Causal Model of a Two-sided Matching Market: Application to Intergenerational Spillover Effects of Education*

Hideyuki Tomiyama
Department of Economics
The Pennsylvania State University
hkt5140@psu.edu

This version: November 16, 2026

Click here for the latest version

Abstract

This paper studies a heterogeneous treatment effect model with instrumental variables in an endogenous one-to-one two-sided matching market. Our leading example is the effect of parental education on children's outcomes, such as income, educational attainment, and health. Under assortative matching, it is difficult to justify the standard instrumental variables methods (e.g., two-stage least squares) that use treatments and instruments from both sides, even when treatment effects are homogeneous. To overcome this problem, we develop a selection model in which each agent selects both their own treatment status and that of their matched partner. We provide point and partial identification results for various causal parameters, including the average treatment effect. We also propose inference procedures for these causal parameters. Finally, we apply our framework to study the treatment effect of parents' college attendance on the educational investment they make in their child, using Japanese household-level data.

*First version: October 25, 2026. I am grateful to Marc Henry for his invaluable guidance and continuous support throughout this project. I am also grateful to Keisuke Hirano, Sung Jae Jun, Joris Pinkse, and Thomas Russell for their helpful discussions and suggestions. I also thank participants at the Penn State econometrics workshop and the Summer Workshop on Economic Theory for helpful comments and suggestions.

1 Introduction

Parental education is widely considered to affect children's outcomes such as educational attainment and health status (e.g., Black and Devereux, 2011; Björklund and Salvanes, 2011). A central concern in studying such treatment effects is endogeneity: unobserved heterogeneity that affect parents' schooling decisions are correlated with their children's outcomes. For instance, parents who care about their own education tend to care about various outcomes of their children as well.

To deal with such endogeneity problems, economists often use instrumental variables methods (e.g., Imbens and Angrist, 1994; Heckman and Vytlacil, 1999, 2001a, 2001b, 2005). In particular, Holmlund et al., 2011 provides a survey of instrumental variables approaches to identifying intergenerational spillover effects of education. For instance, a commonly used instrument for parental education is exposure to a school reform policy (e.g., Black et al., 2005; Oreopoulos et al., 2006; Lindeboom et al., 2009; Chou et al., 2010; McCrary and Royer, 2011; Holmlund et al., 2011; Chevalier et al., 2013; Lundborg et al., 2014). If one focuses specifically on college attendance as the measure of parental education, other possible instruments include the distance to college and local economic conditions in each parent's high-school years (e.g., Currie and Moretti, 2003; Carneiro et al., 2013).

Researchers often focus on only one parent, either the mother or the father. Specifically, these studies typically estimate the causal effect of one parent's education on the child's outcome, without accounting for the other parent. However, not only treatments but also matched pairs are non-randomly determined. Highly educated women tend to be matched with highly educated men and vice versa, which is known as assortative matching (see, for instance, Chiappori, 2020). Therefore, a large estimated effect of paternal education may simply reflect that highly educated men match with highly educated women, with the estimate capturing the mother's influence rather than the father's. Moreover, a person who values a partner's education may also value the child's outcome, further compounding the endogeneity.

Our goal is to develop an econometric framework that jointly considers the influence of both parents while allowing both treatment selection and matching to be endogenous. The first natural approach would be a standard instrumental variable method (e.g., two-stage least squares (TSLS)) that treats the two parents' education as the two endogenous variables with the two parents' instruments (e.g., Breierova and Duflo, 2004; Black et al., 2005; Oreopoulos et al., 2006; Lindeboom et al., 2009; Holmlund et al., 2011; Chevalier et al., 2013). This approach requires independence of instruments at the level of the household, namely that the pair of parental instruments is jointly

independent of both parents' unobserved characteristics (e.g., unobserved preference for education) under the distribution of matched couples.

However, the joint independence is difficult to justify in this setting. For instance, a woman who grows up near a college is more likely to attend college and, through assortative matching, to be paired with a college-graduate man. In turn, his unobserved preference for education tends to be strong because he has decided to attend college. This induces a correlation between the mother's instrument and the father's unobserved heterogeneity even if instruments are realized as if at random for each individual. Therefore, applying the standard instrumental variable methods without explicitly considering how women and men match each other is hard to justify.

To address this problem, we develop a heterogeneous treatment effect model with instruments in an endogenous one-to-one two-sided matching market. The treatment status of each parent is binary. We explicitly consider a selection model in which each person chooses both their own treatment (e.g., whether to attend college) and the treatment status of the partner they match (e.g., whether to match a college-graduate spouse or a non-college-graduate spouse). Here, each person chooses only the observable characteristics of their partner rather than a specific individual, as in many empirical marriage matching models (e.g., Choo and Siow, 2006; Chiappori et al., 2017; Galichon et al., 2019; Galichon and Salanié, 2022; Gualdani and Sinha, 2023).

The potential outcome of each household, namely the child's outcome, is indexed by the pair of parental treatment statuses. For each such pair, the outcome is determined by unobserved factors of both parents that affect the child. These outcome-relevant unobserved factors are correlated with the unobserved heterogeneity that drives each parent's two choices, namely their own education and the education of their partner. This correlation is the source of the endogeneity. Intuitively, a person who cares about their own education or that of their partner tends to care about the child's outcome as well.

To deal with this endogeneity, we introduce, for each individual, instruments that affect both choices (i.e., their own education and that of their partner) but do not directly affect the child's outcome. For example, as noted above, distance to college and local economic conditions shift a person's own education. These instruments also make the person more likely to match with a college-graduate spouse, because they raise the share of college graduates in the region. In contrast to standard instrumental variable methods, our model relies on the independence of instruments at the level of the individual rather than at the level of the household, meaning that instruments are

realized as if at random for each individual.

Our model can be viewed as a causal model with interference, in which interference occurs only between the two matched parents. A related literature studies such causal models of interference within a group or pair using instruments (e.g., Balat and Han, 2023; Hoshino and Yanagi, 2023; Vazquez-Bare, 2023; Acerenza et al., 2025). These approaches also rely on a joint independence condition, namely that the vector of instruments is jointly independent of the unobserved characteristics of all members of the group. They consider a setup in which the pair is formed before the instruments and treatments are realized, under which this joint independence can be justified. In our setting, by contrast, each individual's instrument is realized first, their education is determined by that instrument, and the matching is determined only afterwards. Because the matching is thus endogenous, the joint independence is difficult to justify in this setting.

Under our setup, we study identification of various causal parameters. Our key assumption, matching on observables, requires that who matches whom is as good as random conditional on treatment statuses and other observables. Intuitively, conditional on other observed characteristics including instruments and covariates, (i) college-graduate women who decide to match a non-college-graduate man and (ii) non-college-graduate men who decide to match a college-graduate woman randomly match each other. Thus, we allow for assortative matching on educational levels, namely that college graduates are likely to match each other and vice versa. However, we rule out further sorting on unobservable characteristics.

To convey the key features of the model, we begin with a benchmark under conditional ignorability of the matching response. Intuitively, a person who values their own education may also value the child's outcome, whereas, conditional on this preference over their own education, their preference over having a college-educated partner is independent of how much they value the child's outcome. Then, using a binary instrument on each side, we derive a point identification result on the local average treatment effect (LATE) among complier-complier pairs. Here, in the sense of Imbens and Angrist, 1994, a complier is a person whose educational status is switched as their instrument changes. The identifying formula is the two-sided counterpart of the Wald estimand familiar from the canonical binary-treatment, binary-instrument setup.

Next, we consider the general case in which conditional ignorability of the matching response fails. Intuitively, among people with the same preference over their own education, a person who prefers a college-educated partner may also value the child's outcome more. In this general setting, we pursue

partial identification of various causal parameters of interest (e.g., the average treatment effect (ATE) and the average treatment effect on the treated (ATT)). Each target is a functional of two unknown objects characterizing the economy of interest: (i) the household-level joint distribution of the treatments, instruments, and unobserved characteristics of both sides, and (ii) a function mapping unobserved factors into the potential outcome. We obtain sharp bounds by minimizing and maximizing the functional over the pairs of unknowns consistent with our model and the observed household-level data. However, because each such pair consists of a distribution and a function, this optimization is in principle infinite-dimensional.

We show that this infinite-dimensional problem can be transformed into a finite-dimensional one. The target functional and the constraints depend on these unknowns only through two finite-dimensional summaries: (i) for each side, the probability that an agent's treatment-matching response type—the pattern of how her own and partner treatments would respond to her instrument—takes each value, and (ii) for each pair of response types, the corresponding expected outcome. Then, the bound is obtained by optimizing the target over (i) and (ii), subject to the constraints that link these summaries to the distribution of observables.

To bring this problem into the form of a linear program, we can further impose a parametric restriction on the selection model, namely a specification of how each agent's own and partner treatments are determined from her instrument and her unobserved heterogeneity. Given this parameter, the bound is the optimal value of a linear program in the expected outcomes for each pair of response types, as in Balke and Pearl, 1997, Mogstad et al., 2018, Lafférs, 2019, Gu and Russell, 2023, Kamat et al., 2023, Bai et al., 2024, Han and Yang, 2024, Ura and Zhang, 2024, Acerenza et al., 2025, among others.

