandre Poirier. 2023. "Choosing exogeneity assumptions in potential outcome models." *The Econometrics Journal* 26 (3): 327–349.

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.

Moon, Sarah. 2024. "Partial Identification of Individual-Level Parameters Using Aggregate Data in a Nonparametric Binary Outcome Model." *arXiv preprint arXiv:2403.07236.*

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

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.*

Ponomarev, Kirill. 2024. *Selecting Inequalities for Sharp Identification in Models with Set-Valued Predictions.* http://kponomarev.github.io/files_on_website/sharp%20inequalities.pdf.

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.*

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.

Schaefer, Helmut H., and M. P. Wolff. 1999. *Topological Vector Spaces.* Springer.

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.

Treves, François. 2016. *Topological Vector Spaces, Distributions and Kernels: Pure and Applied Mathematics, Vol. 25.* Vol. 25. Elsevier.

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.

Vikström, Johan, Geert Ridder, and Martin Weidner. 2018. "Bounds on treatment effects on transitions." *Journal of Econometrics* 205 (2): 448–469.

Villani, Cédric, et al. 2009. *Optimal transport: old and new.* Vol. 338. Springer.

Willard, Stephen. 2004. *General topology.* Courier Corporation.

Yildiz, Neşe. 2012. "Consistency of plug-in estimators of upper contour and level sets." *Econometric Theory* 28 (2): 309–327.

# Appendices

## Appendix A Extensions

### A.1 Additional Results on the Roles of Data and Assumptions

This appendix collects complementary results for the discussion in Section 4.

#### A.1.1 Proposition 1 and Existing Bounds

Suppose first that no modeling assumptions are maintained.

**Corollary 1.** *Suppose Assumptions RA and EV hold. If $\mathcal{Y} = \mathbb{R}$, the identified set for $\tau$ is $\mathcal{H}(\tau) = \mathbb{R}$. If $\mathcal{Y} = [0, 1]$:*

$$\mathcal{H}(\tau) = [E_O[YD] - E_O[Y(1-D)] - P_O(D=0), E_O[YD] - E_O[Y(1-D)] + P_O(D=1)]\,. \quad (27)$$

*In both cases, $0 \in \mathcal{H}(\tau)$ and the sign of $\tau$ not identified.*

Corollary 1 shows that if $\mathcal{Y}$ is unbounded and no modeling assumptions are imposed, then $\tau$ is unidentified. If the support is bounded, data combination reproduces bounds of Manski (1990), which utilize only the observational dataset. The bounds remain sharp even when the experimental dataset is added since it brings no identifying power, *on its own*.

Athey et al. (2024, Lemmas 1 and 2) provide bounds on long-term treatment effects in a different setting where $D$ is unobserved in the observational data and experimental compliance is perfect.[17] Their bounds may be narrower than those in Corollary 1, and do not maintain explicit modeling assumptions involving the potential outcomes. However, this does not contradict the result in Proposition 1. Namely, their bounds are derived under assumptions imposed on outcome variables: 1) $Y \perp\!\!\!\perp D|S, G = E$ (statistical surrogacy - Prentice (1989)); 2) $G \perp\!\!\!\perp Y|S$ (comparability). Appendix A.1.4 explains that these assumptions on outcomes imply underlying selection assumptions.

#### A.1.2 Assumption LUC and the Role of Experimental Data

Recall that $\mathcal{H}^O(\tau)$ the identified set for $\tau$ when only observational data are used and no modeling assumptions are imposed, and let $\mathcal{H}^{O/LUC}(\tau)$ denote the identified set under Assumption LUC. Finally, let $\mathcal{H}(\tau)$ be the identified set when combined data are used under Assumption LUC.

17. More precisely, they bound $E_E[Y(1) - Y(0)]$. These bounds remain valid for $\tau$ when Assumption EV is imposed.

**Proposition 2.** *Let Assumptions EV and LUC hold.*

*i) Suppose the observed data distribution $P_O(Y, S, D)$ is such that $V_O[Y|S, D = d] > 0$ $P-a.s.$ for some $d \in \{0, 1\}$ and that $\mathcal{Y}$ is a bounded set. Then $\mathcal{H}^{O/LUC}(\tau) \subsetneq \mathcal{H}^{O}(\tau)$.*

*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 $\tau$ is point-identified, and $\mathcal{H}(\tau) = \mathcal{H}^{O/LUC}(\tau)$.*

A few observations are in order. First, the proposition shows that $\mathcal{H}^{O/LUC}(\tau) \subsetneq \mathcal{H}^{O}(\tau)$ is possible. That is, LUC may have identifying power for $\tau$ for a large class of observable distributions $P_O(Y, S, D)$ even when experimental data are not used. A sufficient condition for this is that $Y$ is bounded, and that $S$ is not a perfect predictor of $Y$ for at least some $D = d$.

Second, Athey, Chetty, and Imbens (2020) show that $\mathcal{H}(\tau)$ is a singleton under combined data and LUC. Since $\mathcal{H}^{O/LUC}(\tau)$ need not be a singleton, we usually have $\mathcal{H}(\tau) \subsetneq \mathcal{H}^{O/LUC}$. Consequently, experimental data may *amplify* the identifying power of LUC.

Third, the proposition shows that $\mathcal{H}(\tau) = \mathcal{H}^{O/LUC}(\tau)$ is possible. That is, short-term experimental data are not necessary for point identification of $\tau$ under LUC. Thus, experimental data do not *necessarily* amplify the identifying power of LUC. This intuitively happens when the short-run outcomes $S$ are not predictive of the mean long-term outcomes $Y$.[18] This condition is strong and may lack practical applicability. However, the result has important theoretical implications in clarifying the role of the experimental data.

### A.1.3 An Example of Nested Misspecification

Section 4 explains that the amplifying role of experimental data has important implications when the modeling assumption fails. Then, adding experimental data may only produce identified sets for $\tau$ that are weakly farther away from the truth. Recalling the notation, $\tilde{\mathcal{H}}$ and $\tilde{\mathcal{H}}^{O/A}$ denote misspecified identified sets for $\tau$ using combined and just observational data. The following example shows that under a standard modeling assumption and a non-pathological data-generating process, $\tilde{\mathcal{H}}^{O/A}$ can be strictly closer to $\tau$ than $\tilde{\mathcal{H}}$. Moreover, $\tilde{\mathcal{H}}^{O/A}$ is informative of the sign of $\tau$.

**Example 5.** Suppose $Y, S \in \{0, 1\}$ and that the researcher maintains Assumption LUC. Let the

18. Observe that no restrictions on $\mathcal{Y}$ are required in this case.

DGP be given by:

$$
\begin{aligned}
&E_O[Y|S=1,D=1]=0.7 && E_O[Y|S=0,D=1]=0.4\\
&E_O[Y|S=1,D=0]=0.4 && E_O[Y|S=0,D=0]=0.2\\
&E[Y(1)|S(1)=1]=0.5 && E[Y(0)|S(0)=1]=0.3\\
&E[Y(1)|S(1)=0]=0.5 && E[Y(0)|S(0)=0]=0.3\\
&P_O[S=1|D=1]=0.6 && P_O[S=1|D=0]=0.4\\
&P[S(1)=1]=0.7 && P[S(0)=1]=0.3\\
&P_O[D=1]=0.5
\end{aligned}
$$

Then $\tau = 0.2$, $\tilde{\mathcal{H}}^{O/A} = [0.15, 0.4]$ and $\tilde{\mathcal{H}} = \{0.35\}$.

### A.1.4 More on Treatment Invariance and Surrogacy

By Lemma 11 *ii)*, Assumption TI is implied by surrogacy when the experiment features perfect compliance. One may thus wish to intuitively interpret TI as stating that the treatment effect on the long-term outcome is fully mediated by the short-term outcome, an interpretation commonly used for the surrogacy assumption. However, surrogacy imposes *selection assumptions* when compliance is imperfect. Then it is immediate that by surrogacy $E_E[Y(1)|S(1)=s, D=1] = E_E[Y(0)|S(0)=s, D=0]$ for $s \in \mathcal{S}$. This is an a priori restriction on the selection mechanism of experimental individuals, because $Y(d)$ are never observed for $G = E$. On the other hand, TI is always a treatment response assumption, restricting only how $(Y(1), Y(0), S(1), S(0))$ are related to each other.

Work relying on surrogacy for identification, such as Athey et al. (2024), commonly also maintains – $G \perp\!\!\!\perp Y|S$ (comparability). Comparability and surrogacy jointly imply a selection assumption even if compliance is perfect. Note that for any $s \in \mathcal{S}$ and $d \in \{0,1\}$:

$$
\begin{aligned}
E[Y(d)|S(d)=s] &= E_O[Y(d)|S(d), D=d]P_O(D=d|S(d)=s)\\
&\quad + E_O[Y(d)|S(d), D\neq d]P_O(D\neq d|S(d)=s)\\
E[Y(d)|S(d)=s] &= E_O[Y(1)|S(1)=s, D=1]P_O(D=1|S=s)\\
&\quad + E_O[Y(0)|S(0)=s, D=0]P_O(D=0|S=s)
\end{aligned}
$$

where the first identity is by the law of iterated expectations (LIE) and the second is by Lemma 11

*vi)* and LIE. Therefore, for any $s$ and $d$ such that $P(D \neq d, S(d) = s) > 0$ by rearranging terms:

$$E_O[Y(d)|S(d) = s, D \neq d] =$$
$$= \frac{E_O[Y|S, D = d]\left(P_O(D = d|S = s) - P_O(D = d|S(d) = s)\right) + E_O[Y|S, D \neq d]P_O(D \neq d|S = s)}{P_O(D \neq d|S(d) = s)}$$

which relates $(Y(1), Y(0), S(1), S(0))$ and $D$ in the observational data, and is hence a *selection assumption*.

## A.2 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 $\lambda : \mathcal{S} \to \mathcal{S}^D := \{1, 2, \ldots, k\}$ 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 $\lambda$ to clarify the subtle differences in applications of results of Section 3.3 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 $\lambda$ 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 2.3. 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 $\lambda$ to be an identity function up with 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 8.** Consider Assumption OLIV with which states that $E[Y(d)|S(d) = s]$ is in $\mathcal{M}^A$ which contains only non-decreasing temporal link functions. Then $E[Y(d)|S^D = s]$ must also be non-decreasing for any order-preserving $\lambda$, so Assumption MA:D holds for an appropriately chosen $\lambda$. 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 3.3 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.*[19] 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 3.3 should be viewed as providing an approximation of the identified set.

## A.3 Reducing Computational Complexity of Bilinear Programming

Depending on the complexity of the constraint set, finding the solution to the generalized bilinear programs may be computationally demanding even with the utilization of CDCs. I provide additional simplifications that exploit the structure of the identified set $\mathcal{H}(m,\gamma)$ and the objective $T$, which may further alleviate computational burden.

First, minimization and maximization problems may be separable into subproblems of lower dimension, which can significantly reduce the computational burden (Nocedal and Wright (1999)). This is possible 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 $(m,\gamma)$ are separable in $d$, it is immediate that the identified set $\mathcal{H}(m,\gamma)$ also becomes rectangular. Then, letting $\mathcal{T}(m_d,\gamma_d) := \int_{\mathcal{S}} m_d(s)d\gamma_d(s)$, we have:

$$\begin{aligned}
\min_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}) &= \min_{(\tilde{m}_1,\tilde{\gamma}_1)\in\mathcal{H}(m_1,\gamma_1)} \mathcal{T}(\tilde{m}_1,\tilde{\gamma}_1) - \max_{(\tilde{m}_0,\tilde{\gamma}_0)\in\mathcal{H}(m_0,\gamma_0)} \mathcal{T}(\tilde{m}_0,\tilde{\gamma}_0) \\
\max_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}) &= \max_{(\tilde{m}_1,\tilde{\gamma}_1)\in\mathcal{H}(m_1,\gamma_1)} \mathcal{T}(\tilde{m}_1,\tilde{\gamma}_1) - \min_{(\tilde{m}_0,\tilde{\gamma}_0)\in\mathcal{H}(m_0,\gamma_0)} \mathcal{T}(\tilde{m}_0,\tilde{\gamma}_0)
\end{aligned} \tag{28}$$

where $\mathcal{H}(m_d,\gamma_d)$ collects all constraints on $(m_d,\gamma_d)$ in (23) with $m_d \in \mathcal{M}_d$. 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, for each feasible $\gamma$, there may be a known $m \in \mathcal{M}^A$ which minimizes or maximizes $T(m,\gamma)$. By appropriately fixing $m$ in the optimization procedure, the size of the parameter space that branch-and-bound algorithms will explore can be reduced. To demonstrate this, the problems can be restated as bilevel programs where the inner problems may have closed-form solutions.[20] 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 $\gamma' \in \mathcal{H}(\gamma)$. The fibers

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

20. Another example of using bilevel optimization problems for identification can be found in Moon (2024).

form a correspondence $\mathcal{H}(m|\cdot) : \mathcal{H}(\gamma) \rightrightarrows \mathcal{M}^A$. The identified set can then be written as:

$$\mathcal{H}(\tau) = \left[ \min_{\tilde{\gamma} \in \mathcal{H}(\gamma)} \min_{\tilde{m} \in \mathcal{H}(\tilde{m}|\tilde{\gamma})} T(\tilde{m}, \tilde{\gamma}), \max_{\tilde{\gamma} \in \mathcal{H}(\gamma)} \max_{\tilde{m} \in \mathcal{H}(\tilde{m}|\tilde{\gamma})} T(\tilde{m}, \tilde{\gamma}) \right]. \tag{29}$$

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

**Definition 4.** (Minimal and Maximal Selectors) Let $\mathcal{H}(m|\cdot) : \mathcal{H}(\gamma) \rightrightarrows \mathcal{M}^A$ be a correspondence defined by fibers of $\mathcal{H}(m, \gamma)$ over its projection $\mathcal{H}(\gamma)$. $L_\gamma$ is a *minimal selector with respect to* $T$ if for any $\gamma \in \mathcal{H}(\gamma)$: $T(L_\gamma, \gamma) \leq T(m, \gamma)$ for all $m \in \mathcal{H}(m|\gamma)$. $U_\gamma$ is a *maximal selector with respect to* $T$ if for any $\gamma \in \mathcal{H}(\gamma)$: $T(U_\gamma, \gamma) \geq T(m, \gamma)$ for all $m \in \mathcal{H}(m|\gamma)$.

**Corollary 2.** *Let conditions of Theorem 2 hold. If $\mathcal{H}(m|\cdot)$ has minimal and maximal selectors with respect to $T$, then:*

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

The corollary demonstrates that whenever $\mathcal{H}(m|\cdot)$ has minimal and maximal selectors with respect to $T$, using them in the optimization procedures will yield the identified set $\mathcal{H}(\tau)$. To operationalize the result, focus on the lower bound $\min_{(\tilde{m}, \tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m}, \tilde{\gamma})$. It is immediate that fixing $\tilde{m} = L_{\tilde{\gamma}}$ in the minimization procedure by definition of $L_{\tilde{\gamma}}$ yields $\min_{(\tilde{m}, \tilde{\gamma}) \in \mathcal{H}(m,\gamma), \tilde{m} = L_{\tilde{\gamma}}} T(\tilde{m}, \tilde{\gamma}) = \min_{\tilde{\gamma} \in \mathcal{H}(\gamma)} T(L_{\tilde{\gamma}}, \tilde{\gamma})$ which is equivalent to the lower bound by Corollary 2. Then, observe that $L_{\tilde{\gamma}} \in \mathcal{M}^A$. Thus, one can simply replace the constraints $m \in \mathcal{M}^A$ in (23) with $m(s) = L_\gamma(s)$ for each $s \in \mathcal{S}$ in the minimization procedure to obtain the lower bound without having to optimize over $m$ for each $\gamma$. Similarly, one can replace $m \in \mathcal{M}^A$ with $m(s) = U_\gamma(s)$ in the maximization procedure to obtain the upper bound. Lemma 12 provides minimal and maximal selectors under Assumptions LIV and TI. Moreover, whenever $m$ is identified, minimal and maximal selectors exist and coincide by definition. For example, under LUC, $L_\gamma(s) = U_\gamma(s) = (E[Y|S = s, D = 0], E[Y|S = s, D = 1])$.

The plug-in estimation procedure in Section 3.4 enables the direct application of simplifications discussed above, leading to estimators that may be less computationally burdensome. If $\mathcal{M}_n^A$ is rectangular, then so is $\mathcal{H}_n(m, \gamma)$ and the optimization becomes separable as in (28). If maximal and minimal selectors $U_{n,\gamma}$ and $L_{n,\gamma}$ with respect to $T$ exist, one can directly use them by appropriately fixing $\tilde{m}$ in the optimization procedures. Since simplifications lead to numerically equivalent results when applicable, their use will also yield consistent estimators.

## A.4 A Criterion-based Estimator

The plug-in estimator $\mathcal{H}_n(\tau)$ proposed in Section 3.4 can produce estimated identified sets that are empty in finite samples, even when the identified set is nonempty. This feature may be viewed as undesirable, and a common way to avoid it is to utilize a criterion-based estimator (e.g. Chernozhukov, Hong, and Tamer (2007), Shi and Shum (2015)).