We also propose inference procedures for the bounds of these causal parameters under the parametric selection model. A first stage estimates the selection-model parameters and the relevant moments of the observed variables. Given these first-stage estimates, the bound is the optimal value of a linear program, and a second stage conducts inference on this value. However, because the linear program need not have unique primal or dual solutions, the standard delta method does not apply to its value. Following Gafarov, 2025, we regularize the objective function with a quadratic penalty to obtain confidence sets that are uniformly valid over a class of data-generating distributions.

The rest of the paper is organized as follows. Using an illustrative example, Section 2 discusses the problem of the standard instrumental variable method that treats the two parents' education as

the two endogenous variables with the two parents' instruments. Section 3 presents the setup of the model. In Section 4, we provide the identification result on LATE among complier-complier pairs under the conditional ignorability of the matching response. Section 5 considers partial identification of various causal parameters (e.g., ATE) without the conditional ignorability of the matching response. In Section 6, we propose inference procedures. Section 7 applies our framework to estimate the causal effect of parental college attendance on the educational investment that couples make in their children, using Japanese data.

Notation

For any distribution F over a random vector $B = (B_1, B_2)$, we denote by F_{B_1} the marginal distribution of B_1, and by $F_{B_1|B_2=b_2}$ the conditional distribution of B_1 given $B_2 = b_2$, both under F. We use $X \perp\!\!\!\perp Y$ to denote that random variables X and Y are statistically independent. We write $X \sim F$ to mean that the random variable X is distributed according to the distribution F. We use $\mathbb{P}_F$ and $\mathbb{E}_F$ to denote the probability and expectation with respect to F, respectively.

2 Motivation

2.1 Standard instrumental variable methods

When we consider causal effects of both parents' education simultaneously, the first natural approach would be to use standard instrumental variable methods that treat the two parents' education (D^w, D^m) as the two endogenous variables with the two parents' instruments (Z^w, Z^m). Here, D^w and D^m denote the mother's and the father's education, and Z^w and Z^m their respective instruments (e.g., the distance to college at the age of seventeen). This section points out the problem of such standard instrumental variable methods.

In particular, the simplest one is the TSLS with the endogenous variables (D^w, D^m) and the instruments (Z^w, Z^m). Indeed, several papers in the literature conduct such TSLS (e.g., Breierova and Duflo, 2004; Black et al., 2005; Oreopoulos et al., 2006; Lindeboom et al., 2009; Holmlund et al., 2011; Chevalier et al., 2013). As a stylized example, we consider the following regression model:

$$Y = \beta_0 + \beta_1 D^w + \beta_2 D^m + \zeta, \qquad \mathbb{E}[\zeta \mid Z^w, Z^m] = 0, \tag{2.1}$$

where Y is the child outcome. In this model, the error term ζ stands for the unobserved heterogeneity

of both parents, which affects both their own education and their child's outcome. Therefore, the moment condition $\mathbb{E}[\zeta \mid Z^w, Z^m] = 0$ intuitively requires the joint independence of the instruments, that is, the pair of instruments (Z^w, Z^m) is jointly independent of the unobserved heterogeneity of both the mother and the father under the distribution of households (i.e., matched pairs).

When the researcher is interested in heterogeneity of treatment effects, a causal interpretation of such TSLS becomes difficult, because the household has the two treatments D^w and D^m. In that case, another possible strategy is to adopt the approach of the literature on causal models for multiple or multi-valued treatments with instruments (e.g., Heckman et al., 2008; Heckman and Pinto, 2018; Lee and Salanié, 2018, 2020; Mountjoy, 2022; Tsuda, 2022; Kamat et al., 2023; Bai et al., 2024; Navjeevan et al., 2026), which regards the combination (D^w, D^m) as the treatment status of each household.

We now assume that the treatment of each parent is binary. We also specify a selection model in which each woman chooses her education according to $D^w = \mathbb{1}(\varepsilon_1^w \leq h_1^w(Z^w))$ and each man analogously. Let $Y(d, d')$ denote the potential outcome of a child when the mother's and the father's education are d and d', respectively. Here, ε_1^w captures the woman's unobserved preference for education, which affects her own education and is correlated with her child's outcomes $Y(\cdot, \cdot)$. As with the TSLS, this approach requires that the pair of instruments (Z^w, Z^m) be jointly independent of the pair of unobserved preferences for education $(\varepsilon_1^w, \varepsilon_1^m)$ under the distribution of households.

Then, under the joint independence $(Z^w, Z^m) \perp\!\!\!\perp (\varepsilon_1^w, \varepsilon_1^m, Y(\cdot, \cdot))$, the papers in this literature provide point and partial identification results for various causal parameters, such as the ATE (i.e., $\mathbb{E}[Y(d, d') - Y(0, 0)]$). In particular, with binary instruments, the LATE among complier-complier parents is point-identified. To describe it, we assume threshold monotonicity, $h_1^s(0) < h_1^s(1)$ for each $s \in \{w, m\}$, under which each parent is an always-taker $(\varepsilon_1^s \leq h_1^s(0))$, a complier $(h_1^s(0) < \varepsilon_1^s \leq h_1^s(1))$, or a never-taker $(\varepsilon_1^s > h_1^s(1))$. The LATE among complier-complier parents is defined as

$$LATE_{JI}(d, d') := \mathbb{E}\big[\, Y(d, d') - Y(0, 0) \;\big|\; h_1^w(0) < \varepsilon_1^w \leq h_1^w(1),\ h_1^m(0) < \varepsilon_1^m \leq h_1^m(1) \,\big].$$

Then we can show that the maternal effect among complier-complier parents can be identified as

$$LATE_{JI}(1, 0) = \frac{\displaystyle\sum_{(z,z') \in \{0,1\}^2} (-1)^{z+z'+1}\, \mathbb{E}\big[\, Y\, \mathbb{1}\{D^m = 0\} \mid Z^w = z,\ Z^m = z' \,\big]}{\displaystyle\sum_{(z,z') \in \{0,1\}^2} (-1)^{z+z'}\, \mathbb{P}\big(D^w = 1,\ D^m = 1 \mid Z^w = z,\ Z^m = z'\big)}.$$

The paternal effect $LATE_{JI}(0,1)$ can be identified analogously. We provide the derivation in Appendix G.1.

However, the joint independence is difficult to justify in this setting. Such an instrument (e.g., the distance to college at the age of seventeen) is realized to each individual rather than to each household, and each individual chooses her own education given her own instrument. The household is formed only afterwards, through the matching market, which is what determines the joint distribution of $(Z^w, D^w, \varepsilon_1^w)$ and $(Z^m, D^m, \varepsilon_1^m)$. Therefore, the joint independence is a restriction on the endogenous matching outcome rather than on the primitives of the model.

Moreover, we can expect a systematic correlation between the mother's instrument and the father's unobserved heterogeneity. For instance, a woman who grows up near a college is more likely to attend college and, through assortative matching, to be matched with a college-graduate man. Because his education is determined based on his unobserved heterogeneity ε_1^m, the endogenous matching process makes her instrument Z^w correlated with ε_1^m under the distribution of households. Therefore, the joint independence may not hold even though the instrument is first assigned to each individual independently of their unobserved preference for education.

2.2 Illustrative example

Beyond the intuition above, we next show that the joint independence often fails to hold in a model of endogenous treatment selection and matching process. First, as in the selection model above, each individual selects their own education, with $D^w = \mathbb{1}(\varepsilon_1^w \leq h_1^w(Z^w))$ for each woman and analogously for each man. Then, each individual chooses the education of their matched partner (e.g., whether a woman matches a college-graduate spouse or a non-college-graduate spouse) according to $T^w = \mathbb{1}(\varepsilon_2^w \leq h_2^w(D^w))$ for each woman and analogously for each man. Here, each individual chooses only the educational type of their partner rather than a specific person in the opposite side.

In particular, we consider the following setup:

(i) $\mathbb{P}(Z^s = 0) = \mathbb{P}(Z^s = 1) = 0.5$ for each $s \in \{w, m\}$;

(ii) ε_1^s and ε_2^s are independent and uniform on $[0,1]$ for each $s \in \{w, m\}$;

(iii) $(h_1^w(0), h_1^w(1)) = (0.2, 0.8)$ and $(h_1^m(0), h_1^m(1)) = (0.4, 0.6)$;

(iv) $(h_2^s(0), h_2^s(1)) = (1-q, q)$ for each $s \in \{w, m\}$, where $q \in [0.5, 1]$,

where the parameter q measures assortativeness, ranging from random matching at $q = 0.5$ to perfectly assortative matching at $q = 1$. Under this parametrization, the market clears for every q, that is, $\mathbb{P}(D^w = d, T^w = d') = \mathbb{P}(D^m = d', T^m = d)$ for every d, d'.

However, the joint independence generically cannot hold under this setting. More formally, we can show that, under assortative matching $q > 0.5$, there does not exist any joint distribution of $(Z^w, D^w, T^w, \varepsilon_1^w, \varepsilon_2^w)$ and $(Z^m, D^m, T^m, \varepsilon_1^m, \varepsilon_2^m)$ such that, in addition to conditions (i)–(iv) above, the following conditions (v)–(vii) hold simultaneously:

(v) $D^s = \mathbb{1}(\varepsilon_1^s \leq h_1^s(Z^s))$ and $T^s = \mathbb{1}(\varepsilon_2^s \leq h_2^s(D^s))$ almost surely for each $s \in \{w, m\}$;

(vi) $D^w = T^m$ and $D^m = T^w$ almost surely;

(vii) the joint independence $(Z^w, Z^m) \perp\!\!\!\perp (\varepsilon_1^w, \varepsilon_1^m)$.

Here, conditions (v)–(vi) require that demand and supply coincide in the matching market. Intuitively, the nonexistence result holds since conditions (v)–(vi) impose

$$\begin{aligned}\mathbb{1}\{\varepsilon_1^w \leq h_1^w(Z^w)\} &= \mathbb{1}\{\varepsilon_2^m \leq h_2^m(D^m)\} \\ &= \mathbb{1}\{\varepsilon_2^m \leq h_2^m(\mathbb{1}\{\varepsilon_1^m \leq h_1^m(Z^m)\})\} \text{ almost surely,}\end{aligned}$$

which induces a correlation between Z^w and ε_1^m even though the instruments are first assigned to each individual at random (i.e., $Z^w \perp\!\!\!\perp \varepsilon_1^w$ and $Z^m \perp\!\!\!\perp \varepsilon_1^m$). This illustrative example suggests that the joint independence is hard to justify in this setting, where the matching is endogenous and sorts individuals on the choices that the instruments shift. Indeed, this difficulty is not specific to this parametrization. We discuss the details and the derivation for the present example in Appendix B.

Finally, we show the bias of the standard instrumental variable methods through this illustrative example. To this end, we now construct a joint distribution of $(Z^w, D^w, T^w, \varepsilon_1^w, \varepsilon_2^w)$ and $(Z^m, D^m, T^m, \varepsilon_1^m, \varepsilon_2^m)$ so that, in addition to conditions (i)–(vi) above, the following conditions (viii)–(ix) are satisfied:

(viii) the marginal independence $Z^w \perp\!\!\!\perp \varepsilon_1^w$ and $Z^m \perp\!\!\!\perp \varepsilon_1^m$;

(ix) women who choose $(D^w, T^w) = (d, d')$ randomly match men who choose $(D^m, T^m) = (d', d)$ for each $d, d' \in \{0, 1\}$.

We then specify the outcome based on the regression model (2.1) with $(\beta_0, \beta_1, \beta_2) = (-0.5, 0.25, 0.25)$. We decompose the error term as $\zeta = \zeta^w + \zeta^m$ with $\zeta^s = 0.25 \cdot \mathbb{1}\{\varepsilon_1^s \leq h_1^s(0)\} - 0.25 \cdot \mathbb{1}\{\varepsilon_1^s > h_1^s(1)\}$

for each $s \in \{w, m\}$. Because ε_1^s is uniform and marginally independent of Z^s, $\mathbb{E}[\zeta^s \mid Z^s] = 0$ for each $s \in \{w, m\}$. In this way, the construction above determines the distribution of the observed variables (Z^w, D^w, Z^m, D^m, Y).

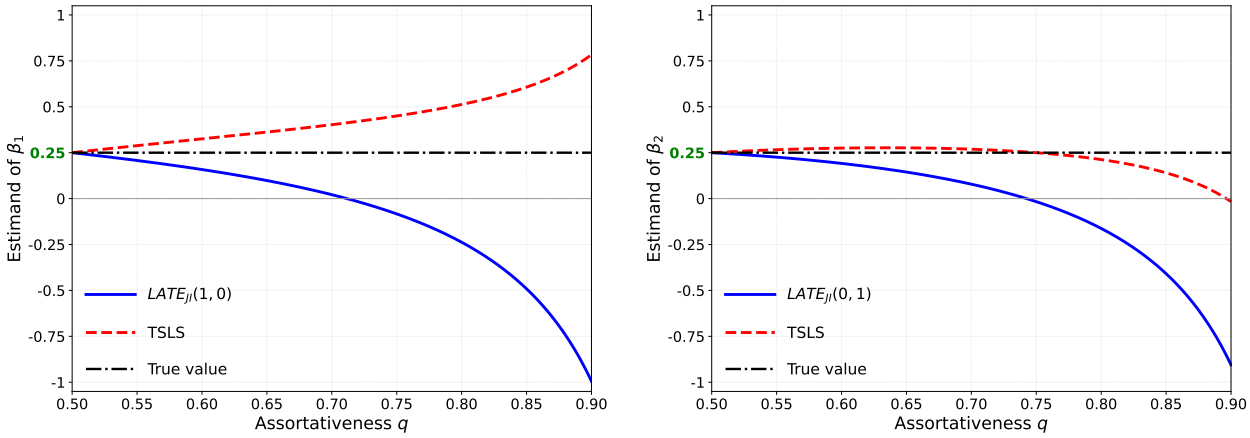
The objects of interest are the maternal effect β_1 and the paternal effect β_2, which are homogeneous under this setup. We try to recover the effects by the following two standard instrumental variable methods. First, we consider the coefficients $(\beta_1^{tsls}, \beta_2^{tsls})$ of the TSLS with treatments and instruments from both sides:

$$(\beta_0^{tsls}, \beta_1^{tsls}, \beta_2^{tsls}) := \mathbb{E}[ZD']^{-1}\mathbb{E}[ZY], \qquad Z := (1, Z^m, Z^w)', \quad D := (1, D^w, D^m)'. \tag{2.2}$$

The TSLS relies on the joint independence $(Z^w, Z^m) \perp\!\!\!\perp (\zeta^w, \zeta^m)$, which gives the moment condition $\mathbb{E}[\zeta^w + \zeta^m \mid Z^w, Z^m] = 0$. Second, we consider $LATE_{JI}(1,0)$ and $LATE_{JI}(0,1)$ defined above, which relies on the joint independence as well.

Figure 1 plots the TSLS and $LATE_{JI}$ estimands against the assortativeness q. At $q = 0.5$, the matching is random and both estimands recover the true effect, because, under random matching, the marginal independence implies the joint independence. At $q > 0.5$, the matching is assortative and both estimands become biased, because the joint independence cannot be satisfied in this setting. Therefore, these standard instrumental variable methods are unreliable under assortative matching.

Figure 1: Bias of TSLS and $LATE_{JI}$ under assortative matching



(a) The maternal effect β_1 (b) The paternal effect β_2

Notes: The TSLS and $LATE_{JI}$ estimands as the assortativeness q varies, for the maternal effect β_1 (panel (a)) and the paternal effect β_2 (panel (b)). The true value is 0.25.

3 Our framework

In Section 2, we showed that applying the standard instrumental variable methods without explicitly modeling how women and men match requires the joint independence of the instruments, which cannot be justified when the matching is assortative. To deal with this problem, we develop a heterogeneous treatment effect model that explicitly considers the endogenous treatment selection and matching process.

3.1 Setup

We consider a continuum of agents. Let $D_i^w \in \{0,1\} =: \mathcal{D}$ denote the treatment status of woman i, let $D_j^m \in \mathcal{D}$ denote the treatment status of man j, and let $Y_{ij} \in \mathcal{Y}$ denote the outcome of the pair (i,j). As the leading example throughout the paper, D_i^w and D_j^m indicate whether woman i and man j attended college, and Y_{ij} is a quantity concerning the child born to the household formed by woman i and man j: for instance, the child's educational attainment, income, or health, or educational investment in them. Throughout the main text, we abstract from observed covariates for simplicity, and develop the model with covariates in Appendix F.

We assume that each woman i selects both her own treatment status $D_i^w \in \mathcal{D}$ (i.e., whether to attend college or not) and that of her matched partner $T_i^w \in \mathcal{D}$ (i.e., whether to match with a college-graduate man or a non-college-graduate man). Similarly, each man j selects $(D_j^m, T_j^m) \in \mathcal{D}^2$. Thus, each person chooses only the treatment status of whom she or he matches instead of choosing a specific individual. This simplification, in which agents are modeled as selecting on the observed characteristics of their match rather than on specific individuals, is common in the literature on empirical matching models (e.g., Choo and Siow, 2006; Chiappori et al., 2017; Chiappori et al., 2018; Galichon et al., 2019; Galichon and Salanié, 2022; Gualdani and Sinha, 2023). We restrict attention to the setup in which every person is matched with exactly one partner on the opposite side, ruling out the possibility of remaining single.

To address the endogeneity, we introduce an instrumental variable Z_i^w, which affects woman i's treatment selection and matching (D_i^w, T_i^w) but does not directly affect her outcomes Y_{ij} for every $j \in \mathcal{J}$. Let $\mathcal{Z}$ be the finite support of Z_i^w. We use $D_i^w(z)$ and $T_i^w(z)$ to denote, respectively, woman i's choice of her own treatment status and her choice of her partner's treatment status when her instrument takes the value $z \in \mathcal{Z}$. Let $K_i^w = \{(D_i^w(z), T_i^w(z))\}_{z \in \mathcal{Z}}$ be the treatment-matching response type of woman i. When no confusion arises, we refer to it as the response type, or simply

as the type. We similarly introduce Z_j^m, $D_j^m(\cdot)$, $T_j^m(\cdot)$, K_j^m for man j. Let $\mathcal{K} \subseteq (\mathcal{D}^2)^{|\mathcal{Z}|}$ be the space of K^w and K^m.