The identified set $\mathcal{H}(m,\gamma)$ can be equivalently represented via a criterion function under the maintained assumptions. Let $\mu_d$ be a $k$-dimensional vector with components $\mu_d(s) = E_O[Y|S = s, D = d]$. Let $\eta_d$ be a $k \times (|\mathcal{Z}| + 1)$ matrix with the elements $(s,z)$ 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 $\beta = (\mu_0, \mu_1, \eta_0, \eta_1, \tilde{\beta}) \in \mathfrak{B}$ where $\tilde{\beta}$ is a vector of other population distribution features that are consistently estimable and used in the definition of $\mathcal{M}^A$.[21] By Assumption E *iii)*, $\mathcal{M}^A = \{m \in \mathcal{M} : h(m,\beta) \geq 0,\ \ g(m,\beta) = 0\}$ for some known linear functions $g$ and $h$. Then:

$$
\begin{aligned}
\mathcal{H}(m,\gamma) &= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(\max_{z\in\mathcal{Z}} P_E(S = s, D = d|Z = z), P_O(S = s, D = d)\right), \\ m_d(s)\gamma_d(s) \geq E_O[Y|S = s, D = d]P_O(S = s, D = d)\ , \\ (1 - m_d(s))\,\gamma_d(s) \geq E_O[1 - Y|S = s, D = d]P_O(S = s, D = d) \end{array} \right\} \\
&= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{Y}^{2k} \times (\Delta(k))^2 : h(m,\beta) \geq 0,\ \ g(m,\beta) = 0, \\ \forall d \in \{0,1\},\ \forall s \in \mathcal{S},\ \forall z \in \{1,\ldots,|\mathcal{Z}| + 1\}, \\ \gamma_d(s) - \eta_d(s,z) \geq 0, \\ m_d(s)\gamma_d(s) - \mu_d(s)\eta_d(s,|\mathcal{Z}| + 1) \geq 0\ , \\ (1 - m_d(s))\gamma_d(s) - (1 - \mu_d(s))\eta_d(s,|\mathcal{Z}| + 1) \geq 0 \end{array} \right\} \\
&= \left\{ \ (m,\gamma) \in \mathcal{Y}^{2k} \times (\Delta(k))^2 : \tilde{h}(m,\gamma,\beta) \geq 0,\ \ g(m,\beta) = 0 \ \right\}.
\end{aligned}
\tag{30}
$$

where $\tilde{h}(m,\gamma,\beta)$ is a vector collecting left-hand sides of all linear and bilinear inequality constraints, and linear inequality restrictions $h(m,\beta) \geq 0$.

Next, convert all inequality constraints $\tilde{h}(m,\gamma,\beta)$ to equality constraints by introducing slackness parameters $\lambda_t \in [0,1]$ for each inequality constraint, as in Shi and Shum (2015, Remark pp. 497). Denote by $\lambda$ the vector of all slackness parameters, and let $\theta = (m,\gamma,\lambda) \in \mathfrak{T}$ be a vector of dimension $d_\theta \times 1$, where $\mathfrak{T}$ implicitly imposes all parameter space constraints on $m$, $\gamma$ and $\lambda$. Write all converted equality constraints $\tilde{h}(m,\gamma,\beta) - \lambda = 0$ and existing equality constraints

21. For example, with Assumption LUC, $\tilde{\beta}$ is not necessary as it only imposes restrictions $m_d(s) = \mu_d(s)$.

$g(m,\beta) = 0$ as $\tilde{g}(\theta,\beta) = \begin{pmatrix} \tilde{h}(m,\gamma,\beta) - \lambda \\ g(m,\beta) \end{pmatrix} = 0$. A criterion function can then be:

$$Q(\theta,\beta) = \tilde{g}(\theta,\beta)'\tilde{g}(\theta,\beta) \tag{31}$$

Now, define $\Theta := \{\theta \in \mathfrak{T} : \tilde{g}(\theta,\beta) = 0\} = \{\theta \in \mathfrak{T} : Q(\theta,\beta) = 0\}$. Under the assumptions, $\Theta$ is non-empty so $\bar{Q} := \min_{\tilde{\theta}\in\mathfrak{T}} Q(\tilde{\theta},\beta) = 0$ and we can write:

$$\Theta = \arg\min_{\theta\in\mathfrak{T}} \tilde{g}(\theta,\beta)'\tilde{g}(\theta,\beta) = \{\theta \in \mathfrak{T} : \ \ Q(\theta,\beta) \leq \bar{Q}\} \tag{32}$$

Note that $\mathcal{H}(m,\gamma) = \{(m,\gamma) : \exists\lambda \text{ such that } (m,\gamma,\lambda) \in \Theta\}$. Therefore:

$$\begin{aligned} \min_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m,\gamma) &= \min_{(m,\gamma,\lambda)\in\Theta} T(m,\gamma) = \min_{(m,\gamma,\lambda)\in\mathfrak{T}} T(m,\gamma) \text{ s.t. } Q(\theta,\beta) \leq \bar{Q}. \\ \max_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m,\gamma) &= \max_{(m,\gamma,\lambda)\in\Theta} T(m,\gamma) = \max_{(m,\gamma,\lambda)\in\mathfrak{T}} T(m,\gamma) \text{ s.t. } Q(\theta,\beta) \leq \bar{Q}. \end{aligned} \tag{33}$$

Let the sample criterion function be obtained via plug-in estimators $\beta_n$ of $\beta$:

$$Q(\theta,\beta_n) = \tilde{g}(\theta,\beta_n)'\tilde{g}(\theta,\beta_n). \tag{34}$$

Denote by $\bar{Q}_n := \min_{\theta\in\mathfrak{T}} Q(\theta,\beta_n)$. Then, let the corresponding criterion estimator proposed by Shi and Shum (2015) is:

$$\Theta_n = \arg\min_{\theta\in\mathfrak{T}} \tilde{g}(\theta,\beta_n)'\tilde{g}(\theta,\beta_n) = \{\theta \in \mathfrak{T} : \ \ Q(\theta,\beta) \leq \bar{Q}_n\} \tag{35}$$

Finally, we can define the corresponding criterion estimator of $\mathcal{H}(\tau)$:

$$\begin{aligned} &\mathcal{H}_n^{crit}(\tau) := \left[\tau_n^{LB}, \tau_n^{UB}\right] \\ &\tau_n^{LB} = \min_{(m,\gamma,\lambda)\in\mathfrak{T}} T(m,\gamma) \text{ s.t. } Q(\theta,\beta_n) \leq \bar{Q}_n \\ &\tau_n^{UB} = \max_{(m,\gamma,\lambda)\in\mathfrak{T}} T(m,\gamma) \text{ s.t. } Q(\theta,\beta_n) \leq \bar{Q}_n. \end{aligned} \tag{36}$$

As shown in the proof of Theorem 3, $d_H(\mathcal{H}_n^{crit}(\tau), \mathcal{H}(\tau)) \xrightarrow{p} 0$ as $n \to \infty$. Moreover, $\mathcal{H}_n^{crit}(\tau) = \mathcal{H}_n(\tau)$ whenever $\mathcal{H}_n(\tau) \neq \emptyset$. Therefore, even if researchers wish to implement an estimator which is always non-empty, it may be desirable to attempt computationally less burdensome estimator $\mathcal{H}_n(\tau)$ and resorting to $\mathcal{H}_n^{crit}(\tau)$ only if the former yields an empty set. It should also be noted that additional structure using minimal and maximal selectors with respect to $T$ may not be directly be used with the criterion estimator as they may with the plug-in one.

However, separability of the optimization problems may still be used.

# Appendix B Proofs

This section contains the proofs of all results. It begins by summarizing notation. Section B.1 collects known supporting results. Section B.2 contains auxiliary results and their proofs. Section B.3 provides proofs of the main results.

## Preliminaries and Notation

I denote laws of random elements using subscripts when the element needs to be specified (e.g. $P_{S(d)}$ is the law of $S(d)$). Laws conditional on an event $\mathcal{E}$ are denoted by $P_{\cdot|\mathcal{E}}$ (e.g. $P_{S(d)|\mathcal{E}}$ is the conditional law of $S(d)$). If the random element is clear, I write $P(\cdot|\mathcal{E}, G = g)$ 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 $E[\cdot|\mathcal{E}, G = g] = E_g[\cdot|\mathcal{E}]$ and $V[\cdot|\mathcal{E}, G = g] = V_g[\cdot|\mathcal{E}]$. Equality of distribution of two random elements or a random element and a law is denoted by $\overset{d}{=}$ (e.g. $Y \overset{d}{=} P_Y$ and $Y \overset{d}{=} Y'$). I denote random sets with boldface letters (e.g. $\mathbf{Y}$), their capacity functionals by boldface $\mathbf{T}$ (e.g. $\mathbf{T}_\mathbf{Y}$) and containment functionals by boldface $\mathbf{C}$ (e.g. $\mathbf{C}_\mathbf{Y}$). I use $(\mathbf{Y}, Z)$ to denote the random set $\mathbf{Y} \times \{Z\}$. $\mathbb{E}(\mathbf{Y}|X)$ is used for the conditional Aumann expectation of a random set $\mathbf{Y}$ given a sigma-algebra generated by a random vector $X$. If a distribution $P_Y$ is selectionable from $\mathbf{Y}$ I write $P_Y \preccurlyeq \mathbf{Y}$. I use $\overset{d}{=}$ to denote that a random element has a law, or an equivalent distribution-determining functional (e.g. $Y \overset{d}{=} P_Y$ and $\mathbf{Y} \overset{d}{=} \mathbf{C}_\mathbf{Y}$). $A$, $B$ and $K$ represent sets. $\mathcal{K}(A)$, $\mathcal{C}(A)$, $\mathcal{O}(A)$, $\mathcal{B}(A)$ are the families of all compact, closed, open and Borel subsets of the set $A$, respectively. $co(A)$ is the closed convex hull of the set $A$. The identified sets for a generic parameter $\theta$ is $\mathcal{H}(\theta)$. The set of distribution functions of random vectors with support $\mathcal{Y}$ is $\mathcal{P}^\mathcal{Y}$. 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:

$$\tilde{Z} = \mathbb{1}[G = E]Z + \mathbb{1}[G = O](sup\mathcal{Z} + 1). \tag{37}$$

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

## B.1 Known Supporting Results

### B.1.1 Random Set Theory Preliminaries

I briefly introduce the necessary concepts, and refer the reader to Molchanov (2017) and Molchanov and Molinari (2018) for a textbook treatment of the topic. More concise overviews are available in Beresteanu, Molchanov, and Molinari (2012) and Molchanov and Molinari (2014).

Define $\mathbf{R} : \Omega \to \mathcal{C}(\mathbb{R}^d)$ to be a measurable correspondence recalling that $\mathcal{C}(\mathbb{R}^d)$ is the collection of all closed subsets of $\mathbb{R}^d$.[22] I refer to $\mathbf{R}$ as a *random (closed) set*. Define the *containment functional* $\mathbf{C}_\mathbf{R} : \mathcal{C}(\mathbb{R}^d) \to [0,1]$ of $\mathbf{R}$ as $\mathbf{C}_\mathbf{R}(B) = P(\mathbf{R} \subseteq B)$, and the *capacity functional* $\mathbf{T}_\mathbf{R} : \mathcal{K}(\mathbb{R}^d) \to [0,1]$ of $\mathbf{R}$ as $\mathbf{T}_\mathbf{R}(K) = P(\mathbf{R} \cap K \neq \emptyset)$, recalling that $\mathcal{K}(\mathbb{R}^d)$ is the collection of all compact subsets of $\mathbb{R}^d$. A *selection* of a random set $\mathbf{R}$ is a random vector $R$ defined on the same probability space such that $P(R \in \mathbf{R}) = 1$. 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 $Sel^1(\mathbf{R})$. Artstein's inequalities (Artstein (1983, Theorem 2.1), Beresteanu, Molchanov, and Molinari (2012, Theorem 2.1)) give an equivalent characterization of the set of distributions of all selections of a random set.

**Lemma 3.** *(Artstein's Inequalities) A probability distribution $\mu$ on a locally compact second countable Hausdorff space $\mathfrak{X}$ is the distribution of a selection of a random closed set $\mathbf{R}$ on the same space if and only if:*

$$\forall B \in \mathfrak{F}_{cont} : \ \mu(B) \geq \mathbf{C}_\mathbf{R}(B) \Leftrightarrow \forall K \in \mathfrak{F}_{cap} : \ \mu(K) \leq \mathbf{T}_\mathbf{R}(K) \tag{38}$$

*where $\mathfrak{F}_{cont} \in \{\mathcal{C}(\mathfrak{X}), \mathcal{O}(\mathfrak{X})\}$ and $\mathfrak{F}_{cap} \in \{\mathcal{C}(\mathfrak{X}), \mathcal{O}(\mathfrak{X}), \mathcal{K}(\mathfrak{X})\}$. If $\mathbf{R}$ is almost surely compact, then (38) is equivalent to:*

$$\forall K \in \mathcal{K}(\mathfrak{X}) : \ \mu(K) \geq \mathbf{C}_\mathbf{R}(K). \tag{39}$$

*Proof.* For proof see Molchanov and Molinari (2018, Theorem 2.13, Corollary 2.14). □

If (38) holds for a distribution function $P_R$, then I call $P_R$ *selectionable* with respect to the distribution of $\mathbf{R}$, and write $P_R \preccurlyeq \mathbf{R}$. $\mu$ is selectionable if and only if there exists a random element $R' \overset{d}{=} P_R$ and a random set $\mathbf{R}' \overset{d}{=} \mathbf{R}$ defined on the same probability space such that $P(R' \in \mathbf{R}') = 1$. Family of all distributions that satisfy (38) are called the *core* of the capacity $\mathbf{T}_\mathbf{R}$. A family of compact sets $\mathcal{K}_{CD} \subseteq \mathcal{K}(\mathfrak{X})$ is a *core-determining class* if $\forall K \in \mathcal{K}_{CD} : \ \mu(K) \leq \mathbf{T}_\mathbf{R}(K)$ implies (38). A core-determining class may reduce the number of conditions that need to be verified to consider $\mu$ selectionable.

22. $\mathbf{R}$ is measurable if for every compact set $K \in \mathcal{K}(\mathbb{R}^d)$: $\{\omega \in \Omega : \mathbf{R}(\omega) \cap K \neq \emptyset\} \in \mathcal{F}$. The codomain of the map $\mathbf{R}$ is equipped by the $\sigma-$algebra generated by the families of sets $\{B \in \mathcal{C}(\mathbb{R}^d) : B \cap K \neq \emptyset\}$ over $K \in \mathcal{K}(\mathbb{R}^d)$.

If $\mathbf{R}$ has at least one integrable selection, that is $Sel^1(\mathbf{R}) \neq \emptyset$, then $\mathbf{R}$ is an *integrable random set*. Whenever the random variable $||\mathbf{R}|| = sup\{||R|| : R \in Sel(\mathbf{R})\}$ is integrable $E[||\mathbf{R}||] < \infty$, then $\mathbf{R}$ is said to be *integrably bounded*.

**Definition 5.** (Aumann Expectation) The Aumann expectation of an integrable random set $\mathbf{R}$ is defined as:

$$\mathbb{E}(\mathbf{R}) = cl\{E[R] : R \in Sel^1(\mathbf{R})\}. \tag{40}$$

If $\mathbf{R}$ is integrably bounded, then:

$$\mathbb{E}(\mathbf{R}) = \{E[R] : R \in Sel(\mathbf{R})\}. \tag{41}$$

Note that when $\mathbf{R}$ is a finite-dimensional and integrably bounded, $\mathbb{E}(\mathbf{R})$ is a closed set, and the closure operator is not used in the definition. (Molchanov (2017, Theorem 2.1.37))

The *support function* for a convex set $A \in \mathbb{R}^{d_A}$ is defined as $h_A(u) = \sup_{a \in A} a'u$ for $u \in \mathbb{R}^{d_A}$. The convex set $A$ is uniquely determined by its its support function via intersections of all half-spaces defined by $h_A$ as:

$$A = \bigcap_{u \in \mathbb{R}^{d_A} : ||u|| = 1} \{a \in \mathbb{R}^{d_A} : a'u \leq h_A(u)\}. \tag{42}$$

If $\mathbf{R}$ is integrably bounded and if either the underlying probability space is non-atomic, or if $\mathbf{R}$ is almost surely convex, then $h_{\mathbb{E}(\mathbf{R})}(u) = E[h_{\mathbf{R}}(u)]$ for all $u \in \mathbb{R}^{d_{\mathbf{R}}}$. (Molchanov and Molinari (2018, Theorem 3.11))

Recalling that $(\Omega, \mathcal{F}, P)$ is the underlying probabilty space, let $\mathcal{F}_0 \subsetneq \mathcal{F}$ be some sub-$\sigma$-algebra.

**Definition 6.** (Conditional Aumann Expectation) Let $\mathbf{R}$ be an integrable random set. For each sub-$\sigma$-algebra $\mathcal{F}_0 \subsetneq \mathcal{F}$, the conditional Aumann expectation of $\mathbf{R}$ given $\mathcal{F}_0$ is the $\mathcal{F}_0$-measurable random set $\mathbb{E}[\mathbf{R}|\mathcal{F}_0]$ such that:

$$Sel^1(\mathbb{E}[\mathbf{R}|\mathcal{F}_0], \mathcal{F}_0) = cl\{E[R|\mathcal{F}_0] : R \in Sel^1(\mathbf{R})\} \tag{43}$$

where $Sel^1(\cdot, \mathcal{F}_0)$ denote the set of integrable selections measurable with respect to $\mathcal{F}_0$ and the closure is taken in $L^1$.

For any integrable random set $\mathbf{R}$, the conditional Aumann expectation $\mathbb{E}[\mathbf{R}|\mathcal{F}_0]$ is integrable, unique and exists. If $\mathbf{R}$ is integrably bounded, so is $\mathbb{E}[\mathbf{R}|\mathcal{F}_0]$ (Molchanov (2017, Theorem 2.1.71)). When $\mathcal{F}_0$ is countably generated, then $cl\{E[R|\mathcal{F}_0] : R \in Sel^1(\mathbf{R})\} = \{E[R|\mathcal{F}_0] : R \in Sel^1(\mathbf{R})\}$.

(Molchanov (2017, pp. 271), Li and Ogura (1998, Theorem 1)) Recall that when $\mathcal{F}_0$ is a $\sigma$-algebra generated by a random vector, it is countably generated. Therefore, for any random vector $W$, $Sel^1(\mathbb{E}[\mathbf{R}|\sigma(W)], \sigma(W))$ is a closed set.

If 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, then the probability measure is said to have not atoms over $\mathcal{F}_0$. Then, $\mathbb{E}[\mathbf{R}|\mathcal{F}_0]$ is almost surely convex and $\mathbb{E}[\mathbf{R}|\mathcal{F}_0] = \mathbb{E}[co(\mathbf{R})|\mathcal{F}_0]$ a.s. (Molchanov (2017, Theorem 2.1.77)) Then, $h_{\mathbb{E}[\mathbf{R}|\mathcal{F}_0]}(u) = h_{\mathbb{E}[co(\mathbf{R})|\mathcal{F}_0]}(u) = E[h_{co(\mathbf{R})}(u)|\mathcal{F}_0]$ a.s. for all $u \in \mathbb{R}^{d_\mathbf{R}}$. (Molchanov (2017, Theorem 2.1.72))[23] Note that this will hold for any sub-$\sigma$-algebra $\mathcal{F}_0$ by Lemma 4 when the probability space is non-atomic.