In the leading example of parental college attendance and child outcomes, candidate instruments for each person include geographic proximity to a college, local college tuition, local economic conditions, and exposure to a school reform policy, each measured at the age of seventeen. These are classical instruments in the literature on individual and intergenerational returns to education (e.g., Card, 1993; Cameron and Heckman, 1998; Currie and Moretti, 2003; Black et al., 2005; Carneiro et al., 2011). Such instruments shift a person's own college attendance and also affect the education of their partner. For example, the distance to college and exposure to a school reform policy affect the supply of college graduates in the local marriage matching market. Hence, even among people with the same education, those in a marriage market with more college graduates are more likely to match with a college-graduate spouse.

Each woman i is endowed with $\boldsymbol{\phi}_i^w := \{\phi_i^w(d,d')\}_{(d,d')\in\mathcal{D}^2}$, where $\phi_i^w(d,d') \in \mathbb{R}$ is her potential factor when her own treatment status and that of whom she matches are d and d', respectively. Similarly, each man j is endowed with $\boldsymbol{\phi}_j^m := \{\phi_j^m(d,d')\}_{(d,d')\in\mathcal{D}^2}$, where $\phi_j^m(d,d') \in \mathbb{R}$ is his potential factor when his own treatment status and that of whom he matches are d' and d, respectively. Let $\Phi := \mathbb{R}^{|\mathcal{D}^2|}$ be the space of $\boldsymbol{\phi}_i^w$ and $\boldsymbol{\phi}_j^m$.

When woman i and man j match, the potential outcome of the household when their respective treatment statuses are d and d' is given by

$$Y_{ij}(d,d') = \pi\big(\phi_i^w(d,d'), \phi_j^m(d,d')\big) \in \mathcal{Y}, \tag{3.1}$$

where π is an unknown Borel-measurable function that maps potential factors of the two sides into the outcome. Let Π be the set of all Borel-measurable functions from $\mathbb{R}^2$ to $\mathcal{Y}$.[1] In the leading example, $\phi_i^w(1,0)$ is woman i's potential factor that contributes to her child's outcome when she attended college and her matched partner did not. The other values $\phi_i^w(d,d')$ and $\phi_j^m(d,d')$ are interpreted in the same way. When woman i attended college and the man j did not, their child outcome $Y_{ij}(1,0)$ is determined by π evaluated at woman i's potential factor $\phi_i^w(1,0)$ and man j's potential factor $\phi_j^m(1,0)$.

[1] At this stage, the factorization is trivially satisfied. That is, for any $\{\{Y_{ij}(d,d')\}_{(d,d')\in\mathcal{D}^2}\}_{(i,j)\in\mathcal{I}\times\mathcal{J}}$, there exists some $\pi \in \Pi$ and $\{\boldsymbol{\phi}_i^w\}_{i\in\mathcal{I}}, \{\boldsymbol{\phi}_j^m\}_{j\in\mathcal{J}}$ such that (3.1) holds for every $(i,j) \in \mathcal{I}\times\mathcal{J}$ and $(d,d') \in \mathcal{D}^2$ as long as the index spaces $\mathcal{I}$ and $\mathcal{J}$ are standard Borel spaces, which holds in all standard settings (e.g., Polish spaces such as $[0,1]$). Combined with the instrument independence condition introduced below (Assumption 3.2), however, the factorization becomes a nontrivial restriction on the data-generating process.

Note that we implicitly impose an exclusion restriction in (3.1). In other words, conditional on the treatment status of both parents, the instrument has no direct effect on the outcome. This is a conventional assumption in causal models with instruments. On top of that, the framework can be extended to accommodate exogenous pair-specific shocks and multidimensional potential factors. We spell out the pair-specific shock in Remark 3.1 below.

A key challenge in this setting is that we observe outcomes only at realized matchings and treatments, while many counterfactual outcomes remain unobserved. For each observed household formed by woman i and man j, we observe the outcome $Y_{ij}(D_i^w, D_j^m)$. However, the outcomes under counterfactual partners, or under counterfactual treatments, can not be observed. Because both the matching and the treatments are chosen by the agents themselves, we must deal with two endogenous selection problems.

Furthermore, assortative matching is a salient feature of marriage markets. College graduates tend to be matched with college graduates, with strong correlations between spouses' educational attainment observed in the data (e.g., Chiappori, 2020). This correlation makes it difficult to disentangle the direct effect of women's or men's treatment from the indirect effect operating through the spouse's treatment.

Although we use parental college attendance and child outcomes as the leading example throughout the paper, the framework applies to a range of related settings. We introduce several such applications below.

Example 3.1 (Peer effects between spouses)**.** In this example, we consider outcomes that pertain to the spouses themselves rather than their children. The treatment D^s again represents educational attainment, and the instrument Z^s can be similar to those discussed above (e.g., geographic proximity to a college or exposure to a school reform policy). The outcome Y may include household income, labor supply decisions, health behaviors, or life expectancy. Additionally, Y could represent fertility outcomes such as the number of children or the timing of childbirth. For instance, there is substantial evidence that college graduates live longer than non-college graduates, and part of this may come from their spouse. Understanding peer effects between spouses is important for evaluating the full returns to education. □

Example 3.2 (Intergenerational spillover effects of smoking)**.** The second example is the causal effects of parental smoking behavior on child outcomes. The treatment D^s indicates whether the parent is a smoker. Potential instruments include cigarette taxes or anti-smoking campaign exposure,

which affect the propensity to smoke but do not directly affect child outcomes. Relevant outcomes Y include birth weight, (non)cognitive development, and the child's own smoking behavior. Similar to the education example, smokers and non-smokers tend to match with each other in marriage markets, complicating the identification of parental smoking effects (e.g., Chiappori et al., 2024). □

3.2 The economy of interest and the main assumptions

In the following, we suppress the individual indices i and j to simplify the notation. Each woman and man is characterized by a vector

$$A^w = (Z^w, D^w, T^w, K^w, \boldsymbol{\phi}^w) \in \mathcal{A}$$
$$A^m = (Z^m, D^m, T^m, K^m, \boldsymbol{\phi}^m) \in \mathcal{A},$$

respectively, where $\mathcal{A} := \mathcal{Z} \times \mathcal{D} \times \mathcal{D} \times \mathcal{K} \times \Phi$. Each household is characterized by $A = (A^w, A^m) \in \mathcal{A}^2$. The distribution of household-level variables A is described by $F \in \mathcal{F}$, where $\mathcal{F}$ denotes the set of all distributions on $\mathcal{A}^2$. The distribution F and the function π together characterize the economy of interest.

The unit of observation is a household. For each household, we observe a random vector

$$(Z^w, D^w, Z^m, D^m, Y) \sim G,$$

where Y denotes the outcome of the household and the remaining variables are interpreted analogously. In what follows, I use $\mathbb{E}[\cdot]$ and $\mathbb{P}(\cdot)$ to denote $\mathbb{E}_G[\cdot]$ and $\mathbb{P}_G(\cdot)$, respectively.

Now we impose several assumptions on (F, π) and G. First, we assume the coherency between them.

Assumption 3.1 (Coherency between G and (F, π))**.** *Suppose that*

$$(Z^w, D^w, T^w, K^w, \boldsymbol{\phi}^w,\ Z^m, D^m, T^m, K^m, \boldsymbol{\phi}^m) \sim F.$$

Then

$$(Z^w, D^w, Z^m, D^m, Y) \sim G,$$

where

$$Y = \pi\big(\phi^w(D^w, D^m),\ \phi^m(D^w, D^m)\big).$$

Assumption 3.1 requires the observed data to be generated by the economy of interest. Tha is, the distribution G is induced by (F, π) on the observed coordinates (Z^w, D^w, Z^m, D^m, Y), with Y realized through π at the realized treatments (D^w, D^m).

Next, we assume the independence of instruments as follows.

Assumption 3.2 (Independence of instruments)**.** $(K^w, \boldsymbol{\phi}^w) \perp\!\!\!\perp Z^w$ *and* $(K^m, \boldsymbol{\phi}^m) \perp\!\!\!\perp Z^m$ *under* F*.*

Assumption 3.2 requires that, under F of interest, for each side of the model, the instruments are independent of the treatment-matching response and potential factor. In the leading example, this requires that a woman's geographic proximity to a college at age seventeen be independent of (i) how much she values attending college herself and how strongly she insists on a college-educated spouse, as well as of (ii) the potential factors that generate her child's outcome with her partner. The same requirement applies to men. This individual-level marginal independence is weaker than the household-level joint independence that the standard instrumental variable approach requires, which is harder to justify.

Next, we define the concept of matching as follows.

Definition 3.1 (Matching)**.** *A distribution* F *on* $\mathcal{A}^2$ *is a matching if*

$$(D^w, D^m) = (T^m, T^w)$$

holds F*-almost surely.*

In words, in a matching, the realized treatment status of each person must coincide with what their partner chose. If a wife is a college graduate, then her husband chose a college graduate wife. If a husband is a non-college graduate, then his wife chose a non-college graduate husband.

Then we define the concept of equilibrium matching as follows.

Definition 3.2 (Equilibrium matching)**.** *A matching* F *is an equilibrium matching if*

$$(D^w, T^w) = (D^w(Z^w), T^w(Z^w))$$
$$(D^m, T^m) = (D^m(Z^m), T^m(Z^m))$$

hold F*-almost surely.*

Definition 3.2 requires the realized choice to be consistent with the response type for each person. The pair (D^s, T^s) is the choice of her or his own treatment status and the treatment status of her or

his partner, whereas $D^s(\cdot)$ and $T^s(\cdot)$, the components of the response type K^s, give her or his choice of these treatment statuses when the instrument takes a given value. In an equilibrium matching, the two coincide at the realized instrument Z^s, that is, $(D^s, T^s) = (D^s(Z^s), T^s(Z^s))$. Here we can interpret that $D^s(\cdot)$ and $T^s(\cdot)$ are not primitives but equilibrium objects, determined endogenously so that demand and supply in the marriage market coincide. For instance, such equilibrium can be derived from empirical matching models (e.g., Galichon and Salanié, 2022). We discuss more details in Appendix D.1 and D.2.

Next, we define the concept of matching on observables as follows.

Definition 3.3. *A matching F satisfies matching on observables if*

$$F_{A|D^w=d,D^m=d',Z^w=z,Z^m=z'} = F_{A^w|D^w=d,T^w=d',Z^w=z} \times F_{A^m|D^m=d',T^m=d,Z^m=z'}$$

for every $d, d' \in \mathcal{D}$, $z, z' \in \mathcal{Z}$, where the equality is in the sense of distribution.

In words, F is a matching on observables if the conditional distribution $F_{A|D^w=d,D^m=d',Z^w=z,Z^m=z'}$ equals the distribution obtained by randomly matching women with $Z^w = z$ who choose $(D^w, T^w) = (d, d')$ to men with $Z^m = z'$ who choose $(D^m, T^m) = (d', d)$. For instance, consider the households in which the woman attended college and the man did not, with the woman's instrument equal to z and the man's equal to z'. Their distribution $F_{A|D^w=1,D^m=0,Z^w=z,Z^m=z'}$ is obtained by randomly matching: college-educated women with $Z^w = z$ who match a non-college man, that is, $(D^w, T^w) = (1, 0)$; and non-college men with $Z^m = z'$ who match a college-educated woman, that is, $(D^m, T^m) = (0, 1)$. This means that assortative matching on the observable characteristics (e.g., educational attainment) is allowed, but the matching does not further depend on unobservables (e.g., response types and potential factors) once we condition on the observable characteristics.

It is convenient to restate the matching on observables condition using the independent coupling of the two sides. To this end, let F^w and F^m denote the marginal distributions of A^w and A^m under F, and let F^* be their product $F^w \times F^m$. In words, F^* is the independent coupling of A^w and A^m, that is, the distribution we would obtain if women and men matched at random. We then have the following lemma.

Lemma 3.1. *A matching F satisfies matching on observables if and only if, for every $d, d' \in \mathcal{D}$, $z, z' \in \mathcal{Z}$,*

$$F_{A|D^w=d,D^m=d',Z^w=z,Z^m=z'} = F^*_{A|D^w=d,T^w=d',Z^w=z,D^m=d',T^m=d,Z^m=z'}.$$

Lemma 3.1 gives an equivalent restatement of the condition in Definition 3.3. F satisfies matching on observables if and only if, conditional on the treatment statuses and instruments, the couple (A^w, A^m) has the same distribution under the actual matching F as under the independent coupling F^*. Compared with the original definition, the right-hand side simply reverses the order of two operations: it first matches women and men at random (forming F^*) and then conditions on each side's status, a woman with $(D^w, T^w, Z^w) = (d, d', z)$ and a man with $(D^m, T^m, Z^m) = (d', d, z')$. This order does not affect the resulting distribution, and hence the two conditions are equivalent, which enables us to replace F with F^* in the identification arguments below.

Finally, we make an assumption that the distribution F of interest is an equilibrium matching and satisfies the matching on observables condition.

Assumption 3.3. *F is an equilibrium matching that satisfies matching on observables.*

The model in this paper involves several objects describing households, the underlying economy, and the observed data. We collect the corresponding notation in Table 1 for reference.

Table 1: Notation of the model

Symbol	Description
$s \in \{w, m\}$	Side of the household: woman (w) or man (m)
D^s	Own binary treatment status of a side s agent
T^s	Binary treatment status of whom a side s agent matches
Y	Outcome of the household
Z^s	Instrument of a side s agent
$D^s(z)$	Treatment response of a side s agent when $Z^s = z$
$T^s(z)$	Matching response of a side s agent when $Z^s = z$
$K^s = \{(D^s(z), T^s(z))\}_{z \in \mathcal{Z}}$	Treatment-matching response type of a side s agent
$\boldsymbol{\phi}^s = \{\phi^s(d, d')\}_{(d,d') \in \mathcal{D}^2}$	Potential factors of a side s agent
$Y(d, d')$	Potential outcome of the household when $(D^w, D^m) = (d, d')$
π	Outcome function: $Y(d, d') = \pi(\phi^w(d, d'), \phi^m(d, d'))$
$A^s = (Z^s, D^s, T^s, K^s, \boldsymbol{\phi}^s)$	Variables characterizing a side-s agent
$A = (A^w, A^m)$	Variables characterizing a household
F	Distribution of A
F^w, F^m	Marginal distributions of A^w and A^m under F
$F^* = F^w \times F^m$	Independent coupling of A^w and A^m
G	Distribution of the observed variables (Z^w, D^w, Z^m, D^m, Y)

Remark 3.1 (Exogenous pair-specific shocks)**.** The factorization in (3.1) extends to an exogenous pair-specific shock ϕ^c_{ij}. In the leading example, this represents a child-specific shock ϕ^c_{ij} that contributes to the child's outcome in addition to the two parents' potential factors. We write the

potential outcome as

$$Y_{ij}(d, d') = \pi\big(\phi_i^w(d, d'),\, \phi_j^m(d, d'),\, \phi_{ij}^c\big),$$

where ϕ_{ij}^c is drawn independently of $A = (A^w, A^m)$ and π is Borel measurable. Let F_ϕ^c denote the distribution of ϕ_{ij}^c. Assumption 3.1 is replaced by its counterpart under this setting, that is, if

$$(Z^w, D^w, T^w, K^w, \boldsymbol{\phi}^w,\, Z^m, D^m, T^m, K^m, \boldsymbol{\phi}^m,\, \phi^c) \sim \tilde{F} := F \times F_\phi^c,$$

then $(Z^w, D^w, Z^m, D^m, Y) \sim G$ with $Y = \pi\big(\phi^w(D^w, D^m),\, \phi^m(D^w, D^m),\, \phi^c\big)$. We also define $\tilde{F}^* := F^* \times F_\phi^c$ as the corresponding independent coupling, where $F^* = F^w \times F^m$. All of the identification results below remain valid under this extension, with F and F^* read as $\tilde{F}$ and $\tilde{F}^*$, respectively. We further discuss this extension in Appendix C. □

4 Conditional ignorability of the matching response

To convey the key feature of the framework, this section studies a simple benchmark under conditional ignorability of the matching response. Under this setup, we consider the identification of LATE among complier-complier pairs with binary instruments from both sides. Then we show a numerical illustration for the comparison between our estimand and conventional approaches.

First, we make the following assumption.

Assumption 4.1 (Conditional ignorability of the matching response)**.** *For each $s \in \{w, m\}$, it holds that $T^s(\cdot) \perp\!\!\!\perp \boldsymbol{\phi}^s$ conditional on $D^s(\cdot)$ under F.*

Assumption 4.1 requires that, conditional on a person's treatment response, their matching response is independent of the potential factors affecting the child's outcome. Intuitively, in the college attendance example, a mother's preference over her own education may be related to how much she values the child's outcome, while, given this preference, her preference over the partner's education is independent of how much she values the child's outcome. Note that, this assumption does not rule out assortative matching on educational attainment since we can still allow that an agent with $D^s = 1$ is more likely to choose $T^s = 1$.

Next, we make the following assumption on the instruments.

Assumption 4.2 (Relevance of binary instruments and overlap condition)**.** *The instruments are binary and relevant, that is, $\mathcal{Z} = \{0, 1\}$, and $p^s(1) > p^s(0)$, where $p^s(z) := \mathbb{P}_F(D^s = 1 \mid Z^s = z)$*

for $s \in \{w, m\}$. *In addition,* $\mathbb{P}(D^w = d, D^m = d', Z^w = z, Z^m = z') > 0$ *for every* $d, d' \in \mathcal{D}$ *and* $z, z' \in \mathcal{Z}$.

Assumption 4.2 requires that the instruments are binary, that the propensity score $p^s(z)$ strictly increases when z increases, and that the mass of households with each combination of the two treatments and the two instruments is positive. Note that $p^s(z)$ can be identified directly from the data and the assumption is testable.

Next, we assume that $D^s(\cdot)$ satisfies the well-known monotonicity property of Imbens and Angrist, 1994, as follows.