### B.1.2 Other Known Results for Reference

**Theorem 4.** *Let $E$, $F$ be metrizable and let $G$ be any topological vector space. If $E$ is a Baire space or if $E$ is barreled and $G$ is locally convex, then every separately equicontinuous family $B$ of bilinear mappings of $E \times F$ into $G$ is equicontinuous.*

*Proof.* See Schaefer and Wolff (1999, Theorem III.5.1). □

**Corollary 3.** *Let $E$, $F$ be metrizable and let $G$ be any topological vector space. If $E$ is a Baire space or if $E$ is barreled and $G$ is locally convex, then every separately continuous bilinear map of $E \times F$ into $G$ is continuous (see also Treves (2016, pp. 425)).*

*Proof.* Direct from Theorem 4. □

## B.2 Auxiliary Lemmas

**Lemma 4.** *Suppose the probability space $(\Omega, \mathcal{F}, P)$ is non-atomic and that $\mathcal{F}_0 \subseteq \mathcal{F}$ is a sub-$\sigma$-algebra. Then $P$ 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.* Pick any $A \in \mathcal{F}$ with positive measure and fix any $B \in \mathcal{F}$ such that $B \subseteq A$ and $0 < P(B) < P(A)$. $B$ exists since $(\Omega, \mathcal{F}, P)$ is non-atomic. Let $C = A \setminus B$ and observe that $A = B \cup C$ and $B \cap C = \emptyset$. Thus, $P(A) = P(B) + P(C)$, $P(C) > 0$ and $P(A|\mathcal{F}_0) = P(B|\mathcal{F}_0) + P(C|\mathcal{F}_0)$ a.s. I proceed by way of contradiction supposing that $P(B|\mathcal{F}_0) = P(A|\mathcal{F}_0)$ a.s. or $P(B|\mathcal{F}_0) = 0$ a.s. Consider $P(B|\mathcal{F}_0) = P(A|\mathcal{F}_0)$ a.s. first. Then, $P(C|\mathcal{F}_0) = 0$ which implies $P(C) = 0$, contradicting $P(C) > 0$. If $P(B|\mathcal{F}_0) = 0$ a.s., then $P(B) = 0$ a.s., contradicting $P(B) > 0$. Therefore, the set $\{\omega \in \Omega : 0 < P(B|\mathcal{F}_0)(\omega) < P(A|\mathcal{F}_0)(\omega)\}$ has positive probability, which concludes the proof. □

23. Theorem 2.1.72 states that $h_{\mathbb{E}[\mathbf{R}|\mathcal{F}_0]}(u) = E[h_\mathbf{R}(u)|\mathcal{F}_0]$ a.s. for all $u \in \mathbb{R}^{d_\mathbf{R}}$. If one wishes to use the support function to determine elements of $\mathbb{E}[\mathbf{R}|\mathcal{F}_0]$, the step $h_{\mathbb{E}[\mathbf{R}|\mathcal{F}_0]}(u) = h_{\mathbb{E}[co(\mathbf{R})|\mathcal{F}_0]}(u)$ by Theorem 2.1.77 is necessary.

**Lemma 5.** *Suppose that Assumption EV holds, and that exerimental data are unobserved. Then the identified set for the distribution function $P_{Y(d),S(d)}$ is:*

$$\mathcal{H}^O(P_{Y(d),S(d)}) = \{\delta \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}} : \delta(B) \geq P_O((Y,S) \in B, D = d)\ \forall B \in \mathcal{C}(\mathcal{Y}\times\mathcal{S})\} \tag{44}$$

*Proof.* The proof proceeds by extending arguments of Beresteanu, Molchanov, and Molinari (2012, Proposition 2.4). Define the random set for $d \in \{0,1\}$:

$$(\mathbf{Y}^O(d), \mathbf{S}^O(d)) = \begin{cases} \{(Y,S)\}, \text{ if } (D,G) = (d,O) \\ \mathcal{Y}\times\mathcal{S}, \text{ otherwise} \end{cases}. \tag{45}$$

By definition, $(\mathbf{Y}^O(d), \mathbf{S}^O(d))$ summarizes all information on $(Y(d), S(d))$ *in the observational data*. Let $\tilde{I}$ be the set of triples random elements $(E_1, E_2, E_3)$ such that $(E_1, E_2, E_3) \in \mathcal{Y}\times\mathcal{S}\times G$ and $(E_1, E_2) \perp\!\!\!\perp E_3$. Then all information in the data and assumptions can be expressed as $(Y(d), S(d), G) \in Sel((\mathbf{Y}^O(d), \mathbf{S}^O(d), G)) \cap \tilde{I}$. Note that this set is non-empty since $(\mathbf{Y}^O(d), \mathbf{S}^O(d))$ produces non-trivial values only for $G = O$.

By Lemma 3, the distribution function $P((Y(d), S(d), G)) \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}\times\{E,O\}}$ characterizes a selection in $Sel((\mathbf{Y}^O(d), \mathbf{S}^O(d), G))$ if and only if:

$$\forall B \in \mathcal{C}(\mathcal{Y}\times\mathcal{S}\times\{E,O\}): \quad P((Y(d), S(d), G) \in B) \geq P((\mathbf{Y}^O(d), \mathbf{S}^O(d), G) \subseteq B) \tag{46}$$

By Molchanov and Molinari (2018, Theorem 2.33), (46) is equivalent to:

$$\begin{aligned} &\forall B \in \mathcal{C}(\mathcal{Y}\times\mathcal{S}): \quad P((Y(d), S(d)) \in B|G) \geq P((\mathbf{Y}^O(d), \mathbf{S}^O(d)) \subseteq B|G)\ P\text{-a.s.} \\ \Leftrightarrow &\forall B \in \mathcal{C}(\mathcal{Y}\times\mathcal{S}): \quad P((Y(d), S(d)) \in B|G) \geq P_O((\mathbf{Y}^O(d), \mathbf{S}^O(d)) \subseteq B) \end{aligned} \tag{47}$$

where the second line follows since experimental data are unobserved, and hence $P(G = O) = 1$.

For $A = \mathcal{Y}\times\mathcal{S}$, $P_O((\mathbf{Y}^O(d), \mathbf{S}^O(d)) \subseteq A) = 1$.[24] For any other closed subset $B \subsetneq \mathcal{Y}\times\mathcal{S}$, the containment functional can be written as:

$$\begin{aligned} P_O((\mathbf{Y}^O(d), \mathbf{S}^O(d)) \subseteq B) &= P_O((Y(d), S(d)) \in B, D = d) \\ &= P_O((Y,S) \in B, D = d). \end{aligned}$$

where the second equality follows by definition of $Y$ and $S$. Hence, the identified set for $P_{Y(d),S(d)}$ follows by (47) and (46). Sharpness follows by construction. For any $P_{Y(d),S(d)} \in$

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

$\mathcal{H}^O(P_{Y(d),S(d)})$ there exist $(Y(d), S(d))$ that are consistent with the data and assumptions such that $(Y(d), S(d)) \overset{d}{=} P_{Y(d),S(d)}$. □

**Lemma 6.** *Suppose that Assumptions RA and EV hold. Then the identified set $\mathcal{H}(P_{Y(d),S(d)})$ for the distribution function $P_{Y(d),S(d)}$ is:*

$$\mathcal{H}(P_{Y(d),S(d)}) = \left\{ \delta \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}} : \begin{array}{l} \forall B \in \mathcal{C}(\mathcal{Y}\times\mathcal{S}): \\ \delta(B) \geq \left[ \begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S]\times \\ \max\left(ess\sup_Z P_E(S\in B_S, D=d|Z), P_O(S\in B_S, D=d)\right)+ \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y}\times B_S] P_O((Y,S)\in B, D=d) \end{array} \right] \end{array} \right\} \tag{48}$$

*Proof.* The proof proceeds by extending arguments of Lemma 5. Define the random set for $d \in \{0,1\}$:

$$(\mathbf{Y}(d), \mathbf{S}(d)) = \begin{cases} \{(Y,S)\}, \text{ if } (D,G) = (d,O) \\ \mathcal{Y}\times\{S\}, \text{ if } (D,G) = (d,E) \\ \mathcal{Y}\times\mathcal{S}, \text{ otherwise} \end{cases} . \tag{49}$$

By definition $(\mathbf{Y}(d), \mathbf{S}(d))$ summarizes all information in the observed data on $(Y(d), S(d))$. Let $\tilde{I}$ be the set of triples 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 $(E_1, E_2) \perp\!\!\!\perp E_3$. Then all information in the data and assumptions can be expressed as $(Y(d), S(d), \tilde{Z}) \in Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}))\cap\tilde{I}$. If Assumptions RA and EV hold, $Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}))\cap\tilde{I} \neq \emptyset$.

By Lemma 3, 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:

$$\forall B \in \mathcal{C}(\mathcal{Y}\times\mathcal{S}\times\tilde{\mathcal{Z}}): \quad P((Y(d), S(d), \tilde{Z}) \in B) \geq P((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}) \subseteq B) \tag{50}$$

By Molchanov and Molinari (2018, Theorem 2.33), (50) is equivalent to:

$$\forall B \in \mathcal{C}(\mathcal{Y}\times\mathcal{S}): \quad P((Y(d), S(d)) \in B|\tilde{Z}) \geq P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) \ P\text{-a.s.} \tag{51}$$

Possible forms that $B$ can take are: 1) $B = \mathcal{Y}\times\mathcal{S}$; 2) $B = \mathcal{Y}\times B_S$ for some $B_S \subsetneq \mathcal{S}$; 3) $B \neq \mathcal{Y}\times B_S$ for any $B_S \subseteq \mathcal{S}$. Now consider the containment functional $P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z})$ for each case.

For $B = \mathcal{Y} \times \mathcal{S}$, $P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) = 1$ $P$-a.s. If $B = \mathcal{Y} \times B_S$ for some $B_S \subsetneq \mathcal{S}$, then $P$-a.s.:

$$\begin{aligned} P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) &= P((Y,S) \in B, D = d|\tilde{Z}) \\ &= P(Y \in \mathcal{Y}, S \in B_S, D = d|\tilde{Z}) \\ &= P(S \in B_S, D = d|\tilde{Z}) \end{aligned}$$

where the first equality is by definition of the random set, the second is by the fact that $B = \mathcal{Y} \times B_S$, and the third by definition of $\mathcal{Y}$. Finally, if for all $B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S$:

$$P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) = \begin{cases} 0, \text{ if } \tilde{Z} \in \mathcal{Z} \text{ (i.e. } G = E) \\ P_O((Y,S) \in B, D = d), \text{ if } \tilde{Z} \notin \mathcal{Z} \text{ (i.e. } G = O) \end{cases}.$$

To see why the first case holds, define the fiber of $B$ at point $s$ as $B_Y(s) = \{y : (y,s) \in B\}$. Observe that if for all $B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S$, then for some $s$ it must be that $B_Y(s) \subsetneq \mathcal{Y}$. Therefore, whenever $G = E$ (or equivalently $\tilde{Z} \in \mathcal{Z}$), the random set $(\mathbf{Y}(d), \mathbf{S}(d)) = \mathcal{Y} \times \{S\} \not\subseteq B$. Hence, only if $G = O$ can the containment functional be positive, that is, when $\tilde{Z} \notin \mathcal{Z}$. That $P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) = P_O((Y,S) \in B, D = d)$ when $G = O$ is immediate by definitions of $Y$, $S$, $\tilde{Z}$ and the random set.

Collect the relevant cases to characterize the containment functional:

$$\begin{aligned} P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) = &\begin{cases} P(S \in B_S, D = d|\tilde{Z}), \text{ if } B = \mathcal{Y} \times B_S \text{ for some } B_S \subseteq \mathcal{S} \\ \mathbb{1}[\tilde{Z} \notin \mathcal{Z}] P_O((Y,S) \in B, D = d), \text{ otherwise} \end{cases} \\ = &\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] P(S \in B_S, D = d|\tilde{Z}) + \\ &\mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] \mathbb{1}[\tilde{Z} \notin \mathcal{Z}] P_O((Y,S) \in B, D = d) \end{aligned}$$

Hence, 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 $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ $P$-a.s.:

$$P((Y(d), S(d)) \in B|\tilde{Z}) \geq \left[ \begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] P(S \in B_S, D = d|\tilde{Z}) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] \mathbb{1}[\tilde{Z} \notin \mathcal{Z}] P_O((Y,S) \in B, D = d) \end{array} \right].$$

Finally, 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:

$$P((Y(d),S(d))\in B) \geq ess\sup_{\tilde{z}} \left[ \begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S] P(S\in B_S, D=d|\tilde{Z}) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y}\times B_S] \mathbb{1}[\tilde{Z}\notin \mathcal{Z}] P_O((Y,S)\in B, D=d) \end{array} \right].$$

$$= \left[ \begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S] ess\sup_{\tilde{Z}} P(S\in B_S, D=d|\tilde{Z}) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y}\times B_S] P_O((Y,S)\in B, D=d) \end{array} \right]$$

$$= \left[ \begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S] \times \\ \max\left(ess\sup_Z P_E(S\in B_S, D=d|Z), P_O(S\in B_S, D=d)\right) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y}\times B_S] P_O((Y,S)\in B, D=d) \end{array} \right]$$

where the first line follows by the fact that $\tilde{Z} \perp\!\!\!\perp (Y(d),S(d))$, the second by the fact that $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S]$ and $\mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y}\times B_S]$ refer to mutually exclusive deterministic events, and the third by definition of $\tilde{Z}$ and the fact that $P(G=g)>0$ for $g\in\{O,E\}$. Hence, the identified set for $P_{Y(d),S(d)}$ follows by (51) and (50). Sharpness follows by construction. For any $P_{Y(d),S(d)} \in \mathcal{H}^O(P_{Y(d),S(d)})$ there exist $(Y(d),S(d))$ that are consistent with the data and assumptions such that $(Y(d),S(d)) \overset{d}{=} P_{Y(d),S(d)}$. □

**Lemma 7.** *Let $\mathcal{H}^O(P_{Y(d)})$ and $\mathcal{H}(P_{Y(d)})$ be the sets of marginals of distributions in $\mathcal{H}^O(P_{Y(d),S(d)})$ and $\mathcal{H}(P_{Y(d),S(d)})$. Then:*

$$\mathcal{H}^O(P_{Y(d)}) = \mathcal{H}(P_{Y(d)}) = \{\delta\in\mathcal{P}^{\mathcal{Y}} : \delta(B) \geq P_O(Y\in B, D=d)\ \forall B\in\mathcal{C}(\mathcal{Y})\}. \tag{52}$$

*Proof.* For any Borel set $B\in\mathcal{B}(\mathbb{R})$, by definition of a marginal distribution function:

$$P(Y(d)\in B) = P(Y(d)\in B, S(d)\in\mathcal{S}) = P((Y(d),S(d))\in B\times\mathcal{S}) \tag{53}$$

where the last line is by equivalence of events $\{Y(d)\in B, S(d)\in\mathcal{S}\}$ and $\{(Y(d),S(d))\in B\times\mathcal{S}\}$. Lemma 5 yields the identified set for joint distributions $P_{Y(d),S(d)}$ using only observational data:

$$\mathcal{H}^O(P_{Y(d),S(d)}) = \{\delta\in\mathcal{P}^{\mathcal{Y}\times\mathcal{S}} : \delta(B) \geq P_O((Y,S)\in B, D=d)\ \forall B\in\mathcal{C}(\mathcal{Y}\times\mathcal{S})\}. \tag{54}$$

Hence the identified set for marginals $P_{Y(d)}$ using only observational data is:

$$\begin{aligned} \mathcal{H}^O(P_{Y(d)}) &= \{P_{Y(d)}\in\mathcal{P}^{\mathcal{Y}} : \exists\delta\in\mathcal{H}^O(P_{Y(d),S(d)}) \text{ s.t } P(Y(d)\in B) = \delta(B\times\mathcal{S})\ \forall B\in\mathcal{B}(\mathbb{R})\} \\ &= \{P_{Y(d)}\in\mathcal{P}^{\mathcal{Y}} : P(Y(d)\in B) \geq P_O((Y,S)\in B\times\mathcal{S}, D=d)\ \forall B\in\mathcal{C}(\mathcal{Y})\} \\ &= \{P_{Y(d)}\in\mathcal{P}^{\mathcal{Y}} : P(Y(d)\in B) \geq P_O(Y\in B, D=d)\ \forall B\in\mathcal{C}(\mathcal{Y})\} \end{aligned}$$

where the first line is by definition of a marginal distribution, second is by Lemma 5 and the third is by (53).