Assumption 4.3 (Monotonicity of treatment response)**.** *For each* $s \in \{w, m\}$, $D^s(1) \geq D^s(0)$ F*-almost surely.*

Under Assumptions 4.2 and 4.3, the treatment response types for each side are partitioned into three groups, namely always-takers ($D^s(0) = D^s(1) = 1$), compliers ($D^s(0) = 0 < 1 = D^s(1)$), and never-takers ($D^s(0) = D^s(1) = 0$). A complier-complier pair refers to a pair in which both the woman and the man are compliers. We define LATE among complier-complier pairs under F^* as follows.

$$LATE^*(d, d') := \mathbb{E}_{F^*}[Y(d, d') - Y(0, 0) \mid D^w(1) \neq D^w(0), D^m(1) \neq D^m(0)]$$

Because F^* is the independent coupling of F^w and F^m, conditioning on both sides being compliers describes the population obtained by collecting the female compliers and the male compliers and paring them at random. $LATE^*(d, d')$ is therefore the average treatment effect of the treatments (d, d') relative to $(0, 0)$ among these randomly matched complier-complier pairs.

To derive formulae that identify $LATE^*(d, d')$, we start from the observed conditional mean of the outcome with both treatments and both instruments set to one:

$$\begin{aligned}
&\mathbb{E}[Y \mid D^w = 1, D^m = 1, Z^w = 1, Z^m = 1] \\
&= \mathbb{E}_{F^*}[\pi(\phi^w(1,1), \phi^m(1,1)) \mid D^w = 1, D^m = 1, T^w = 1, T^m = 1, Z^w = 1, Z^m = 1] \\
&= \mathbb{E}_{F^*}[\pi(\phi^w(1,1), \phi^m(1,1)) \mid D^w(1) = 1, D^m(1) = 1].
\end{aligned}$$

The first equality holds because the observed distribution is coherent with (F, π) (Assumption 3.1) and F is a matching on observables (Assumption 3.3); the second holds because the instruments are

independent (Assumption 3.2), the matching is exogenous conditional on own treatment (Assumption 4.1), and the treatment response is monotone (Assumption 4.3). The left-hand side is identified directly from the distribution G of the observed variables, while the right-hand side is a mixture, with identifiable weights, of the local average treatment responses at $(D^w, D^m) = (1, 1)$ over the AT-AT, AT-CO, CO-AT, and CO-CO pairs under F^*, where AT and CO denote the always-taker and the complier.

We can derive similar equations on $\mathbb{E}[Y \mid D^w = d, D^m = d', Z^w = z, Z^m = z']$ for each $d, d' \in \mathcal{D}$, $z, z' \in \mathcal{Z}$. By combining these equations, we obtain the following identification results of $LATE^*(d, d')$.

Theorem 4.1 ($LATE$ among complier-complier pairs under F^*)**.** *Suppose that Assumptions 3.1, 3.2, 3.3, 4.1, 4.2, and 4.3 hold. Then,*

$$
\begin{aligned}
&LATE(1,1) \\
&= \sum_{(z,z')\in\{0,1\}^2} \frac{(-1)^{z+z'}\{e_{11}(z,z')\cdot p^w(z)\cdot p^m(z') - e_{00}(z,z')\cdot(1-p^w(z))\cdot(1-p^m(z'))\}}{\{p^w(1)-p^w(0)\}\cdot\{p^m(1)-p^m(0)\}}
\end{aligned}
$$

$$
\begin{aligned}
&LATE(1,0) \\
&= \sum_{(z,z')\in\{0,1\}^2} \frac{(-1)^{z+z'+1}\{e_{10}(z,z')\cdot p^w(z)\cdot(1-p^m(z')) + e_{00}(z,z')\cdot(1-p^w(z))\cdot(1-p^m(z'))\}}{\{p^w(1)-p^w(0)\}\cdot\{p^m(1)-p^m(0)\}}
\end{aligned}
$$

$$
\begin{aligned}
&LATE(0,1) \\
&= \sum_{(z,z')\in\{0,1\}^2} \frac{(-1)^{z+z'+1}\{e_{01}(z,z')\cdot(1-p^w(z))\cdot p^m(z') + e_{00}(z,z')\cdot(1-p^w(z))\cdot(1-p^m(z'))\}}{\{p^w(1)-p^w(0)\}\cdot\{p^m(1)-p^m(0)\}}
\end{aligned}
$$

where

$$
e_{dd'}(z,z') := \mathbb{E}[Y \mid D^w = d, D^m = d', Z^w = z, Z^m = z'].
$$

Theorem 4.1 expresses LATE among complier-complier pairs as a closed-form functional of the conditional outcome means $e_{dd'}(\cdot,\cdot)$ and the propensity scores $p^w(\cdot)$ and $p^m(\cdot)$, each of which is identified directly from the distribution G of the observed variables.

The formula is the two-sided version of the well-known LATE formula of Imbens and Angrist,

1994 for a binary treatment and a binary instrument. Their formula divides the difference in the mean outcome between the two instrument values by the difference in the propensity score. Then, under monotonicity condition, the numerator corresponds to the share of compliers times the average treatment effect among them, and the denominator corresponds to the share of compliers.[2]

Theorem 4.1 keeps this ratio with a maternal and a paternal instrument. Its numerator equals $\mathbb{E}_{F^*}[\{Y(1,1)-Y(0,0)\}\{D^w(1)-D^w(0)\}\{D^m(1)-D^m(0)\}]$, that is, the share of complier-complier pairs times the average treatment effect of $(1,1)$ relative to $(0,0)$ among them. Its denominator is $\mathbb{E}_{F^*}[\{D^w(1)-D^w(0)\}\{D^m(1)-D^m(0)\}]$, which equals $\{p^w(1)-p^w(0)\}\{p^m(1)-p^m(0)\}$, the share of complier-complier pairs under the independent coupling F^*.

Next, we define the counterpart under the actual matching F as

$$LATE(d,d') := \mathbb{E}_F\left[Y(d,d')-Y(0,0) \mid D^w(0)\neq D^w(1),\ D^m(0)\neq D^m(1)\right].$$

Its target population is described by the distribution of actual households in which both the mother and the father are compliers. $LATE(d,d')$ is therefore the average treatment effect of (d,d') relative to $(0,0)$ in that population. Then, we have the following result on the equivalence between $LATE(d,d')$ and $LATE^*(d,d')$.

Theorem 4.2 (the equivalence of LATE under F and F^*)**.** *Suppose that Assumptions 3.1, 3.2, 3.3, 4.1, 4.2, and 4.3 hold. Then, for each $d,d' \in \mathcal{D}$, $LATE(d,d') = LATE^*(d,d')$.*

The two parameters are the average treatment effect over different populations. The target population of $LATE^*(d,d')$ is the female compliers and the male compliers matched to one another at random, while that of $LATE(d,d')$ is the actual households in which both the mother and the father are compliers. Combined with Theorem 4.1, this equivalence shows that the same formula identifies $LATE(d,d')$, and the average treatment effect among complier-complier pairs under F is identified from the distribution of observed variables. Note that the equivalence holds only under conditional ignorability of the matching response.

[2] In the canonical setup with a binary treatment D and a binary instrument Z, their formula is

$$\frac{\mathbb{E}[Y \mid Z=1]-\mathbb{E}[Y \mid Z=0]}{\mathbb{E}[D \mid Z=1]-\mathbb{E}[D \mid Z=0]} = \sum_{z\in\{0,1\}} \frac{(-1)^{z+1}\{e_1(z)p(z)+e_0(z)(1-p(z))\}}{p(1)-p(0)},$$

where $e_d(z) := \mathbb{E}[Y \mid D=d, Z=z]$ and $p(z) := \mathbb{P}[D=1 \mid Z=z]$. Its numerator equals $\mathbb{E}[\{Y(1)-Y(0)\}\{D(1)-D(0)\}]$ and the denominator equals to $\mathbb{E}[D(1)-D(0)]$.

Remark 4.1. The matching on observables condition in Definition 3.3 requires the equality of the two distributions $F_{A|D^w=T^m=d,T^w=D^m=d',Z^w=z,Z^m=z'} = F^*_{A|D^w=T^m=d,T^w=D^m=d',Z^w=z,Z^m=z'}$ for every $(d,d',z,z') \in \mathcal{D}^2 \times \mathcal{Z}^2$ (Lemma 3.1). However, for the identification result in Theorem 4.1, we only need a weaker condition that the conditional expectations of potential outcomes are equal between F and F^*:

$$\begin{aligned}&\mathbb{E}_F[\pi(\phi^w(d,d'),\phi^m(d,d')) \mid D^w = T^m = d, T^w = D^m = d', Z^w = z, Z^m = z'] \\ =&\mathbb{E}_{F^*}[\pi(\phi^w(d,d'),\phi^m(d,d')) \mid D^w = T^m = d, T^w = D^m = d', Z^w = z, Z^m = z']\end{aligned}$$

for each $(d,d',z,z') \in \mathcal{D}^2 \times \mathcal{Z}^2$. This weaker condition allows for dependence in higher moments of potential outcomes between matched pairs. Note that this condition is not enough for the equivalence of $LATE$ and $LATE^*$ in Theorem 4.2. The equivalence relies on the matching on observables condition in Definition 3.3. □

Remark 4.2. Monotonicity of $D^s(\cdot)$ (i.e., Assumption 4.3) can be rationalized by a sequential threshold crossing model. To be more specific, for each $s \in \{w,m\}$,

$$D^s(z) = \mathbb{1}\left(\varepsilon_1^s \leq h_1^s(z)\right), \quad T^s(z) = \mathbb{1}\left(\varepsilon_2^s \leq h_2^s\left(D^s(z), z\right)\right), \tag{4.1}$$

where $h_1^s : \mathcal{Z} \to \mathbb{R}$ and $h_2^s : \mathcal{D} \times \mathcal{Z} \to \mathbb{R}$ are functions that determine the thresholds and $\varepsilon^s = (\varepsilon_1^s, \varepsilon_2^s)$ is continuously distributed unobserved heterogeneity. In the first stage of their life, each agent determines their treatment status (e.g., education), and in the second stage, she determines that of whom she matches. We can interpret that these thresholds are not primitives but equilibrium objects. For instance, they arise from a transferable matching model in which the thresholds h_1^s, h_2^s are determined so that demand and supply in the marriage market (e.g., the demand and supply of college-educated women) coincide. Appendix D.1 provides a more detailed discussion on the micro-foundation of the sequential threshold crossing model.