Lemma 6 yields the identified set for joint distributions $P_{Y(d),S(d)}$ using combined data:

$$\mathcal{H}(P_{Y(d),S(d)}) = \left\{ \delta \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}} : \begin{array}{l} \forall B \in \mathcal{C}(\mathcal{Y}\times\mathcal{S}): \\ \delta(B) \geq \left[ \begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S]\times \\ \max\left(ess\sup_Z P_E(S\in B_S, D=d|Z), P_O(S\in B_S, D=d)\right)+ \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y}\times B_S] P_O((Y,S)\in B, D=d) \end{array} \right] \end{array} \right\} \tag{55}$$

Observe that the marginals are fully defined by Borel sets of the form $B\times\mathcal{S}$ with $B \subsetneq \mathcal{Y}$, which means that for all sets of interest $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S] = 0$ in the expression above. Thus, the identified set for marginals $P_{Y(d)}$ using combined data is:

$$\begin{aligned} \mathcal{H}(P_{Y(d)}) &= \{P_{Y(d)} \in \mathcal{P}^{\mathcal{Y}} : \exists \delta \in \mathcal{H}(P_{Y(d),S(d)}) \text{ s.t } P(Y(d)\in B) = \delta(B\times\mathcal{S})\ \forall B \in \mathcal{B}(\mathbb{R})\} \\ &= \{P_{Y(d)} \in \mathcal{P}^{\mathcal{Y}} : P(Y(d)\in B) \geq P_O((Y,S)\in B\times\mathcal{S}, D=d)\ \forall B \in \mathcal{C}(\mathcal{Y})\} \\ &= \{P_{Y(d)} \in \mathcal{P}^{\mathcal{Y}} : P(Y(d)\in B) \geq P_O(Y\in B, D=d)\ \forall B \in \mathcal{C}(\mathcal{Y})\} \end{aligned}$$

where the first line is by definition of a marginal distribution, second is by Lemma 6 and the fact that $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S] = 0$, and the third is by (53). It is then immediate that $\mathcal{H}(P_{Y(d)}) = \mathcal{H}^O(P_{Y(d)})$ □

**Remark 9.** The formulation of the identified sets $\mathcal{H}^O(P_{Y(d)})$ and $\mathcal{H}(P_{Y(d)})$ coincides by application of (38) to the random set:

$$\mathbf{Y}(d) = \begin{cases} \{Y\}, \text{ if } (D,G) = (d,O) \\ \mathcal{Y}, \text{ otherwise} \end{cases}.$$

**Lemma 8.** *Let $\mathcal{H}^O(P_{S(d)})$ and $\mathcal{H}(P_{S(d)})$ be the sets of marginals of distributions in $\mathcal{H}^O(P_{Y(d),S(d)})$ and $\mathcal{H}(P_{Y(d),S(d)})$. Then:*

$$\mathcal{H}^O(P_{S(d)}) = \{\delta \in \mathcal{P}^{\mathcal{S}} : \delta(B) \geq P_O(S\in B, D=d)\ \forall B \in \mathcal{C}(\mathcal{S})\} \tag{56}$$

$$\mathcal{H}(P_{S(d)}) = \left\{ \delta \in \mathcal{P}^{\mathcal{S}} : \begin{array}{l} \forall B \in \mathcal{C}(\mathcal{S}): \\ \delta(B) \geq \max\left(ess\sup_Z P_E(S\in B, D=d|Z), P_O(S\in B, D=d)\right) \end{array} \right\} \tag{57}$$

*Let $\mathcal{H}^E(P_{S(d)})$ be the identified set for $P_E(S(d))$ obtained using only experimental data. Then:*

$$\mathcal{H}^E(P_{S(d)}) = \{\delta \in \mathcal{P}^{\mathcal{S}} : ess\sup_Z P_E(S \in B, D = d|Z)\ \forall B \in \mathcal{C}(\mathcal{S})\}. \tag{58}$$

*Proof.* For any Borel set $B \in \mathcal{B}(\mathbb{R})$, by definition of a marginal distribution:

$$P(S(d) \in B) = P(S(d) \in \mathcal{Y}, S(d) \in B) = P((Y(d), S(d)) \in \mathcal{Y} \times B) \tag{59}$$

where the last line is by equivalence of events $\{Y(d) \in \mathcal{Y}, S(d) \in B\}$ and $\{(Y(d), S(d)) \in \mathcal{Y} \times B\}$. Lemma 5 yields the identified set for joint distributions $P_{Y(d),S(d)}$ using only observational data:

$$\mathcal{H}^O(P_{Y(d),S(d)}) = \{\delta \in \mathcal{P} : \delta(B) \geq P_O((Y, S) \in B, D = d)\ \forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})\}. \tag{60}$$

Hence the identified set for marginals $P(S(d))$ using only observational data is:

$$\begin{aligned}
\mathcal{H}^O(P_{S(d)}) &= \{P(S(d)) \in \mathcal{P}^{\mathcal{S}} : \exists \delta \in \mathcal{H}^O(P_{Y(d),S(d)}) \text{ s.t } P(S(d) \in B) = \delta(\mathcal{Y} \times B)\ \forall B \in \mathcal{B}(\mathbb{R})\} \\
&= \{P(S(d)) \in \mathcal{P}^{\mathcal{S}} : P(S(d) \in B) \geq P_O((Y, S) \in \mathcal{Y} \times B, D = d)\ \forall B \in \mathcal{C}(\mathcal{S})\} \\
&= \{P(S(d)) \in \mathcal{P}^{\mathcal{S}} : P(S(d) \in B) \geq P_O(S \in B, D = d)\ \forall B \in \mathcal{C}(\mathcal{S})\}
\end{aligned}$$

where the first line is by definition of a marginal distribution, second is by Lemma 5 and the third is by (59).

Lemma 6 yield the identified set for joint distributions $P_{Y(d),S(d)}$ using combined data:

$$\mathcal{H}(P_{Y(d),S(d)}) = \left\{ \delta \in \mathcal{P} : \begin{array}{l} \forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S}) : \\ \delta(B) \geq \left[ \begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] \times \\ \max\left(ess\sup_Z P_E(S \in B_S, D = d|Z), P_O(S \in B_S, D = d)\right) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] P_O((Y, S) \in B, D = d) \end{array} \right] \end{array} \right\} \tag{61}$$

Observe that marginals are defined by Borel sets of the form $\mathcal{Y} \times B$, which means that for all sets of interest $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] = 1$ in the expression above. Thus, the identified set

for marginals $P(S(d))$ using combined data is:

$$\begin{aligned}\mathcal{H}(P_{S(d)}) &= \{P(S(d)) \in \mathcal{P}^{\mathcal{S}} : \exists \delta \in \mathcal{H}(P_{Y(d),S(d)}) \text{ s.t } P(S(d) \in B) = \delta(\mathcal{Y} \times B)\ \forall B \in \mathcal{B}(\mathbb{R})\} \\ &= \{\delta \in \mathcal{P}^{\mathcal{S}} : \delta(B) \geq P_O(S \in B, D = d)\ \forall B \in \mathcal{C}(\mathcal{S})\} \\ &= \left\{ \delta \in \mathcal{P}^{\mathcal{S}} : \begin{array}{l} \forall B \in \mathcal{C}(\mathcal{S}) : \\ \delta(B) \geq \max\left(ess\sup_Z P_E(S \in B_S, D = d|Z), P_O(S \in B_S, D = d)\right) \end{array} \right\}.\end{aligned}$$

where the first line is by definition of a marginal distribution, and the second is by Lemma 6 and the fact that $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] = 1$.

For $\mathcal{H}^E(P_{S(d)})$ a simplified version of the argument for Lemma 6 applies, and is therefore omitted. □

**Remark 10.** The formulation of the identified sets $\mathcal{H}^O(P_{S(d)})$ and $\mathcal{H}(P_{S(d)})$ coincides by application of (38) to the random set:

$$(\mathbf{S}(d), \tilde{Z}) = \begin{cases} \{S\}, \text{ if } (D, G) \in \{(d, E), (d, O)\} \\ \mathcal{S}, \text{ otherwise} \end{cases}$$

and finding the set of selections $Sel(\mathbf{S}(d), \tilde{Z}) \cap I_S$ where $I_S$ is the set of random elements $(E_1, E_2)$ such that $E_1 \perp\!\!\!\perp E_2$.

**Lemma 9.** *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 $P(E_O[Y|S, D = d] \geq \sup \mathcal{Y}) > 0$. Then by definition of $\mathcal{Y}$, $P(E_O[Y|S, D = d] = \sup \mathcal{Y}) > 0$, so there exists a Borel subset $B \subseteq \mathcal{B}(\mathcal{S})$ 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 $P(Y = \sup \mathcal{Y}|S \in$

$B, D = d) = 1$. Suppose not, so that $P(Y = \sup \mathcal{Y} | S \in B, D = d) < 1$, then:

$$\begin{aligned} E_O[Y|S \in B, D = d] =& E_O[Y|Y = \sup \mathcal{Y}, S \in B, D = d] P(Y = \sup \mathcal{Y}|S \in B, D = d) + \\ & E_O[Y|Y \neq \sup \mathcal{Y}, S \in B, D = d] P_O(Y \neq \sup \mathcal{Y}|S \in B, D = d) \\ =& E_O[Y|Y = \sup \mathcal{Y}, S \in B, D = d] P_O(Y = \sup \mathcal{Y}|S \in B, D = d) + \\ & E_O[Y|Y < \sup \mathcal{Y}, S \in B, D = d] P_O(Y < \sup \mathcal{Y}|S \in B, D = d) \\ =& \sup \mathcal{Y} P_O(Y = \sup \mathcal{Y}|S \in B, D = d) + \\ & E_O[Y|Y < \sup \mathcal{Y}, S \in B, D = d] P_O(Y < \sup \mathcal{Y}|S \in B, D = d) \\ <& \mathcal{Y} \end{aligned} \tag{62}$$

where the first equality is by LIE, second is by definition of $\mathcal{Y}$, third by $E_O[Y|S \in B, D = d] = \sup \mathcal{Y}$, and the fourth by $E_O[Y|Y < \sup \mathcal{Y}, S \in B, D = d] < \sup \mathcal{Y}$ and $P(Y = \mathcal{Y}|S \in B, D = d) < 1$. By assumption, $E_O[Y|S \in B, D = d] = \sup \mathcal{Y}$. Then (62) yields a contradiction, showing that $P(Y = \sup \mathcal{Y}|S \in B, D = d) = 1$. But then $V_O[Y|S \in B, D = d] = 0$ and $P_O(S \in B|D = d) > 0$, so $P(V_O[Y|S \in B, D = d] = 0) > 0$ which contradicts $V_O[Y|S \in B, D = d] > 0$ $P-$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.

□

**Lemma 10.** *For any $\gamma_d$ that is a distribution of a selection in $Sel((\mathbf{S}(d), \tilde{Z})) \cap I$, there exists a $\gamma_d-$integrable function $\pi_{\gamma_d}$ such that for any measurable set $B \in \mathcal{B}(\mathcal{S})$:*

$$P_O(S \in B, D = d) = \int_B \pi_{\gamma_d} d\gamma_d. \tag{63}$$

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

$$P_O(D = d|\varsigma'_d) = \pi_{\gamma'_d}(\varsigma'_d) \ a.s. \tag{64}$$

*Proof.* Fix any $\gamma_d$ such that $\exists \varsigma_d \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ and $\gamma_d \overset{d}{=} \varsigma_d$. 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 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 $\gamma_d$. Then, by the Radon-Nikodym theorem there exists a measurable function $\pi_{\gamma_d}$ 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 $\gamma'_d$ that is a distribution of a selection in $Sel((\mathbf{S}(d), \tilde{Z})) \cap I$, 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 $\pi_{\gamma'_d}(s) = P_O(D = d|\varsigma_d = s)$, $\gamma'_d-$a.e. $s \in \mathcal{S}$, which concludes the proof. □

**Lemma 11.** *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* $E_O[Y|S = s] = E_g[Y(d)|S(d) = s]$ *for all* $s \in \mathcal{S}$ *and* $g \in \{O, E\}$.

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

$$E_E[Y|S] = E_E[Y|S, D = d] = E_E[Y(d)|S(d), D = d] = E_E[Y(d)|S(d)] \tag{65}$$

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:

$$\begin{aligned} E_O[Y|S=s] &= E_E[Y|S=s] \\ &= E_E[Y(1)|S(1)=s, D=1]P_E(D=1|S=s) \\ &+ E_E[Y(0)|S(0)=s, D=0]P_E(D=0|S=s) \\ &= E_E[Y(1)|S(1)=s]P_E(D=1|S=s) + E_E[Y(0)|S(0)=s]P_E(D=0|S=s) \end{aligned} \tag{66}$$

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.

□

## B.3 Main Results

**Theorem 1.** *Let Assumptions RA, EV and MA hold. If* $\mathbf{Y}(d)$ *is integrably bounded, the identified set for* $(m, \gamma)$ *is:*

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \forall B \in \mathcal{C}(\mathcal{S}), \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D=d|Z), P_O(S \in B, D=d)\right), \\ \forall u \in \{-1,1\}\text{:}\ u m_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1-\pi_{\gamma_d}(s))\ \gamma_d\text{−}a.e. \end{array} \right\} \tag{15}$$

*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* $\forall B \in \mathfrak{C}$.

*Proof of Theorem 1.* The proof proceeds through a series of steps:

1. Restrictions on $m_d$ given a selection $\varsigma_d \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ are equivalently stated using the conditional Aumann expectation;

2. The restrictions on $m_d$ given $\varsigma_d$ are restated using the support function of the conditional Aumann expectation via the convexification property on non-atomic probability spaces;

3. Restrictions on $\gamma_d$ given a selection $\varsigma_d \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ are stated using Artstein's theorem;

4. Restrictions on $(m, \gamma)$ are shown to be invariant to the selection $\varsigma_d$.

Steps 1 and 2 remove the need to search over selections $\upsilon_d \in Sel^1(\mathbf{Y}(d))$. Steps 3 allows the removal of search over selections $\varsigma_d \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$. This is formalized in Step 4.

**Step 1:** Reformulating restrictions on $m_d$ given $\varsigma_d$ as a conditional Aumann expectation.

Fix an arbitrary $d \in \{0,1\}$ and $\varsigma_d$ such that $(\varsigma_d, \tilde{Z}) \in Sel(\mathbf{S}(d)) \times \{\tilde{Z}\} \cap I$. Let $\sigma(\varsigma_d | G = O)$ be the sub-$\sigma$-algebra generated by $\varsigma_d$ 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. Since $\mathbf{Y}(d)$ is integrably bounded, so is $\mathbb{E}_O[\mathbf{Y}(d)|S(d)]$ (Molchanov (2017, Theorem 2.1.71)). $\varsigma_d$ is a measurable selection, hence a random vector. Therefore, the conditioning sub-$\sigma$-algebra $\sigma(\varsigma_d|G = O)$ of $\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]$ is generated by a random vector and is thus countably generated. Then since $\mathbf{Y}(d)$ is integrably bounded and defined on $\mathbb{R}$, $\{E_O[\upsilon_d|\varsigma_d] : \upsilon_d \in Sel^1(\mathbf{Y}(d))\}$ is a closed set, so $\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d] = \{E_O[\upsilon_d|\varsigma_d] : \upsilon_d \in Sel^1(\mathbf{Y}(d))\}$ (Li and Ogura (1998, Theorem 1), Molchanov (2017, Theorem Section 2.1.6)).

Since $\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d] = \{E_O[\upsilon_d|\varsigma_d] : \upsilon_d \in Sel^1(\mathbf{Y}(d))\}$, it is then immediate that:

$$\exists \upsilon_d \in Sel^1(\mathbf{Y}(d)) : \quad m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d] \text{ a.s.} \Leftrightarrow m_d(\varsigma_d) \in \mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d] \text{ a.s.} \tag{67}$$

Therefore:

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\}, \ \exists(\varsigma_d, \tilde{Z}) \in Sel(\mathbf{S}(d), \tilde{Z})) \cap I, \\ \exists \upsilon_d \in Sel^1(\mathbf{Y}(d)), \ \gamma_d \overset{d}{=} \varsigma_d \ m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d] \text{ a.s.} \end{array} \right\} \tag{68}$$

$$= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\}, \ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \gamma_d \overset{d}{=} \varsigma_d, \ m_d(\varsigma_d) \in Sel^1(\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]) \end{array} \right\} \tag{69}$$

$$= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\}, \ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \gamma_d \overset{d}{=} \varsigma_d, \ m_d(\varsigma_d) \in Sel(\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]) \end{array} \right\} \tag{70}$$

where the first line is by Lemma 1, the second is by (67) and the third follows since $\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]$ is integrably bounded.

**Step 2:** Representation of restrictions on $m_d$ given $\varsigma_d$ using the support function.

By assumption, the probability space is non-atomic. By Lemma 4, $P$ has no atoms over $\sigma(\varsigma_d|G = O)$ for any measurable selection $\varsigma_d$. 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 $h_{\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]}(u) = \mathbb{E}_O[co(\mathbf{Y}(d))|\varsigma_d] = E_O[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d]$ a.s. for all $u \in \mathbb{R}$.

Fix an arbitrary $\varsigma_d$ such that $(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ and $\varsigma_d \overset{d}{=} \gamma_d$. Then:

$$\begin{aligned} & m_d(\varsigma_d) \in \mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d] \text{ a.s.} \\ \Leftrightarrow & \forall u \in \{-1,1\}: \ \ um_d(\varsigma_d) \leq h_{\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]}(u) \text{ a.s.} \\ \Leftrightarrow & \forall u \in \{-1,1\}: \ \ um_d(\varsigma_d) \leq E_O[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d] \text{ a.s.} \end{aligned} \tag{71}$$

where the second line is by Rockafellar (1970, Theorem 13.1) and almost sure convexity of $\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]$, and the third 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:

$$\begin{aligned} E_O[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d] &= E_O[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d, D=d]P_O(D=d|\varsigma_d) \\ &\quad + E_O[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d, D\neq d]P_O(D\neq d|\varsigma_d) \\ &= uE_O[Y|\varsigma_d, D=d]P_O(D=d|\varsigma_d) + h_{co(\mathcal{Y})}(u)P_O(D\neq d|\varsigma_d) \\ &= uE_O[Y|S, D=d]P_O(D=d|\varsigma_d) + h_{co(\mathcal{Y})}(u)P_O(D\neq d|\varsigma_d) \\ &= u\mu_d(\varsigma_d)P_O(D=d|\varsigma_d) + h_{co(\mathcal{Y})}(u)P_O(D\neq d|\varsigma_d) \\ &= u\mu_d(\varsigma_d)\pi_{\gamma_d}(\varsigma_d) + h_{co(\mathcal{Y})}(u)(1-\pi_{\gamma_d}(\varsigma_d)) \end{aligned} \tag{72}$$

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 $\mu_d$ and $P_O(\varsigma_d = S|D = d) = 1$. The final equality is by Lemma 10. Then observe that:

$$\begin{aligned} & \forall u \in \{-1,1\}: \ \ um_d(\varsigma_d) \leq E_O[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d] \text{ a.s.} \\ \Leftrightarrow & \forall u \in \{-1,1\}: \ \ um_d(\varsigma_d) \leq u\mu_d(\varsigma_d)\pi_{\gamma_d}(\varsigma_d) + h_{co(\mathcal{Y})}(u)(1-\pi_{\gamma_d}(\varsigma_d)) \text{ a.s.} \\ \Leftrightarrow & \forall u \in \{-1,1\}: \ \ um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1-\pi_{\gamma_d}(s)) \ \gamma_d\text{–a.e.} \end{aligned} \tag{73}$$

where the second line follows by (72) and the third by $\varsigma_d \overset{d}{=} \gamma_d$. Therefore:

$$\begin{aligned} & \mathcal{H}(m,\gamma) \\ &= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\}, \ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \ \gamma_d \overset{d}{=} \varsigma_d, \\ \forall u \in \{-1,1\}: \ \ um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1-\pi_{\gamma_d}(s)), \ \gamma_d\text{–a.e.} \end{array} \right\} \end{aligned} \tag{74}$$

**Step 3:** Representation of restriction on $\gamma_d$ given $\varsigma_d$ using Artstein's theorem.

Note that for any $(m,\gamma) \in \mathcal{H}(m,\gamma)$, there exists $(\varsigma_d, \tilde{Z}) \in Sel(\mathbf{S}(d), \tilde{Z})) \cap I$ such that $\gamma_d \overset{d}{=} \varsigma_d$. I follow similar steps to those in the proof of Lemma 6 to characterize restrictions imposed on $\gamma_d$ by this condition.