Since ε^s determines the treatment-matching response type K^s, we can reformulate the model so that each side-s agent is characterized by $A^s = (Z^s, D^s, T^s, \varepsilon^s, \boldsymbol{\phi}^s)$ for each $s \in \{w,m\}$. Then, Assumption 4.1 is also satisfied if $(\varepsilon_1^s, \boldsymbol{\phi}^s)$, ε_2^s, and Z^s are mutually independent under F for each $s \in \{w,m\}$. Importantly, unlike the equivalence result in a canonical binary treatment setup by Vytlacil, 2002, the sequential threshold crossing model is just an example under which Assumption

4.3 can be rationalized.

□

5 General case without conditional ignorability

Recall that Section 4 considers the LATE under a simplified setup in which the matching response is conditionally ignorable. In this section, we consider a more general setup in which conditional ignorability fails. For instance, in the college attendance example, even conditional on her preference over her own education, a mother's potential factors that affect her child's outcome may be correlated with her preference over the partner's education. This is plausible if a mother who cares more about her child's outcome also has a stronger preference for a college-educated spouse. Under this setup, we consider partial identification of various causal parameters such as the average treatment effect $\mathbb{E}_F\left[Y(d,d') - Y(0,0)\right]$.

5.1 ATE bounds under heterogeneous treatment effect

In what follows, we assume that Y is binary, that is, $\mathcal{Y} = \{0,1\}$. Let $\Theta := \mathcal{F} \times \Pi$ be the space of (F,π). Then, we define the identified set of (F,π) as follows.

Definition 5.1. *Given the distribution of observed variables G, the identified set of (F,π) is defined as*

$$\Theta_I := \{(F,\pi) \in \Theta : \text{ Assumption 3.1, 3.2, and 3.3 hold under } G \text{ and } (F,\pi).\}.$$

Θ_I collects every $(F,\pi) \in \Theta$ that is compatible with the distribution of observed variables G (Assumption 3.1) and whose F satisfies the independence of instruments (Assumption 3.2) and is an equilibrium matching and a matching on observables (Assumption 3.3).

Recall that $F^* = F^w \times F^m$, where F^w and F^m are the marginals on A^w and A^m under F, respectively. For the moment, we consider the lower bound of the ATE under the independent coupling F^*

$$\begin{aligned} &\inf_{(F,\pi)\in\Theta_I} \mathbb{E}_{F^*}\left[Y(d,d') - Y(0,0)\right] \\ =&\inf_{(F,\pi)\in\Theta_I} \mathbb{E}_{F^*}\left[\pi\left(\phi^w\left(d,d'\right),\phi^m\left(d,d'\right)\right) - \pi\left(\phi^w\left(0,0\right),\phi^m\left(0,0\right)\right)\right]. \end{aligned} \tag{5.1}$$

Focusing on the ATE under F^* here is purely a matter of notational convenience. The same key idea, through a parallel argument, also yields bounds for the ATE under F (replacing $\mathbb{E}_{F^*}$ with $\mathbb{E}_F$) and for other causal parameters.

Since this is an infinite dimensional problem, we wish to convert the problem into a finite dimensional problem. To this end, we will make some additional notations. Recall that $\mathcal{K} \subseteq (\mathcal{D}^2)^{|\mathcal{Z}|}$ is the space of the treatment-matching response type, i.e., K^w and K^m. We express each $k \in \mathcal{K}$ as $k = \{(d_z, t_z)\}_{z \in \mathcal{Z}}$, where (d_z, t_z) represents the treatment status of their own and their matched partner when the value of the instrument is z. Let $\mathcal{K}(d, t \mid z) := \{k \in \mathcal{K} : k_z = (d, t)\}$, where k_z is the z-th coordinate of k for $z \in \mathcal{Z}$.

We consider any $(F, \pi) \in \Theta_I$. Then, let

$$e(d, d' \mid k, k') := \mathbb{E}_{F^*}[Y(d, d') \mid K^w = k, K^m = k']$$

for each $d, d' \in \mathcal{D}$ and each $k, k' \in \mathcal{K}$. Note that $e(d, d' \mid k, k')$ is the average treatment response to $(D^w, D^m) = (d, d')$ for a pair whose woman has response type k and man has response type k'. Since Y is assumed to be binary, $0 \leq e(d, d' \mid k, k') \leq 1$ for each $d, d' \in \mathcal{D}$ and $k, k' \in \mathcal{K}$. We use $\mathbf{E} := [0, 1]^{|\mathcal{D}|^2 |\mathcal{K}|^2}$ to denote the space of $\{e(d, d' \mid k, k')\}_{d, d' \in \mathcal{D}, k, k' \in \mathcal{K}}$.

Next, we define

$$p_k^s := \mathbb{P}_F(K^s = k)$$

as the probability that the response type is k for each side $s \in \{w, m\}$. We use $\mathbf{P} := [0, 1]^{2|\mathcal{K}|}$ to denote the space of the response type probability vectors $(\{p_k^w\}_{k \in \mathcal{K}}, \{p_k^m\}_{k \in \mathcal{K}})$.

Under this setup, for any $(F, \pi) \in \Theta_I$, we can show that our target, $\mathbb{E}_{F^*}[Y(d, d') - Y(0, 0)]$, can be expressed as a function of $(e, p) \in \mathbf{E} \times \mathbf{P}$ as follows.

$$ATE^*(e, p) = \sum_{k \in \mathcal{K}} \sum_{k' \in \mathcal{K}} p_k^w p_{k'}^m \{e(d, d' \mid k, k') - e(0, 0 \mid k, k')\}. \tag{5.2}$$

The above objective function takes a product form $p_k^w p_{k'}^m$ because, under the independent coupling F^*, the treatment-matching response types of women and men are independently matched.

The values of (e, p) cannot be chosen freely. They must be consistent with the distribution of the observed data. We now derive two sets of constraints that any (e, p) induced by a $(F, \pi) \in \Theta_I$ must satisfy. Then, we consider two constraints to be satisfied under $(F, \pi) \in \Theta_I$. First, for each $\tilde{d}, \tilde{d}' \in \mathcal{D}$

and $z, z' \in \mathcal{Z}$, we can show that the conditional expectation of the outcome can be expressed as

$$\begin{aligned}
&\mathbb{E}\left[Y \mid D^w = \tilde{d}, D^m = \tilde{d}', Z^w = z, Z^m = z'\right] \\
&= \sum_{k \in \mathcal{K}(\tilde{d}, \tilde{d}' \mid z)} \sum_{k' \in \mathcal{K}(\tilde{d}', \tilde{d} \mid z')} \frac{p_k^w p_{k'}^m e(\tilde{d}, \tilde{d}' \mid k, k')}{\mathbb{P}\left(D^w = \tilde{d}, D^m = \tilde{d}' \mid Z^w = z\right) \mathbb{P}\left(D^w = \tilde{d}, D^m = \tilde{d}' \mid Z^m = z'\right)}.
\end{aligned} \tag{5.3}$$

The left-hand side of (5.3) is a conditional expectation that can be identified directly from the observed data, while the right-hand side is a function of the unknown objects (e, p) given identified objects $\mathbb{P}(D^w = \tilde{d}, D^m = \tilde{d}' \mid Z^s = z)$. Thus, the equality (5.3) serves as a constraint that links the unobserved treatment response parameters to the observed distribution. Note that the equality (5.3) holds since F satisfies the exogeneity of instruments and matching on observables.

Next, for each $s, s' \in \{w, m\}$ such that $s \neq s'$ and $\tilde{d}, \tilde{d}' \in \mathcal{D}$, we can express the propensity scores as

$$\mathbb{P}\left(D^s = \tilde{d}, D^{s'} = \tilde{d}' \mid Z^s = z\right) = \sum_{k \in \mathcal{K}(\tilde{d}, \tilde{d}' \mid z)} p_k^s. \tag{5.4}$$

The equality (5.4) requires that the probability that a woman chooses $(D^w, T^w) = (\tilde{d}, \tilde{d}')$ conditional on $Z^w = z$ is the probability that each woman's treatment-matching response type is included in $\mathcal{K}(\tilde{d}, \tilde{d}' \mid z)$, that is, the set of treatment-matching response types who choose $(\tilde{d}, \tilde{d}')$ when the value of the instruments is z. Note that the equality (5.4) holds since instruments are exogenous under F.