By Lemma 3, a distribution function characterizes a selection in $Sel((\mathbf{S}(d), \tilde{Z}))$ if and only if:

$$\forall B \in \mathcal{C}(\mathcal{S} \times \tilde{\mathcal{Z}}): \quad P((S(d), \tilde{Z}) \in B) \geq P((\mathbf{S}(d), \tilde{Z}) \subseteq B) \tag{75}$$

$$\Leftrightarrow \forall B \in \mathcal{C}(\mathcal{S}): \quad P(S(d) \in B | \tilde{Z}) \geq P(\mathbf{S}(d) \subseteq B | \tilde{Z}) \text{ a.s.} \tag{76}$$

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 $B \subseteq \mathcal{S}$, then $P(\mathbf{S}(d) \subseteq B|\tilde{Z}) = P(S \subseteq B, D = d|\tilde{Z})$. 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:

$$\forall B \in \mathcal{C}(\mathcal{S}): \quad P(\varsigma_d \in B|\tilde{Z}) \geq P(S \subseteq B, D = d|\tilde{Z}) \text{ a.s.} \tag{77}$$

Since $(\varsigma_d, \tilde{Z}) \in I$, (77) is equivalent to:

$$\forall B \in \mathcal{C}(\mathcal{S}): \quad P(\varsigma_d \in B) \geq ess\sup_{\tilde{Z}} P(S \subseteq B, D = d|\tilde{Z}) \tag{78}$$

$$= \max\left(ess\sup_{Z} P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right) \tag{79}$$

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

Therefore, write:

$$\begin{aligned} &\exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I \text{ s.t. } \gamma_d \overset{d}{=} \varsigma_d \\ &\Leftrightarrow \forall B \in \mathcal{C}(\mathcal{S}): \quad P(\varsigma_d \in B) \geq \max\left(ess\sup_{Z} P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right) \end{aligned} \tag{80}$$

By definition, if $\mathfrak{C}$ is a core determining class, (80) is equivalent to:

$$\begin{aligned} &\exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I \text{ s.t. } \gamma_d \overset{d}{=} \varsigma_d \\ &\Leftrightarrow \forall B \in \mathfrak{C}: \quad P(\varsigma_d \in B) \geq \max\left(ess\sup_{Z} P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right) \end{aligned} \tag{81}$$

Recall that for any $(m, \gamma) \in \mathcal{H}(m, \gamma)$, there exists $(\varsigma_d, \tilde{Z}) \in Sel(\mathbf{S}(d), \tilde{Z})) \cap I$ such that $\gamma_d \overset{d}{=} \varsigma_d$. Then each such $\gamma_d$ must satisfy the conditions (80). Hence, by the characterization of $\mathcal{H}(m, \gamma)$

in (74) it follows that:

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I,\ \forall B \in \mathcal{C}(\mathcal{S}), \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right),\ \gamma_d \overset{d}{=} \varsigma_d, \\ \forall u \in \{-1,1\}:\ \ um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s)),\ \gamma_d\text{−a.e.} \end{array} \right\} \tag{82}$$

**Step 4:** Removing search over selections $\varsigma_d$.

It remains to show that $\mathcal{H}(m,\gamma) = \mathcal{H}^I$ where:

$$\mathcal{H}^I = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\}, \forall B \in \mathcal{C}(\mathcal{S}), \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ \forall u \in \{-1,1\}:\ \ um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s)),\ \gamma_d\text{−a.e.} \end{array} \right\} \tag{83}$$

First, pick $(m,\gamma) \in \mathcal{H}(m,\gamma)$. Then, since the conditions imposed on elements of $\mathcal{H}^I$ is a strict subset of those imposed on elements of $\mathcal{H}(m,\gamma)$, it must be that $(m,\gamma) \in \mathcal{H}^I$. Conversely, pick $(m,\gamma) \in \mathcal{H}^I$. By (80) and Lemma 3, for every $d \in \{0,1\}$ there exists $\varsigma_d$ such that $(\varsigma_d, \tilde{Z}) \in Sel(\mathbf{S}(d), \tilde{Z})) \cap I$, $\gamma_d \overset{d}{=} \varsigma_d$. Therefore $(m,\gamma) \in \mathcal{H}(m,\gamma)$ and thus $\mathcal{H}^I = \mathcal{H}(m,\gamma)$.

If $\mathfrak{C}$ is a core determining class, then by similar arguments and (81) it follows that:

$$\mathcal{H}^I = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\}, \forall B \in \mathfrak{C}, \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ \forall u \in \{-1,1\}:\ \ um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s)),\ \gamma_d\text{−a.e.} \end{array} \right\} \tag{84}$$

□

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

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \ \exists S(d) \in Sel(\mathbf{S}(d)) \cap \bar{I}, \\ \exists Y(d) \in Sel^1(\mathbf{Y}(d)),\ \ \gamma_d \overset{d}{=} S(d),\ \ m_d(S(d)) = E_O[Y(d)|S(d)]\ a.s. \end{array} \right\}. \tag{17}$$

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

*Proof of Lemma 1.* Let $I$ be the set of random elements $(E_1, E_2) \in \mathcal{S} \times \tilde{\mathcal{Z}}$ such that $E_1 \perp\!\!\!\perp E_2$. Recalling the definition of $\tilde{Z}$ in (37), note that $S(d) \in Sel(\mathbf{S}(d)) \cap \bar{I}$ can be equivalently stated as $(S(d), \tilde{Z}) \in Sel(\mathbf{S}(d), \tilde{Z}) \cap I$.

The proof then proceeds through a series of steps:

1. Find the set of $(m_d, \gamma_d)$ which are consistent with the data, Assumptions RA and EV, and $E[|Y(d)|] < \infty$ in terms of measurable selections $(\upsilon_d, \varsigma_d)$ of $(\mathbf{Y}(d), \mathbf{S}(d))$;

2. Equivalently characterize the set, removing redundant restrictions;

3. Find the set of corresponding $(m, \gamma)$ consistent with the data, Assumptions RA and EV, and $E[|Y(d)|] < \infty$;

4. Collect all $(m, \gamma)$ that satisfy Assumption MA to obtain $\mathcal{H}(m, \gamma)$.

**Step 1:** Restrictions on $(m_d, \gamma_d)$ without the modeling assumption and integrability.
I use the random set:

$$(\mathbf{Y}(d), \mathbf{S}(d)) = \begin{cases} \{(Y, S)\}, \text{ if } (D, G) = (d, O) \\ \mathcal{Y} \times \{S\}, \text{ if } (D, G) = (d, E) \\ \mathcal{Y} \times \mathcal{S}, \text{ otherwise} \end{cases} \tag{85}$$

which summarizes all information on $(Y(d), S(d))$ contained in the data, by definition. Recall from the proof of Lemma 6 that all restrictions imposed by data and Assumptions RA and EV on $(Y(d), S(d))$ can be expressed as $(Y(d), S(d), \tilde{Z}) \in Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})) \cap \tilde{I}$ where $\tilde{I}$ is the set of all random elements $(E_1, E_2, E_3) \in \mathcal{S} \times \mathcal{Y} \times \tilde{\mathcal{Z}}$ such that $E_3 \perp\!\!\!\perp (E_1, E_2)$. Then, the set of $(m_d, \gamma_d)$ consistent with the data and Assumptions RA and EV follows by definition as:

$$\mathcal{H}^{EV/RA}(m_d, \gamma_d) = \left\{ \begin{array}{l} (m_d, \gamma_d) \in \mathcal{M}_d \times \mathcal{P}^{\mathcal{S}} : \exists (\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})) \cap \tilde{I}, \\ \gamma_d \overset{d}{=} \varsigma_d, \;\; m_d(\varsigma_d) = E[\upsilon_d | \varsigma_d] \text{ a.s.} \end{array} \right\} \tag{86}$$

where $\mathcal{M}_d$ is the projection of $\mathcal{M}$ onto its first component. Next, recall the definition of random sets:

$$\mathbf{Y}(d) = \begin{cases} \{Y\}, \text{ if } (D, G) = (d, O) \\ \mathcal{Y}, \text{ otherwise} \end{cases}, \quad \mathbf{S}(d) = \begin{cases} \{S\}, \text{ if } (D, G) \in \{(d, E), (d, O)\} \\ \mathcal{S}, \text{ otherwise} \end{cases} \tag{87}$$

so that $(\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}) = \mathbf{Y}(d) \times \mathbf{S}(d) \times \{\tilde{Z}\}$ for any $d \in \{0, 1\}$.

I now show that $Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})) = Sel(\mathbf{Y}(d)) \times Sel(\mathbf{S}(d)) \times \{\tilde{Z}\}$. Fix an arbitrary $(\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}))$. Then:

$$\begin{aligned} 1 &= P\Big((\upsilon_d, \varsigma_d, \tilde{Z}) \in (\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})\Big) \\ &= P\Big(\upsilon_d \in \mathbf{Y}(d), \varsigma_d \in \mathbf{S}(d), \tilde{Z} \in \{\tilde{Z}\}\Big) \\ &= P\Big(\upsilon_d \in \mathbf{Y}(d), \varsigma_d \in \mathbf{S}(d)\Big) \\ &\leq P\Big(\upsilon_d \in \mathbf{Y}(d)\Big). \end{aligned} \tag{88}$$

where the first line follows since $(\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}))$, the second is by $(\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}) = \mathbf{Y}(d) \times \mathbf{S}(d) \times \{\tilde{Z}\}$, the third and fourth are by observation. Hence $P\Big(\upsilon_d \in \mathbf{Y}(d)\Big) = 1$. By a similar argument, $P\Big(\varsigma_d \in \mathbf{S}(d)\Big) = 1$. Therefore $(\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel(\mathbf{Y}(d)) \times Sel(\mathbf{S}(d)) \times \{\tilde{Z}\}$.

Next, fix an arbitrary $(\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel(\mathbf{Y}(d)) \times Sel(\mathbf{S}(d)) \times \{\tilde{Z}\}$. Then:

$$\begin{aligned}
1 &= P\Big(\upsilon_d \in \mathbf{Y}(d)\Big) \\
&= P\Big(\upsilon_d \in \mathbf{Y}(d), \varsigma_d \in \mathbf{S}(d)\Big) + P\Big(\upsilon_d \in \mathbf{Y}(d), \varsigma_d \notin \mathbf{S}(d)\Big) \\
&= P\Big(\upsilon_d \in \mathbf{Y}(d), \varsigma_d \in \mathbf{S}(d), \tilde{Z} \in \{\tilde{Z}\}\Big) + P\Big(\upsilon_d \in \mathbf{Y}(d), \varsigma_d \notin \mathbf{S}(d)\Big) \\
&= P\Big(\upsilon_d \in \mathbf{Y}(d), \varsigma_d \in \mathbf{S}(d), \tilde{Z} \in \{\tilde{Z}\}\Big) \\
&= P\Big((\upsilon_d, \varsigma_d, \tilde{Z}) \in (\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})\Big).
\end{aligned} \tag{89}$$

where the first line is since $\upsilon_d \in Sel(\mathbf{Y}(d))$, second and third are by observation, fourth is since $P\Big(\upsilon_d \in \mathbf{Y}(d), \varsigma_d \notin \mathbf{S}(d)\Big) \leq P\Big(\varsigma_d \notin \mathbf{S}(d)\Big) = 0$ given that $\varsigma_d \in Sel(\mathbf{S}(d))$, and the last is by $(\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}) = \mathbf{Y}(d) \times \mathbf{S}(d) \times \{\tilde{Z}\}$. Thus, $Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})) = Sel(\mathbf{Y}(d)) \times Sel(\mathbf{S}(d)) \times \{\tilde{Z}\}$. Then write:

$$\begin{aligned}
\mathcal{H}^{EV/RA}(m_d, \gamma_d) &= \left\{ \begin{array}{l} (m_d, \gamma_d) \in \mathcal{M}_d \times \mathcal{P}^{\mathcal{S}} : \exists (\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel(\mathbf{Y}(d)) \times Sel(\mathbf{S}(d)) \times \{\tilde{Z}\} \cap \tilde{I}, \\ \gamma_d \overset{d}{=} \varsigma_d, \;\; m_d(\varsigma_d) = E[\upsilon_d | \varsigma_d] \text{ a.s.} \end{array} \right\} \\
&= \left\{ \begin{array}{l} (m_d, \gamma_d) \in \mathcal{M}_d \times \mathcal{P}^{\mathcal{S}} : \;\; \exists (\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \exists \upsilon_d \in Sel(\mathbf{Y}(d)), \;\; (\upsilon_d, \varsigma_d) \perp\!\!\!\perp \tilde{Z}, \;\; \gamma_d \overset{d}{=} \varsigma_d, \;\; m_d(\varsigma_d) = E[\upsilon_d | \varsigma_d] \text{ a.s.} \end{array} \right\} \\
&= \left\{ \begin{array}{l} (m_d, \gamma_d) \in \mathcal{M}_d \times \mathcal{P}^{\mathcal{S}} : \;\; \exists (\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \exists \upsilon_d \in Sel(\mathbf{Y}(d)), \;\; (\upsilon_d, \varsigma_d) \perp\!\!\!\perp \tilde{Z}, \;\; \gamma_d \overset{d}{=} \varsigma_d, \;\; m_d(\varsigma_d) = E_O[\upsilon_d | \varsigma_d] \text{ a.s.} \end{array} \right\}
\end{aligned} \tag{90}$$

where the first line holds by $Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})) = Sel(\mathbf{Y}(d)) \times Sel(\mathbf{S}(d)) \times \{\tilde{Z}\}$, second is by rearrangement, and third is by $(\upsilon_d, \varsigma_d) \perp\!\!\!\perp \tilde{Z}$.

**Step 2:** Equivalent restrictions on $(m_d, \gamma_d)$ without the modeling assumption.

I show that $\mathcal{H}^{EV/RA}(m_d, \gamma_d)$ is equivalent to:

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

First fix $(m_d, \gamma_d) \in \mathcal{H}^{EV/RA}$. Then, there exist $(\upsilon_d, \varsigma_d)$ such that $m_d(\varsigma_d) = E_O[\upsilon_d | \varsigma_d]$ a.s. and $\gamma_d \overset{d}{=} \varsigma_d$, $(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ and $\upsilon_d \in Sel(\mathbf{Y}(d))$. Hence $(m_d, \gamma_d) \in \tilde{\mathcal{H}}^{EV/RA}(m_d, \gamma_d)$.

Next, fix $(m_d, \gamma_d) \in \tilde{\mathcal{H}}^{EV/RA}(m, \gamma)$ and let $(\upsilon_d, \varsigma_d)$ be the corresponding selections in $Sel(\mathbf{Y}(d)) \times$

$\left[Sel((\mathbf{S}(d),\tilde{Z}))\cap I\right]$ that generate them. I show that there exist $(\upsilon'_d,\varsigma'_d)$ such that: 1)$(\varsigma'_d,\tilde{Z})\in Sel((\mathbf{S}(d),\tilde{Z}))\cap I$ and $\upsilon'_d\in Sel(\mathbf{Y}(d))$; 2) $m_d(\varsigma'_d)=E_O[\upsilon'_d|\varsigma'_d]$ and $\varsigma'_d\overset{d}{=}\gamma_d$; 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 $(\varsigma'_d,\tilde{Z})\in Sel((\mathbf{S}(d),\tilde{Z}))\cap I$ 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 6, by Lemma 3 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.:

$$P_{\upsilon_d,\varsigma_d}(B|\tilde{Z})\geq\left[\begin{array}{l}\mathbb{1}[\exists B_S\subseteq\mathcal{S}:B=\mathcal{Y}\times B_S]P(S\in B_S,D=d|\tilde{Z})+\\ \mathbb{1}[\forall B_S\subseteq\mathcal{S}:B\neq\mathcal{Y}\times B_S]\mathbb{1}[\tilde{Z}\notin\mathcal{Z}]P_O((Y,S)\in B,D=d)\end{array}\right]. \tag{91}$$

Since $(\upsilon_d,\varsigma_d,\tilde{Z})\in Sel(\mathbf{Y}(d))\times\left[Sel(\mathbf{S}(d))\times\{\tilde{Z}\}\cap I\right]$, it must be that $P_{\varsigma_d}(\cdot|\tilde{Z})=P_{\varsigma_d}(\cdot)$ $P-$a.s.. This a restriction on the marginal of $P_{\upsilon_d,\varsigma_d}$, hence for any $B\in\mathcal{C}(\mathcal{Y}\times\mathcal{S})$ such that $B=\mathcal{Y}\times B_S$ for some $B_S\subseteq\mathcal{S}$:[25]

$$P_{\upsilon_d,\varsigma_d}(B|\tilde{Z})=P_{\varsigma_d}(B_S|\tilde{Z})=P_{\varsigma_d}(B_S)=P_{\upsilon_d,\varsigma_d}(B) \tag{92}$$

Where the first equality is by definition of a marginal distribution and $B=\mathcal{Y}\times B_S$, the second is because $P_{\varsigma_d}(\cdot|\tilde{Z})=P_{\varsigma_d}(\cdot)$ $P-$a.s., and the third is by definition of a marginal distribution and $B=\mathcal{Y}\times B_S$.

By (91) and (92), $(\upsilon_d,\varsigma_d,\tilde{Z})\in Sel(\mathbf{Y}(d))\times\left[Sel(\mathbf{S}(d))\times\{\tilde{Z}\}\cap I\right]$ only if $\forall B\in\mathcal{C}(\mathcal{Y}\times\mathcal{S})$ $P$-a.s.:

$$\begin{aligned}P_{\upsilon_d,\varsigma_d}(B|\tilde{Z})&\geq\left[\begin{array}{l}\mathbb{1}[\exists B_S\subseteq\mathcal{S}:B=\mathcal{Y}\times B_S]ess\sup_{\tilde{Z}}P(S\in B_S,D=d|\tilde{Z})+\\ \mathbb{1}[\forall B_S\subseteq\mathcal{S}:B\neq\mathcal{Y}\times B_S]\mathbb{1}[\tilde{Z}\notin\mathcal{Z}]P_O((Y,S)\in B,D=d)\end{array}\right]\\&=\left[\begin{array}{l}\mathbb{1}[\exists B_S\subseteq\mathcal{S}:B=\mathcal{Y}\times B_S]\times\\ \max\left(ess\sup_Z P_E(S\in B_S,D=d|Z),P_O(S\in B_S,D=d)\right)+\\ \mathbb{1}[\forall B_S\subseteq\mathcal{S}:B\neq\mathcal{Y}\times B_S]\mathbb{1}[\tilde{Z}\notin\mathcal{Z}]P_O((Y,S)\in B,D=d)\end{array}\right].\end{aligned} \tag{93}$$

Observe that by (93) $\forall B\in\mathcal{C}(\mathcal{Y}\times\mathcal{S})$:

$$P_{\upsilon_d,\varsigma_d}(B|G=O)\geq\left[\begin{array}{l}\mathbb{1}[\exists B_S\subseteq\mathcal{S}:B=\mathcal{Y}\times B_S]\times\\ \max\left(ess\sup_Z P_E(S\in B_S,D=d|Z),P_O(S\in B_S,D=d)\right)+\\ \mathbb{1}[\forall B_S\subseteq\mathcal{S}:B\neq\mathcal{Y}\times B_S]P_O((Y,S)\in B,D=d)\end{array}\right]. \tag{94}$$

25. Note that the condition need not hold for *every* $B\in\mathcal{C}(\mathcal{Y}\times\mathcal{S})$, only for $B$ s.t. $B=\mathcal{Y}\times B_S$ for some $B_S\subseteq\mathcal{S}$.

Then for any $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ $P-$a.s.:

$$P_{\upsilon'_d,\varsigma'_d}(B) = P_{\upsilon'_d,\varsigma'_d}(B|\tilde{Z}) = P_{\upsilon_d,\varsigma_d}(B|G = O) \geq P_O((Y,S) \in B, D = d) \tag{95}$$

where the first equality is by $P_{\upsilon'_d,\varsigma'_d}(\cdot|\varsigma'_d, \tilde{Z}) = P_{\upsilon'_d,\varsigma'_d}(\cdot|\varsigma'_d)$ and $P_{\varsigma'_d}(\cdot|\tilde{Z}) = P_{\varsigma_d}(\cdot)$, the second is by $P_{\upsilon'_d,\varsigma'_d}(\cdot|\varsigma'_d = s, \tilde{Z} = z) = P_{\upsilon_d,\varsigma_d}(\cdot|\varsigma_d = s, G = O)$, and the inequality is by (93).

For any $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ such that $B = \mathcal{Y} \times B_S$ for some $B_S \subset \mathcal{S}$:

$$\begin{aligned} P_{\upsilon'_d,\varsigma'_d}(B|\tilde{Z}) &= P_{\varsigma'_d}(B_S|\tilde{Z}) = P_{\varsigma_d}(B_S) = P_{\upsilon_d,\varsigma_d}(B) \\ &\geq \max\left(ess\sup_Z P_E(S \in B_S, D = d|Z), P_O(S \in B_S, D = d)\right). \end{aligned} \tag{96}$$

where the first equality follows by definition of a marginal distribution and $B = \mathcal{Y} \times B_S$, the second is by $P_{\varsigma'_d}(B_S|\tilde{Z}) = P_{\varsigma_d}(\cdot)$, third is by definition of a marginal distribution and $B = \mathcal{Y} \times B_S$, and the inequality is by (93).

By (95) and (96) $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$:

$$P_{\upsilon'_d,\varsigma'_d}(B) \geq \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] \times \\ \max\left(ess\sup_Z P_E(S \in B_S, D = d|Z), P_O(S \in B_S, D = d)\right) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] P_O((Y,S) \in B, D = d) \end{array}\right]. \tag{97}$$

Then recall that by Lemma 6, $\exists(\upsilon'_d, \varsigma'_d, \tilde{Z}) \in Sel(\mathbf{Y}(d)) \times Sel(\mathbf{S}(d)) \times \{\tilde{Z}\} \cap \tilde{I}$ if and only if $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ (97) holds. Therefore, there exist $(\varsigma'_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ and $\upsilon'_d \in Sel(\mathbf{Y}(d))$ such that $(\upsilon'_d, \varsigma'_d) \overset{d}{=} P_{\upsilon'_d,\varsigma'_d}$.

Next, note that since $P_{\upsilon'_d,\varsigma'_d}(\cdot) = P_{\upsilon_d,\varsigma_d}(\cdot|G = O)$, then $m_d(\varsigma'_d) = E_O[\upsilon'_d|\varsigma'_d]$ a.s. Because $P_{\varsigma'_d}(\cdot|\tilde{Z}) = P_{\varsigma_d}(\cdot) = P_{\varsigma'_d}(\cdot)$, $\varsigma'_d \overset{d}{=} \varsigma_d \overset{d}{=} \gamma_d$. 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 $(m_d, \gamma_d) \in \tilde{\mathcal{H}}^{EV/RA}(m_d, \gamma_d)$, then $(m_d, \gamma_d) \in \mathcal{H}^{EV/RA}(m_d, \gamma_d)$.

Hence:

$$\mathcal{H}^{EV/RA}(m_d, \gamma_d) = \left\{ \begin{array}{l} (m_d, \gamma_d) \in \mathcal{M}_d \times \mathcal{P}^{\mathcal{S}} : \ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \exists \upsilon_d \in Sel(\mathbf{Y}(d)), \ \gamma_d \overset{d}{=} \varsigma_d, \ m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d] \text{ a.s.}. \end{array} \right\} \tag{98}$$

Finally, impose $E[|Y(d)|] < \infty$. This can equivalently be restated as $Y(d) \in Sel^1(\mathbf{Y}(d))$. Then the identified set for $(m_d, \gamma_d)$ under Assumptions RA and EV, and $E[|Y(d)|] < \infty$ is:

$$\mathcal{H}^{EV/RA/Int}(m_d,\gamma_d)=\left\{\begin{array}{l}(m_d,\gamma_d)\in\mathcal{M}_d\times\mathcal{P}^{\mathcal{S}}:\ \exists(\varsigma_d,\tilde{Z})\in Sel((\mathbf{S}(d),\tilde{Z}))\cap I,\\ \exists\upsilon_d\in Sel^1(\mathbf{Y}(d)),\ \gamma_d\overset{d}{=}\varsigma_d,\ m_d(\varsigma_d)=E_O[\upsilon_d|\varsigma_d]\text{ a.s.}\end{array}\right\}. \tag{99}$$

**Step 3:** Restrictions on $(m,\gamma)$ without the modeling assumption.

Since the data never reveal $(S(0),Y(0))$ and $(S(1),Y(1))$ jointly, Assumptions RA and EV, and $E[|Y(d)|]<\infty$ do not impose cross-restrictions on them. Then the set of all $(m,\gamma)$ consistent with the data Assumptions RA and EV, and $E[|Y(d)|]<\infty$ is:

$$\begin{aligned}&\mathcal{H}^{EV/RA/Int}(m,\gamma)=\mathcal{H}^{EV/RA/Int}(m_0,\gamma_0)\times\mathcal{H}^{EV/RA/Int}(m_1,\gamma_1)\\&=\left\{\begin{array}{l}(m,\gamma_d)\in\mathcal{M}\times(\mathcal{P}^{\mathcal{S}}))^2:\ \forall d\in\{0,1\},\ \exists(\varsigma_d,\tilde{Z})\in Sel((\mathbf{S}(d),\tilde{Z}))\cap I,\\ \exists\upsilon_d\in Sel^1(\mathbf{Y}(d)),\ \gamma_d\overset{d}{=}\varsigma_d,\ m_d(\varsigma_d)=E_O[\upsilon_d|\varsigma_d]\text{ a.s.}\end{array}\right\}.\end{aligned} \tag{100}$$

**Step 4:** Identified set $\mathcal{H}(m,\gamma)$.

It only remains to impose Assumption MA. To do so, observe that a valid identified set is:

$$\begin{aligned}&\mathcal{H}(m,\gamma)=\mathcal{H}^{EV/RA/Int}(m,\gamma)\cap(\mathcal{M}^A\times(\mathcal{P}^{\mathcal{S}})^2)\\&=\left\{\begin{array}{l}(m,\gamma_d)\in\mathcal{M}^A\times(\mathcal{P}^{\mathcal{S}}))^2:\ \forall d\in\{0,1\},\ \exists(\varsigma_d,\tilde{Z})\in Sel((\mathbf{S}(d),\tilde{Z}))\cap I,\\ \exists\upsilon_d\in Sel^1(\mathbf{Y}(d)),\ \gamma_d\overset{d}{=}\varsigma_d,\ m_d(\varsigma_d)=E_O[\upsilon_d|\varsigma_d]\text{ a.s.}\end{array}\right\}.\end{aligned} \tag{101}$$

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

**Theorem 2.** *Let Assumptions RA, EV, and MA hold. Suppose $\mathcal{S}$ is a finite set and that $\mathcal{M}^A$ is closed and convex. Then:*

$$\mathcal{H}(\tau)=\left[\min_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)}T(\tilde{m},\tilde{\gamma}),\max_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)}T(\tilde{m},\tilde{\gamma})\right] \tag{22}$$

*where:*

$$\mathcal{H}(m,\gamma)=\left\{\begin{array}{l}(m,\gamma)\in\mathcal{M}^A\times(\Delta(k))^2:\forall d\in\{0,1\},\ \forall s\in\mathcal{S},\\ \gamma_d(s)\geq\max\left(ess\sup_Z P_E(S=s,D=d|Z),P_O(S=s,D=d)\right),\\ m_d(s)\gamma_d(s)\geq E_O[Y|S=s,D=d]P_O(S=s,D=d)\ ,\\ (1-m_d(s))\,\gamma_d(s)\geq E_O[1-Y|S=s,D=d]P_O(S=s,D=d)\end{array}\right\} \tag{23}$$

*Proof of Theorem 2.* The proof proceeds through a series of steps:

1. Characterizing $\mathcal{H}(m,\gamma)$;

2. Proving that $T(m,\gamma)$ is jointly continuous;

3. Proving that $\mathcal{H}(m,\gamma)$ is convex;

4. Proving that $\mathcal{H}(m,\gamma)$ is compact;

5. Proving that $\mathcal{H}(\tau)$ is an interval.

**Step 1:** Characterizing $\mathcal{H}(m,\gamma)$.

For any selection $v_d \in \mathbf{Y}(d)$, $v_d \in \mathcal{Y}$ so $E[|v_d|] \leq |\sup \mathcal{Y}| < \infty$ where the strict inequality follows by boundedness of $\mathcal{Y}$. Hence, $\mathbf{Y}(d)$ is integrably bounded. Since $\mathcal{S} = \{1, 2, \ldots, k\}$, represent $\gamma_d$ 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:

$$\begin{aligned}\mathcal{H}(m,\gamma) &= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\, \forall B \in \mathcal{C}(\mathcal{S}), \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ \forall u \in \{-1,1\}\text{: } um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s))\ \gamma_d\text{−a.e.} \end{array} \right\} \\ &= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\, \forall B \in \mathcal{C}(\mathcal{S}), \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ \forall u \in \{-1,1\}\text{: } um_d(s) \leq u\mu_d(s)\frac{P_O(S=s,D=d)}{\gamma_d(s)} + h_{co(\mathcal{Y})}(u)\left(1 - \frac{P_O(S=s,D=d)}{\gamma_d(s)}\right)\ \gamma_d\text{−a.e.} \end{array} \right\} \end{aligned} \tag{102}$$

where the first line is by Theorem 1. The second is by definition of $\pi_{\gamma_d}(s)$ and $\gamma_d$ 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:

$$\begin{aligned}
\mathcal{H}(m,\gamma) &= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\, \forall s \in \mathcal{S},\, \forall u \in \{-1,1\}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right), \\ um_d(s) \leq u\mu_d(s)\frac{P_O(S=s,D=d)}{\gamma_d(s)} + h_{co(\mathcal{Y})}(u)\left(1 - \frac{P_O(S=s,D=d)}{\gamma_d(s)}\right)\ \gamma_d\text{−a.e.} \end{array} \right\} \\
&= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\, \forall s \in \mathcal{S},\, \forall u \in \{-1,1\}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right), \\ um_d(s) \leq uE[Y|S=s,D=d]\frac{P_O(S=s,D=d)}{\gamma_d(s)} + h_{co(\mathcal{Y})}(u)\left(1 - \frac{P_O(S=s,D=d)}{\gamma_d(s)}\right)\ \gamma_d\text{−a.e.} \end{array} \right\} \\
&= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\, \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right), \\ m_d(s)\gamma_d(s) \geq E_O[Y|S=s,D=d]P_O(S=s,D=d)\ \gamma_d\text{−a.e.}, \\ (1-m_d(s))\gamma_d(s) \geq E_O[1-Y|S=s,D=d]P_O(S=s,D=d)\ \gamma_d\text{−a.e.} \end{array} \right\} \\
&= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\, \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right), \\ m_d(s)\gamma_d(s) \geq E_O[Y|S=s,D=d]P_O(S=s,D=d), \\ (1-m_d(s))\gamma_d(s) \geq E_O[1-Y|S=s,D=d]P_O(S=s,D=d) \end{array} \right\}.
\end{aligned} \tag{103}$$

where the first line is by Theorem 1 and (102), 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.

**Step 2:** $T$ is jointly continuous.

Endow the set of reals with its natural topology, making it a locally convex topological vector space (t.v.s.). By bilinearity of the Riemann-Stieltjes integral in the integrand and integrator, $T(m,\gamma)$ is a bilinear map. Since $T$ is a bilinear map in a finite-dimensional space, it is separately continuous in each argument. Note that $T : \mathbb{R}^{2d_s} \times \mathbb{R}^{2d_s} \to \mathbb{R}$ and that $\mathbb{R}^{2d_s}$ is Polish (separable and completely metrizable), and hence metrizable. By a corollary of the first Baire category theorem, every Polish space is a Baire space, so $\mathbb{R}^{2d_s}$ is a Baire space (Willard (2004, Corollary 25.4)). By Corollary 3, $T$ is jointly continuous since every separately continuous bilinear map from a product of a Baire space and a metrizable space to a locally convex t.v.s. is jointly continuous.

**Step 3:** $\mathcal{H}(m,\gamma)$ is convex.

Define the following set:

$$\begin{aligned}\mathcal{H}^{WC}(m,\gamma) :=& \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (int(\Delta(k)))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right), \\ m_d(s)\gamma_d(s) \geq E_O[Y|S=s,D=d]P_O(S=s,D=d)\ , \\ (1-m_d(s))\gamma_d(s) \geq E_O[1-Y|S=s,D=d]P_O(S=s,D=d) \end{array} \right\} \\ =& \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (int(\Delta(k)))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right), \\ m_d(s) \geq E_O[Y|S=s,D=d]\frac{P_O(S=s,D=d)}{\gamma_d(s)} \\ m_d(s) \leq E_O[Y|S=s,D=d]\frac{P_O(S=s,D=d)}{\gamma_d(s)} + 1 - \frac{P_O(S=s,D=d)}{\gamma_d(s)}, \end{array} \right\}. \end{aligned} \tag{104}$$

where the second line is by rearrangement and the fact that $\gamma \in (int(\Delta(k)))^2$ so $\gamma_d(s) > 0$ for any $s \in \mathcal{S}$ and $d \in \{0,1\}$. It is immediate that $cl\left(\mathcal{H}^{WC}(m,\gamma)\right) = \mathcal{H}(m,\gamma)$. I first prove that $\mathcal{H}^{WC}(m,\gamma)$ is convex, which is sufficient for $\mathcal{H}(m,\gamma)$ to be convex. Pick any $(m,\gamma),(m',\gamma') \in \mathcal{H}^{WC}(m,\gamma)$ and fix $a \in (0,1)$. It remains to show that $a(m,\gamma)+(1-a)(m',\gamma') \in \mathcal{H}^{WC}(m,\gamma)$.

$\mathcal{M}^A$ is convex by assumption so $am+(1-a)m' \in \mathcal{M}^A$. $\Delta(k)$ is the $k$-dimensional simplex, and thus convex. The interior of a convex set is convex, so $int(\Delta(k))$ and $(int(\Delta(k)))^2$ are convex. Therefore, $a\gamma+(1-a)\gamma \in (int(\Delta(k)))^2$. Observe that for any $d \in \{0,1\}$ and $s \in \mathcal{S}$:

$$a\gamma_d(s)+(1-a)\gamma'_d(s) \geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right) \tag{105}$$

since both $\gamma_d$ and $\gamma'_d$ satisfy the same condition. Next, note that for any $d \in \{0,1\}$ and $s \in \mathcal{S}$, recalling that $\gamma_d(s) > 0$ and $\gamma'_d(s) > 0$:

$$\begin{aligned} &\frac{a\gamma'_d(s)+(1-a)\gamma_d(s)}{\gamma_d(s)\gamma'_d(s)} - \frac{1}{a\gamma_d(s)+(1-a)\gamma'_d(s)} \\ &= \frac{(a\gamma'_d(s)+(1-a)\gamma_d(s))(a\gamma_d(s)+(1-a)\gamma'_d(s)) - \gamma_d(s)\gamma'_d(s)}{\gamma_d(s)\gamma'_d(s)(a\gamma_d(s)+(1-a)\gamma'_d(s))} \\ &= \frac{(a^2+(1-a)^2-1)\gamma'_d(s)\gamma_d(s) + a(1-a)(\gamma'_d(s)^2+\gamma_d(s)^2)}{\gamma_d(s)\gamma'_d(s)(a\gamma_d(s)+(1-a)\gamma'_d(s))} \\ &= \frac{2a(a-1)\gamma'_d(s)\gamma_d(s) + a(1-a)(\gamma'_d(s)^2+\gamma_d(s)^2)}{\gamma_d(s)\gamma'_d(s)(a\gamma_d(s)+(1-a)\gamma'_d(s))} \\ &= \frac{a(1-a)(\gamma_d(s)-\gamma_d(s)')^2}{\gamma_d(s)\gamma'_d(s)(a\gamma_d(s)+(1-a)\gamma'_d(s))} \geq 0. \end{aligned} \tag{106}$$

Then for any $d \in \{0,1\}$ and $s \in \mathcal{S}$:

$$\begin{aligned} am_d(s) + (1-a)m'_d(s) &\geq E_O[Y|S=s, D=d]P_O(S=s, D=d)\left(\frac{a}{\gamma_d(s)} + \frac{1-a}{\gamma'_d(s)}\right) \\ &= E_O[Y|S=s, D=d]P_O(S=s, D=d)\left(\frac{a\gamma'_d(s) + (1-a)\gamma_d(s)}{\gamma_d(s)\gamma'_d(s)}\right) \\ &\geq E_O[Y|S=s, D=d])\frac{P_O(S=s, D=d)}{a\gamma_d(s) + (1-a)\gamma'_d(s)} \end{aligned} \tag{107}$$

where the first line follows by $(m,\gamma),(m',\gamma') \in \mathcal{H}^{WC}(m,\gamma)$, second is by observation and the third is by (106). Finally, for any $d \in \{0,1\}$ and $s \in \mathcal{S}$:

$$\begin{aligned} am_d(s) + (1-a)m'_d(s) &\leq (E_O[Y|S=s, D=d] - 1)\, P_O(S=s, D=d)\left(\frac{a}{\gamma_d(s)} + \frac{1-a}{\gamma'_d(s)}\right) + 1 \\ &\leq (E_O[Y|S=s, D=d] - 1)\,\frac{P_O(S=s, D=d)}{a\gamma_d(s) + (1-a)\gamma'_d(s)} + 1 \end{aligned} \tag{108}$$

where $(m,\gamma),(m',\gamma') \in \mathcal{H}^{WC}(m,\gamma)$ yields the first line, and the second follows by (106) and $(E_O[Y|S=s, D=d] - 1) \leq 0$. Hence, $a(m,\gamma) + (1-a)(m',\gamma') \in \mathcal{H}^{WC}(m,\gamma)$, and $\mathcal{H}^{WC}(m,\gamma)$ is convex. Since closure preserves convexity, $\mathcal{H}(m,\gamma) = cl\left(\mathcal{H}^{WC}(m,\gamma)\right)$ is convex.

**Step 4:** $\mathcal{H}(m,\gamma)$ is compact

It is immediate that $\mathcal{H}(m,\gamma)$ is bounded since $\mathcal{H}(m,\gamma) \subseteq [0,1]^k \times (\Delta(k))^2$, by bounded support of $Y(d)$ and $\mathcal{S}$ being a finite set. That it is closed is immediate by definition of a closure. Then, $\mathcal{H}(m,\gamma)(m,\gamma)$ is compact.

**Step 5:** $\mathcal{H}(\tau)$ is an interval.

$T$ was shown to be a continuous map, so it preserves connectedness. Hence, $\mathcal{H}(\tau) = \{T(m,\gamma) : (m,\gamma) \in \mathcal{H}(m,\gamma)\}$ is a connected set. Since $\mathcal{H}(\tau) \subseteq \mathbb{R}$, it is an interval. Continuous images preserve compactness, so the $\mathcal{H}(\tau)$ is a compact interval, so:

$$\mathcal{H}(\tau) = \left[\inf_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m,\gamma), \sup_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m,\gamma)\right] \tag{109}$$

$$= \left[\min_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m,\gamma), \max_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m,\gamma)\right] \tag{110}$$

where the second line follows by continuity of $T$ and compactness of $\mathcal{H}(m,\gamma)$. □

**Theorem 3.** *Let Assumptions RA, EV, MA, and E hold. Then as $n \to \infty$:*

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

*Proof of Theorem 3.* Note that $\mathcal{H}(\tau)$ and $\mathcal{H}_n(\tau)$ are both closed intervals by Theorem 2 and the definition of the latter. Then, by definition of the Hausdorff distance, it is sufficient to show that boundaries of $\mathcal{H}_n(\tau)$ converge in probability to boundaries of $\mathcal{H}(\tau)$ as $n \to \infty$. Considering the upper bounds of $\mathcal{H}(\tau)$ and $\mathcal{H}_n(\tau)$, I show that for any $\varepsilon > 0$:

$$\limsup_{n \to \infty} P\left( \left| \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}_n(m,\gamma)} T(\tilde{m},\tilde{\gamma}) - \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}) \right| > \varepsilon \right) = 0 \tag{111}$$

and the argument for the lower bounds is symmetric. Fix any $\varepsilon > 0$ and note that:

$$\left| \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}_n(m,\gamma)} T(\tilde{m},\tilde{\gamma}) - \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}) \right| \leq \sup_{||(m,\gamma)-(m',\gamma')|| \leq d_H(\mathcal{H}(m,\gamma),\mathcal{H}_n(m,\gamma))} |T(m,\gamma) - T(m',\gamma')|. \tag{112}$$

$\mathcal{M}$ is a set of finite-dimensional vectors with finite components and therefore compact. $\mathcal{M} \times (\Delta(k))^2$ is then also compact. Proof of Theorem 2 shows that $T$ is a jointly continuous functional under the maintained assumptions. Hence, $T : \mathcal{M} \times (\Delta(k))^2 \to \mathbb{R}$ is uniformly continuous over its domain by the Heine-Cantor theorem. For the fixed $\varepsilon$, let $\varepsilon' = 2\varepsilon > 0$. By uniform continuity, there exists a $\delta' > 0$ such that $||(m,\gamma) - (m',\gamma')|| < \delta'$ implies $|T(m,\gamma) - T(m',\gamma')| < \varepsilon'$. Let $\delta = \delta'/2 > 0$. If $|T(m,\gamma) - T(m',\gamma')| \geq \varepsilon' > \varepsilon$ it must be that $d_H(\mathcal{H}(m,\gamma), \mathcal{H}_n(m,\gamma)) \geq \delta' > \delta$. Therefore:

$$P\left( \left| \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}_n(m,\gamma)} T(\tilde{m},\tilde{\gamma}) - \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}) \right| > \varepsilon \right) \leq P\left( d_H(\mathcal{H}(m,\gamma), \mathcal{H}_n(m,\gamma)) > \delta \right) \tag{113}$$

Thus, to prove (111), it is sufficient to show that given $\delta > 0$:

$$\limsup_{n \to \infty} P(d_H(\mathcal{H}(m,\gamma), \mathcal{H}_n(m,\gamma)) > \delta) = 0. \tag{114}$$

Therefore to prove (111), it is sufficient to prove $d_H(\mathcal{H}(m,\gamma), \mathcal{H}_n(m,\gamma)) \xrightarrow{p} 0$. To do so, I adapt the arguments in Russell (2021, Theorem 2). Let $\mu_d$ be a $k$-dimensional vector with components $\mu_d(s) = E_O[Y|S = s, D = d]$. Let $\eta_d$ be a $k \times |\tilde{\mathcal{Z}}|$ matrix with the element $(s,\tilde{z})$ being $\eta_d(s,\tilde{z}) = P(S = s, D = d|\tilde{Z} = z)$. Finally, collect $\beta = (\mu_0, \mu_1, \eta_0, \eta_1, \tilde{\beta}) \in \mathfrak{B}$ where $\tilde{\beta}$ is a vector of other population distribution features that are consistently estimable and used in the definition of $\mathcal{M}^A$. By Assumption E *iii)*, there is an estimator $\tilde{\beta}_n$ such that $\tilde{\beta}_n \xrightarrow{p} \tilde{\beta}$ as

$n \to \infty$. Therefore, by elementary arguments, also $\beta_n \xrightarrow{p} \beta$ as $n \to \infty$. By the same assumption, $\mathcal{M}^A = \{m \in \mathcal{M} : h(m,\beta) \geq 0, \;\; g(m,\beta) = 0\}$ for some known linear functions $g$ and $h$. Then:

$$
\begin{aligned}
\mathcal{H}(m,\gamma) &= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\}, \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(\max_{z \in \mathcal{Z}} P_E(S=s, D=d|Z=z), P_O(S=s, D=d)\right), \\ m_d(s)\gamma_d(s) \geq E_O[Y|S=s, D=d]P_O(S=s, D=d) \;, \\ (1-m_d(s))\,\gamma_d(s) \geq E_O[1-Y|S=s, D=d]P_O(S=s, D=d) \end{array} \right\} \\
&= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{Y}^{2k} \times (\Delta(k))^2 : h(m,\beta) \geq 0, \;\; g(m,\beta) = 0, \;\; \forall d \in \{0,1\}, \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(\max_{z \in \mathcal{Z}} P_E(S=s, D=d|Z=z), P_O(S=s, D=d)\right), \\ m_d(s)\gamma_d(s) \geq E_O[Y|S=s, D=d]P_O(S=s, D=d) \;, \\ (1-m_d(s))\,\gamma_d(s) \geq E_O[1-Y|S=s, D=d]P_O(S=s, D=d) \end{array} \right\} \\
&= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{Y}^{2k} \times (\Delta(k))^2 : h(m,\beta) \geq 0, \;\; g(m,\beta) = 0, \\ \forall d \in \{0,1\}, \forall s \in \mathcal{S}, \forall \tilde{z} \in \tilde{\mathcal{Z}}, \\ \gamma_d(s) \geq \eta_d(s,\tilde{z}), \\ m_d(s)\gamma_d(s) \geq \mu_d(s)\eta_d(s, \sup \tilde{\mathcal{Z}} + 1) \;, \\ (m_d(s)-1)\gamma_d(s) \leq (\mu_d(s)-1)\eta_d(s, \sup \tilde{\mathcal{Z}} + 1) \end{array} \right\}
\end{aligned}
\tag{115}
$$

where the first line follows by Theorem 2 and Assumption E *ii)*, the second line is by definition of $\mathcal{M}^A$, and the third is by definition $\eta_d$ and $\mu_d$ and $\tilde{Z}$. Hence, $\mathcal{H}(m,\gamma)$ can be equivalently represented through a set of equality and inequality constraints as:

$$
\mathcal{H}(m,\gamma) = \left\{ \;\; (m,\gamma) \in \mathcal{Y}^{2k} \times (\Delta(k))^2 : \tilde{h}(m,\gamma,\beta) \geq 0, \;\; g(m,\beta) = 0 \;\; \right\}. \tag{116}
$$

where $\tilde{h}(m,\gamma,\beta)$ collects all linear inequality restrictions $h(m,\beta) \geq 0$ and remaining linear and bilinear inequality constraints in (115).

Next, convert all inequality constraints $\tilde{h}(m,\gamma,\beta)$ to equality constraints by introducing slackness parameters $\lambda_t \in [0,1]$ for each inequality constraint, as in Shi and Shum (2015, Remark pp. 497).[26] Denote by $\lambda$ the vector of all slackness parameters, and let $\theta = (m,\gamma,\lambda) \in \mathfrak{T}$ be a vector of dimension $d_\theta \times 1$. Write all converted equality constraints and existing equality constraints $g(m,\beta)$ as $\tilde{g}(\theta,\beta) = 0$. Define also the inequality constraints $h_\lambda(\theta) \geq 0$, which collect non-negativity constraints $\lambda_t \geq 0$. Now define $\Theta = \{\theta : \tilde{g}(\theta,\beta) = 0, \;\; h_\lambda(\theta) \geq 0.\}$. Under the

26. Note that for proofs of consistency, it is sufficient to just add slackness parameters to each inequality constraint.

assumptions, both $\mathcal{H}(m,\gamma)$ and $\Theta$ are non-empty. Therefore, equivalently write:

$$\Theta = \underset{\theta\in\mathfrak{T}:h_\lambda(\theta)\geq 0}{\arg\min}\ \tilde{g}(\theta,\beta)'\tilde{g}(\theta,\beta) \tag{117}$$

and let the corresponding estimator be:

$$\Theta_n = \underset{\theta\in\mathfrak{T}:h_\lambda(\theta)\geq 0}{\arg\min}\ \tilde{g}(\theta,\beta_n)'\tilde{g}(\theta,\beta_n). \tag{118}$$

Note that for any $(m,\gamma)\in\mathcal{H}(m,\gamma)$ if and only if $(m,\gamma,\lambda)\in\Theta$ for some feasible $\lambda$. Then, the projection of $\Theta$ onto the first two components $(m,\gamma)$ is $\mathcal{H}(m,\gamma)$. Therefore, whenever $\tilde{\mathcal{H}}_n(m,\gamma)\neq\emptyset$, $\tilde{\mathcal{H}}_n(m,\gamma)$ is numerically equivalent to the projection of $\Theta_n$ onto $(m,\gamma)$. Moreover, since $\beta_n \xrightarrow{p} \beta$ as $n\to\infty$, $P(\tilde{\mathcal{H}}_n(m,\gamma)\neq\emptyset)\to 1$ (see Yildiz (2012, Footnote 10)). Thus, for (114) and therefore (111), it is sufficient to show that $d_H(\Theta_n,\Theta)\xrightarrow{p}0$. This follows immediately by verifying the conditions of Shi and Shum (2015, Theorem 2.1).

First, the preceding arguments argue that $\beta_n\xrightarrow{p}\beta$. Second, for $d\in\{0,1\}$, $\mu_d,m_d\in[0,1]^k$, $\eta_d,\gamma_d\in\Delta(k)$, $\lambda_t\in[0,1]$ for all $t<\infty$, hence the parameter spaces for $\mathfrak{T}$ and $\mathfrak{B}$ are compact. Third, $\tilde{g}(\cdot,\beta)$ is continuously differentiable for $\beta\in\mathfrak{B}$ as it is bilinear in $\theta$; $h_\lambda(\cdot)$ is linear in $\theta$ and hence continuous. Applying identical arguments of Step 4 in the proof Russell (2021, Theorem 2) then yields $d_H(\Theta_n,\Theta)\xrightarrow{p}0$.

□

**Proposition 1.** *Suppose Assumptions RA and EV hold. Then:*

*i)* $\mathcal{H}^O(\tau)=\mathcal{H}(\tau)$;

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

*Proof of Proposition 1.* I show that $\mathcal{H}^O(P_{Y(0),Y(1)})=\mathcal{H}(P_{Y(0),Y(1)})$, which immediately yields $\mathcal{H}^O(\tau)=\mathcal{H}(\tau)$.

The data never reveal $(Y(0),Y(1))$ jointly, so the data and assumptions do not impose cross-restrictions on $Y(0)$ and $Y(1)$. Then the identified set for $P_{Y(0),Y(1)}$ *given* $P_{Y(1)}$ and $P_{Y(0)}$ is the set of all joint distributions consistent with the marginals $P_{Y(1)}$ and $P_{Y(0)}$. The identified set for $P_{Y(0),Y(1)}$ is the union of such sets over all possible $(P_{Y(0)},P_{Y(1)})$.

To that end, let $\Pi(\nu_0,\nu_1)$ be the set of couplings of probability measures $\nu_0$ and $\nu_1$ defined as (Villani et al. (2009, Definition 1.1)):

$$\Pi(\nu_0,\nu_1)=\left\{\delta\in\mathcal{P}^{\mathcal{Y}}\times\mathcal{P}^{\mathcal{Y}}:\ \forall A\subseteq\mathcal{Y}\quad \begin{array}{l}\delta(A\times\mathcal{Y})=\nu_0(A),\\ \delta(\mathcal{Y}\times A)=\nu_1(A)\end{array}\right\}. \tag{119}$$

$\Pi(\nu_0, \nu_1)$ is always non-empty (Galichon (2018, Section 2.1)). Equivalently, the identified set for $P_{Y(0),Y(1)}$ *given* $P_{Y(1)}$ and $P_{Y(0)}$ is $\Pi(P_{Y(0)}, P_{Y(1)})$. Using the identified sets for the marginals $P_{Y(d)} \in \mathcal{H}^O(P_{Y(d)})$ for $d \in \{0,1\}$, the identified set $\mathcal{H}^O(P_{Y(0),Y(1)})$ is then the union of all possible couplings:

$$\mathcal{H}^O(P_{Y(0),Y(1)}) = \bigcup_{(\nu_0,\nu_1)\in\mathcal{H}^O(P_{Y(0)})\times\mathcal{H}^O(P_{Y(1)})} \Pi(\nu_0, \nu_1). \tag{120}$$

Similarly for $\mathcal{H}(P_{Y(0),Y(1)})$:

$$\mathcal{H}(P_{Y(0),Y(1)}) = \bigcup_{(\nu_0,\nu_1)\in\mathcal{H}(P_{Y(0)})\times\mathcal{H}(P_{Y(1)})} \Pi(\nu_0, \nu_1). \tag{121}$$

Lemma 7 shows that $\mathcal{H}^O(P_{Y(d)}) = \mathcal{H}(P_{Y(d)})$ for any $d \in \{0,1\}$. That $\mathcal{H}^O(P_{Y(0),Y(1)}) = \mathcal{H}(P_{Y(0),Y(1)})$ follows.

Next, observe that $\tau$ is a functional of $P_{Y(0),Y(1)}$. It is then immediate that $\mathcal{H}^O(\tau) = \mathcal{H}(\tau)$,

**Remark 11.** The same result may be obtained directly by defining the random set:

$$(\mathbf{Y}, \mathbf{S}) = \begin{cases} \mathcal{S} \times \{S\} \times \mathcal{Y} \times \{Y\}, \text{ if } (D,G) = (1,O) \\ \mathcal{S} \times \{S\} \times \mathcal{Y} \times \mathcal{Y}, \text{ if } (D,G) = (1,E) \\ \{S\} \times \mathcal{S} \times \{Y\} \times \mathcal{Y}, \text{ if } (D,G) = (0,O) \\ \{S\} \times \mathcal{S} \times \mathcal{Y} \times \mathcal{Y}, \text{ if } (D,G) = (0,E) \end{cases} \tag{122}$$

which summarizes all information on $(S(0), S(1), Y(0), Y(1))$, and retracing the steps of Lemmas 5, 6 and 7 for the joint distribution $P_{Y(0),Y(1)}$.

□

**Lemma 2.** *(Nested Misspecification) Let $\tilde{\mathcal{H}} \subseteq \tilde{\mathcal{H}}^{O/A}$ be misspecified identified sets for some parameter $\tau$. Let $d$ be the point-to-set distance defined as $d(A,t) := \inf\{\|t-a\| : a \in A\}$ for $A \subseteq \mathbb{R}$ and $t \in \mathbb{R}$. Then:*

$$d(\tilde{\mathcal{H}}^{O/A}, \tau) \leq d(\tilde{\mathcal{H}}, \tau)$$

*Proof of Lemma 2.*

$$d(\tilde{\mathcal{H}}^{O/A}, \tau) = \inf\left\{\|t-\tau\| : t \in \tilde{\mathcal{H}}^{O/A}\right\} \leq \inf\left\{\|t-\tau\| : t \in \tilde{\mathcal{H}}\right\} = d(\tilde{\mathcal{H}}, \tau) \tag{123}$$

where the inequality follows by $\tilde{\mathcal{H}} \subseteq \tilde{\mathcal{H}}^{O/A}$. □

**Corollary 1.** *Suppose Assumptions RA and EV hold. If $\mathcal{Y} = \mathbb{R}$, the identified set for $\tau$ is $\mathcal{H}(\tau) = \mathbb{R}$. If $\mathcal{Y} = [0,1]$:*

$$\mathcal{H}(\tau) = [E_O[YD] - E_O[Y(1-D)] - P_O(D=0), E_O[YD] - E_O[Y(1-D)] + P_O(D=1)]\,. \quad (27)$$

*In both cases, $0 \in \mathcal{H}(\tau)$ and the sign of $\tau$ not identified.*

*Proof of Corollary 1.* Suppose first that Assumptions RA and EV hold. Assume that $\mathcal{Y} = \mathbb{R}$ and pick an arbitrary $\tilde{c} \in \mathbb{R}$. I show that $\tilde{c} \in \mathcal{H}^O(\tau)$ which is equivalent to $\tilde{c} \in \mathcal{H}(\tau)$ by Proposition 1.

Define a distribution function for any $(a,d) \in \mathbb{R} \times \{0,1\}$:

$$\gamma_{a|d}(B) = P_O(Y \in B, D = d) + \mathbb{1}[a \in B] P_O(D \neq d)$$

for any Borel set $B \in \mathcal{B}(\mathcal{Y})$. Recall from Lemma 7 that:

$$\mathcal{H}^O(P_{Y(d)}) = \{\gamma \in \mathcal{P}^{\mathcal{Y}} : \gamma(B) \geq P_O(Y \in B, D = d)\ \forall B \in \mathcal{C}(\mathcal{Y})\}.$$

Since $\mathcal{C}(\mathcal{Y}) \subseteq \mathcal{B}(\mathcal{Y})$, then $\gamma_{a|d}(B) \in \mathcal{H}^O(P_{Y(d)})$ for any $(a,d) \in \mathbb{R} \times \{0,1\}$. Note also that any coupling of $\gamma_{a|1} \in \mathcal{H}^O(P_{Y(1)})$ and $\gamma_{a'|0} \in \mathcal{H}^O(P_{Y(0)})$ is compatible with the observed data.

Next, observe that $\gamma_{a|d}$ is a pushforward measure of the random variable $Y\mathbb{1}[D=d] + a\mathbb{1}[D \neq d]$, which has the expectation of $E[Y\mathbb{1}[D=d]] + aP_O(D \neq d)$. Let $c = \frac{\tilde{c} - E_O[YD] + E_O[Y(1-D)]}{P_O(D \neq d)} \in \mathbb{R}$. Then $\gamma_{c|1}$ yields the expected value:

$$E[YD] + \frac{\tilde{c} - E_O[YD] + E_O[Y(1-D)]}{P_O(D=0)} P_O(D=0) = \tilde{c} + E_O[Y(1-D)].$$

Similarly, $\gamma_{0|0}$ yields the expected value $E[Y(1-D)]$.

Now take $\gamma_{c|1} \in \mathcal{H}^O(P_{Y(1)})$ and $\gamma_{0|0} \in \mathcal{H}^O(P_{Y(0)})$ as distribution functions of $Y(1)$ and $Y(0)$, recalling that any coupling of $\gamma_{c|1}$ and $\gamma_{c|0}$ is compatible with the observed data. It follows that $\tau = E[Y(1) - Y(0)] = \tilde{c}$. Since $\tilde{c}$ was arbitrary, $\mathcal{H}^O(\tau) = \mathbb{R}$. By Proposition 1, $\mathcal{H}^O(\tau) = \mathcal{H}(\tau)$.

Next, let $\mathcal{Y} = [0,1]$. Since Proposition 1 holds for any $\mathcal{Y} \subseteq \mathbb{R}$, I can again recover $\mathcal{H}(\tau)$ by using only distributions in $\mathcal{H}^O(P_{Y(d)})$. Equivalently, I can find $\mathcal{H}(\tau)$ by utilizing only information in the observational data. Then, by elementary arguments as in Manski (1990), the bounds in (27) follow. □

**Proposition 2.** *Let Assumptions EV and LUC hold.*

*i) Suppose the observed data distribution $P_O(Y,S,D)$ is such that $V_O[Y|S,D=d] > 0$ $P-$a.s. for some $d \in \{0,1\}$ and that $\mathcal{Y}$ is a bounded set. Then $\mathcal{H}^{O/LUC}(\tau) \subsetneq \mathcal{H}^O(\tau)$.*

*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 $\tau$ is point-identified, and $\mathcal{H}(\tau) = \mathcal{H}^{O/LUC}(\tau)$.*

*Proof of Proposition 2.* I prove the claims in order.

*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 $\mathcal{H}^O(\tau)$ is:

$$\tau \leq E_O[Y(2D-1)] + \sup \mathcal{Y} P_O(D=1) - \inf \mathcal{Y} P_O(D=0) = \sup \mathcal{H}^O(\tau). \tag{124}$$

By Lemma 9 $V_O[Y|S,D=d] > 0$ $P-$a.s. implies $E_O[Y|S,D=d] < \sup \mathcal{Y}$ $P-$a.s. If there exists $d \in \{0,1\}$ s.t. $V[Y|S,D=d] > 0$ $P-$a.s., then it must be that for every Borel subset $B \subseteq \mathcal{B}(\mathcal{S})$ with $P_O(S \in B|D=d) > 0$ we have $E_O[Y|S \in B, D=d] < \sup \mathcal{Y}$. Under Assumption LUC then:

$$\begin{aligned} E_O[Y(d)|D \neq d] &= E_O[E_O[Y(d)|S(d), D \neq d]|D \neq d] P_O(D \neq d) \\ &= E_O[E_O[Y(d)|S(d), D = d|D \neq d]] P_O(D \neq d) \\ &= E_O[E_O[Y|S, D=d]|D \neq d] P_O(D \neq d) \\ &< \sup \mathcal{Y} P_O(D \neq d) \end{aligned} \tag{125}$$

where the first line is by LIE, second by Assumption LUC, third by definition, and the fourth since $E[Y|S \in B, D=d] < \sup \mathcal{Y}$ for every Borel set $B$ of positive measure. Then under LUC:

$$\begin{aligned} E[Y(d)] &= E_O[Y \mathbb{1}[D=d]] + E[Y(d)|D \neq d] P_O(D \neq d) \\ &< E_O[Y \mathbb{1}[D=d]] + \sup \mathcal{Y} P_O(D \neq d). \end{aligned} \tag{126}$$

Therefore, under Assumption LUC:

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

where the inequality follows by (126). Thus $\sup \mathcal{H}^{O/LUC}(\tau) < \sup \mathcal{H}^O(\tau)$. So there must exist a point in $\mathcal{H}(\tau)$ which is not contained in $\mathcal{H}^{O/LUC}(\tau)$. Conclude that $\mathcal{H}^{O/LUC}(\tau) \subsetneq \mathcal{H}^O(\tau)$.

*ii)*

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

Then, following the same steps as in (125):

$$\begin{aligned} E_O[Y(d)|\neq d] &= E_O[E_O[Y(d)|S(d), D\neq d]|D\neq d]P_O(D\neq d) \\ &= E_O[E_O[Y(d)|S(d), D= d]|D\neq d]]P_O(D\neq d) \\ &= E_O[E_O[Y|S, D= d]|D\neq d]P_O(D\neq d) \\ &= yP_O(D\neq d) \end{aligned} \tag{127}$$

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

**Corollary 2.** *Let conditions of Theorem 2 hold. If $\mathcal{H}(m|\cdot)$ has minimal and maximal selectors with respect to $T$, then:*

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

*Proof of Corollary 2.* By observation, it is immediate that iterated and joint minima and maxima search over all values of the set $\mathcal{H}(m,\gamma)$ Thus:

$$\mathcal{H}(\tau) = \left[\min_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}), \max_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma})\right] \tag{128}$$

$$= \left[\min_{\tilde{\gamma}\in\mathcal{H}(\gamma)}\min_{\tilde{m}\in\mathcal{H}(m|\tilde{\gamma})} T(\tilde{m},\tilde{\gamma}), \max_{\tilde{\gamma}\in\mathcal{H}(\gamma)}\max_{\tilde{m}\in\mathcal{H}(m|\tilde{\gamma})} T(\tilde{m},\tilde{\gamma})\right]. \tag{129}$$

By definition of $L_\gamma$ and $U_\gamma$:

$$\forall \tilde{m} \in \mathcal{H}(m|\gamma): \;\; T(L_\gamma,\gamma) \leq T(\tilde{m},\gamma) \leq T(U_\gamma,\gamma). \tag{130}$$

Therefore:

$$\begin{aligned} \min_{\tilde{m}\in\mathcal{H}(m|\gamma)} T(\tilde{m},\gamma) &= T(L_\gamma,\gamma) \\ \max_{\tilde{m}\in\mathcal{H}(m|\gamma)} T(\tilde{m},\gamma) &= T(U_\gamma,\gamma). \end{aligned} \tag{131}$$

□

**Lemma 12.** *Let Assumptions RA, and EV hold. Suppose that $\mathcal{S}$ is a finite set.*

*i) Suppose that Assumption LIV holds and that $d_s = 1$. Then for $s \in \mathcal{S}$:*

$$
\begin{aligned}
L_\gamma(s) = \Bigg( & \min_{s' \geq s} E_O[Y|S = s', D = 0] \frac{P_O(S = s', D = 0)}{\gamma_0(s')} + 1 - \frac{P_O(S = s', D = 0)}{\gamma_0(s')}, \\
& \max_{s' \leq s} E_O[Y|S = s', D = 1] \frac{P_O(S = s', D = 1)}{\gamma_1(s')} \Bigg), \\
U_\gamma(s) = \Bigg( & \max_{s' \leq s} E_O[Y|S = s', D = 0] \frac{P_O(S = s', D = 0)}{\gamma_0(s')}, \\
& \min_{s' \geq s} E_O[Y|S = s', D = 1] \frac{P_O(S = s', D = 1)}{\gamma_1(s')} + 1 - \frac{P_O(S = s', D = 1)}{\gamma_1(s')} \Bigg),
\end{aligned}
$$

*ii) Suppose that Assumption TI holds. Then $L_{\gamma'} = (m^{TI,L,\gamma'}, m^{TI,L,\gamma'})$ and $U_{\gamma'} = (m^{TI,U,\gamma'}, m^{TI,U,\gamma'})$ for:*

$$
\begin{aligned}
m(s)^{TI,L,\gamma'} &= m(s)^{L,\gamma'} \mathbb{1}[\gamma_1'(s) \geq \gamma_0'(s)] + m(s)^{U,\gamma'} \mathbb{1}[\gamma_1'(s) < \gamma_0'(s)] \\
m(s)^{TI,U,\gamma'} &= m(s)^{L,\gamma'} \mathbb{1}[\gamma_1'(s) < \gamma_0'(s)] + m(s)^{U,\gamma'} \mathbb{1}[\gamma_1'(s) \geq \gamma_0'(s)],
\end{aligned} \tag{132}
$$

*where:*

$$
\begin{aligned}
m(s)^{L,\gamma'} &= \max_{d \in \{0,1\}} E_O[Y|S = s, D = d] \frac{P_O(S = s, D = d)}{\gamma_d'(s)}, \\
m(s)^{U,\gamma'} &= \min_{d \in \{0,1\}} E_O[Y|S = s, D = d] \frac{P_O(S = s, D = d)}{\gamma_d'(s)} + 1 - \frac{P_O(S = s, D = d)}{\gamma_d'(s)}.
\end{aligned} \tag{133}
$$

*Proof. i)*

Fix any $\gamma'$ such that there exists $(m, \gamma') \in \mathcal{H}(m, \gamma)$. Then $\mathcal{H}(m|\gamma') \neq \emptyset$. By bounded $\mathcal{Y}$, $h_{co(\mathcal{Y})}(-1) = 0$ and $h_{co(\mathcal{Y})}(1) = 1$. By Theorem 2, $\forall s \in \mathcal{S}$ restrictions imposed by data on $m_d(s)$ can then be equivalently stated for $d \in \{0, 1\}$ as:

$$
\begin{aligned}
m_d(s) \in \Bigg[ & E_O[Y|S = s, D = d] \frac{P_O(S = s, D = d)}{\gamma_d'(s)}, \\
& E_O[Y|S = s, D = d] \frac{P_O(S = s, D = d)}{\gamma_d'(s)} + 1 - \frac{P_O(S = s, D = d)}{\gamma_d'(s)} \Bigg].
\end{aligned} \tag{134}
$$

By Manski and Pepper (2000, Proposition 1) under Assumption LIV the sharp bound on $m_d(s)$

is:

$$
\begin{aligned}
m_d(s) \geq m_d(s)^{LIV,L,\gamma'} &:= \sup_{s' \leq s} E_O[Y|S=s', D=d] \frac{P_O(S=s', D=d)}{\gamma'_d(s')} \\
m_d(s) \leq m_d(s)^{LIV,U,\gamma'} &:= \inf_{s' \geq s} E_O[Y|S=s', D=d] \frac{P_O(S=s', D=d)}{\gamma'_d(s')} + 1 - \frac{P_O(S=s', D=d)}{\gamma'_d(s')}.
\end{aligned} \tag{135}
$$

First, note that both $m_d^{LIV,L,\gamma'}$ and $m_d^{LIV,U,\gamma'}$ are non-decreasing in $s$ by definition for all $d \in \{0,1\}$. Thus, $L_{\gamma'} := (m_0^{LIV,U,\gamma'}, m_1^{LIV,L,\gamma'}) \in \mathcal{M}^A$ and $U_{\gamma'} := (m_0^{LIV,L,\gamma'}, m_1^{LIV,U,\gamma'}) \in \mathcal{M}^A$. Hence $(L_{\gamma'}, \gamma'), (U_{\gamma'}, \gamma') \in \mathcal{H}(m, \gamma)$. Since $\gamma'$ was arbitrary, $(L_{\gamma'}, \gamma'), (U_{\gamma'}, \gamma') \in \mathcal{H}(m, \gamma)$ for any $\gamma' \in \mathcal{H}(\gamma)$. Therefore, $L_{\gamma'}$ and $U_{\gamma'}$ are selectors of $\mathcal{H}(m|\cdot)$. Then, observe that $T$ is non-decreasing in $m_1(s)$ and non-increasing in $m_0(s)$ for each $s \in \mathcal{S}$. Therefore, $\forall m \in \mathcal{H}(m|\gamma')\ T(L_{\gamma'}, \gamma') \leq T(m, \gamma')$, so $L_{\gamma'}$ is a minimal selector with respect to $T$. Similarly, $\forall m \in \mathcal{H}(m|\gamma')\ T(U_{\gamma'}, \gamma') \geq T(m, \gamma')$, so $U_{\gamma'}$ is a maximal selector with respect to $T$. Since $\mathcal{S}$ is a finite set, infima and suprema may be replaced by minima and maxima.

*ii)*

As in proof of *i)*, fix any $\gamma'$ such that there exists $(m, \gamma') \in \mathcal{H}(m, \gamma)$, so $\mathcal{H}(m|\gamma') \neq \emptyset$. Assumption TI maintains that $m_1 = m_0$. Then write for any $s \in \mathcal{S}$ and $d \in \{0,1\}$:

$$
T(m, \gamma') = \int_{\mathcal{S}} m_1(s) d\gamma'_1(s) - \int_{\mathcal{S}} m_0(s) d\gamma'_0(s) = \int_{\mathcal{S}} m_d(s)(d\gamma'_1(s) - d\gamma'_0(s)). \tag{136}
$$

Define:

$$
\begin{aligned}
m(s)^{L,\gamma'} &:= \max_{d \in \{0,1\}} E_O[Y|S=s, D=d] \frac{P_O(S=s, D=d)}{\gamma'_d(s)}, \\
m(s)^{U,\gamma'} &:= \min_{d \in \{0,1\}} E_O[Y|S=s, D=d] \frac{P_O(S=s, D=d)}{\gamma'_d(s)} + 1 - \frac{P_O(S=s, D=d)}{\gamma'_d(s)}.
\end{aligned} \tag{137}
$$

Next let for any $s \in \mathcal{S}$:

$$
\begin{aligned}
m(s)^{TI,L,\gamma'} &:= m(s)^{L,\gamma'} \mathbb{1}[\gamma'_1(s) \geq \gamma'_0(s)] + m(s)^{U,\gamma'} \mathbb{1}[\gamma'_1(s) < \gamma'_0(s)] \\
m(s)^{TI,U,\gamma'} &:= m(s)^{L,\gamma'} \mathbb{1}[\gamma'_1(s) < \gamma'_0(s)] + m(s)^{U,\gamma'} \mathbb{1}[\gamma'_1(s) \geq \gamma'_0(s)]
\end{aligned} \tag{138}
$$

and $L_{\gamma'} := (m^{TI,L,\gamma'}, m^{TI,L,\gamma'})$ and $U_{\gamma'} := (m^{TI,U,\gamma'}, m^{TI,U,\gamma'})$.

By Theorem 2, it is immediate that for $\mathcal{H}(m|\gamma') = \{m \in \mathcal{M} : m_1 = m_0,\ \forall d \in \{0,1\},\ \forall s \in \mathcal{S},\ m_d(s) \geq m(s)^{L,\gamma'},\ m_d(s) \leq m_d(s)^{U,\gamma'}\}$. Hence $(L_{\gamma'}, \gamma'), (U_{\gamma'}, \gamma') \in \mathcal{H}(m|\gamma')$. Since $\gamma'$ was arbitrary, $(L_{\gamma'}, \gamma'), (U_{\gamma'}, \gamma') \in \mathcal{H}(m|\gamma')$ for any $\gamma' \in \mathcal{H}(\gamma)$. Therefore, $L_{\gamma'}$ and $U_{\gamma'}$ are selectors of $\mathcal{H}(m|\cdot)$.

Then observe that by (136), $\forall m \in \mathcal{H}(m|\gamma')\ T(L_{\gamma'}, \gamma') \leq T(m, \gamma')$, so $L_{\gamma'}$ is a minimal selector with respect to $T$. Similarly, $\forall m \in \mathcal{H}(m|\gamma')\ T(U_{\gamma'}, \gamma') \geq T(m, \gamma')$, so $U_{\gamma'}$ is a maximal selector with respect to $T$. □