The above arguments imply that $\inf_{(F,\pi) \in \Theta_I} \mathbb{E}_{F^*}\left[Y(d, d') - Y(0, 0)\right] \geq \inf_{(e,p) \in \mathbf{E} \times \mathbf{P}} ATE^*(e, p)$ subject to (5.3)–(5.4) for every $\tilde{d}, \tilde{d}' \in \mathcal{D}$, $z, z' \in \mathcal{Z}$, and $s, s' \in \{w, m\}$ such that $s \neq s'$. In fact, we can show that the opposite inequality also holds, and the values of these two optimization problems are equivalent. The reverse direction holds because, for any (e, p) satisfying the constraints, one can construct $(F, \pi) \in \Theta_I$ that achieves the corresponding value of the objective function. Furthermore, the infimum and the supremum are attained for both problems. More formally, we have the following equivalence results.

Theorem 5.1. *Suppose that $\Theta_I \neq \emptyset$. Then, the lower bound and the upper bound of the ATE under F^* can be expressed as*

$$\begin{aligned}
&\min_{(F,\pi) \in \Theta_I} \mathbb{E}_{F^*}\left[Y(d, d') - Y(0, 0)\right] \\
&= \min_{(e,p) \in \mathbf{E} \times \mathbf{P}} ATE^*(e, p) \text{ subject to (5.3)–(5.4) for every } \tilde{d}, \tilde{d}' \in \mathcal{D},\ z, z' \in \mathcal{Z},\ s, s' \in \{w, m\} \text{ with } s \neq s',
\end{aligned}$$

and

$$\max_{(F,\pi)\in\Theta_I} \mathbb{E}_{F^*}\left[Y(d,d') - Y(0,0)\right]$$
$$= \max_{(e,p)\in\mathbf{E}\times\mathbf{P}} ATE^*(e,p) \text{ subject to (5.3)–(5.4) for every } \tilde{d},\tilde{d}' \in \mathcal{D},\ z,z' \in \mathcal{Z},\ s,s' \in \{w,m\} \text{ with } s \neq s'.$$

Theorem 5.1 states that the optimization problem (5.1) can be transformed into a finite dimensional problem, and the values of the two problems are equivalent. We can find similar results in many papers on causal models with instruments (e.g., Balke and Pearl, 1997; Bai et al., 2024).

Other causal parameters can be handled in the same way. Each is obtained by replacing the objective function $ATE^*(e,p)$ with a suitable alternative, while the constraints (5.3)–(5.4) remain unchanged. We provide the corresponding objective functions for these causal parameters in Appendix E.

5.2 Parametric restrictions on the selection model

We can further tighten the bounds by imposing additional assumptions. First, we consider the following parametric restrictions on the selection model.

Condition 5.1 (Parametric restrictions on the selection model)**.** *For each $s \in \{w,m\}$, there exist a distribution $F^s_\varepsilon(\cdot;\sigma)$ on $\mathbb{R}^{d_\varepsilon}$ and Borel-measurable functions $d^s(\cdot,\cdot;\gamma), t^s(\cdot,\cdot;\gamma) : \mathcal{Z}\times\mathbb{R}^{d_\varepsilon} \to \mathcal{D}$, known up to finite-dimensional parameters (σ,γ), such that*

$$\{(d^s(z,\varepsilon^s;\gamma), t^s(z,\varepsilon^s;\gamma))\}_{z\in\mathcal{Z}} \sim F^s_K \text{ for } \varepsilon^s \sim F^s_\varepsilon(\cdot;\sigma), \tag{5.5}$$

where F^s_K is the marginal distribution over K^s under F.

The condition requires that (D^w,T^w) and (D^m,T^m) are chosen according to the functions (d^w,t^w) and (d^m,t^m), which are known up to a finite dimensional parameter γ.[3] Also, the unobserved heterogeneity ε^w and ε^m follow $F^w_\varepsilon(\cdot;\sigma)$ and $F^m_\varepsilon(\cdot;\sigma)$, which are also known up to a finite dimensional parameter σ. We define $\delta = (\sigma,\gamma) \in \Delta$ as the parameter which determines the marginal distribution of the treatment-matching response type F^s_K for each side $s \in \{w,m\}$, where $\Delta := \Sigma\times\Gamma$ is the compact parameter space of δ. Note that, given $\delta \in \Delta$, the unobserved heterogeneity ε^s characterizes the treatment-matching response type K^s of each person in each side $s \in \{w,m\}$.

[3]Note that this setup includes a nonparametric specification of (d^s,t^s) since $\mathcal{Z}$ is finite.

Such parametric restrictions on the selection model are common when we consider nonbinary treatments (e.g., Kamat et al., 2023). There are many candidates for the selection model (d^s, t^s). Here we introduce sequential decision models and additive random utility models.

Example 5.1 (Sequential threshold-crossing model)**.** The first example is a sequential threshold-crossing model discussed in Remark 4.2. For each $s \in \{w, m\}$, there exist $h_1^s : \mathcal{Z} \to \mathbb{R}$ and $h_2^s : \mathcal{Z} \times \mathcal{D} \to \mathbb{R}$, known up to γ, such that

$$d^s(z, \varepsilon^s; \gamma) = \mathbb{1}\left(\varepsilon_1^s \leq h_1^s(z; \gamma)\right)$$

$$t^s(z, \varepsilon^s; \gamma) = \mathbb{1}\left(\varepsilon_2^s \leq h_2^s\left(z, d^s(z, \varepsilon^s; \gamma); \gamma\right)\right),$$

where $\varepsilon^s = (\varepsilon_1^s, \varepsilon_2^s) \sim F_\varepsilon^s(\cdot; \sigma)$ is the 2-dimensional unobserved heterogeneity. □

Example 5.2 (Additive random utility model)**.** Second, we consider an additive random utility model. To be more specific, for each $(d, t) \in \mathcal{D}^2$, there exist $U_{dt}^w : \mathcal{Z} \to \mathbb{R}$ and $U_{td}^m : \mathcal{Z} \to \mathbb{R}$, which are known up to γ, such that

$$(d^w(z, \varepsilon^w; \gamma), t^w(z, \varepsilon^w; \gamma)) = \underset{(d,t)\in\mathcal{D}\times\mathcal{D}}{\arg\max}\, U_{dt}^w(z; \gamma) + \varepsilon_{dt}^w$$

$$(d^m(z, \varepsilon^m; \gamma), t^m(z, \varepsilon^m; \gamma)) = \underset{(d,t)\in\mathcal{D}\times\mathcal{D}}{\arg\max}\, U_{td}^m(z; \gamma) + \varepsilon_{td}^m,$$

where $\varepsilon^w = \{\varepsilon_{dt}^w\}_{(d,t)\in\mathcal{D}^2} \sim F_\varepsilon^w(\cdot; \sigma)$ and $\varepsilon^m = \{\varepsilon_{td}^m\}_{(d,t)\in\mathcal{D}^2} \sim F_\varepsilon^m(\cdot; \sigma)$ are the unobserved taste shocks of each woman and man, respectively.

The utility that a woman with instrument vector z derives from selecting a treatment profile (d, t) is given by an additively separable form $U_{dt}^w(z; \gamma) + \varepsilon_{dt}^w$, where $U_{dt}^w(z; \gamma)$ is the deterministic component, and ε_{dt}^w captures unobserved heterogeneity. The utility of a man who selects own treatment d and partner treatment t is $U_{td}^m(z; \gamma) + \varepsilon_{td}^m$, where the subscript is the household cell (t, d) formed by his choice (his own treatment is the second coordinate). Each agent selects the treatment status of their own and that of their matched partner simultaneously in order to maximize their random utility.

Selection with additive random utility models is studied in, for instance, Heckman et al., 2008, Lee and Salanié, 2018, 2020, Tsuda, 2022, and Kamat et al., 2023. Importantly, each agent's utility depends not on the identity of the partner but on the treatment status of the partner. This formulation aligns with the literature on empirical matching models (e.g., Choo and Siow, 2006;

Chiappori et al., 2017; Chiappori et al., 2018; Galichon et al., 2019; Galichon and Salanié, 2022; Gualdani and Sinha, 2023). As in the sequential case, we can interpret these utilities as a result of some (incompletely) transferable matching model, with the systematic utilities U^s determined so that demand and supply in the marriage market coincide; Appendix D.2 provides a more detailed discussion. □

Under this setup, we define the restricted identified set of (F, π) as

$$\tilde{\Theta}_I := \{(F,\pi) \in \Theta_I : \text{ There exists } \delta \in \Delta \text{ such that (5.5) holds for each } s \in \{w, m\}.\}.$$

Note that $\tilde{\Theta}_I$ collects the elements of the original identified set Θ_I that are compatible with the parametric restriction in Condition 5.1. We wish to obtain the lower and upper bound of $\mathbb{E}_{F^*}[Y(d,d') - Y(0,0)]$ over $(F,\pi) \in \tilde{\Theta}_I$.

To this end, for each side $s \in \{w, m\}$, we use $\mathcal{E}^s(k;\gamma)$ to denote the set of values of the unobserved heterogeneity ε^s that induce the response type k. More formally,

$$\mathcal{E}^s(k;\gamma) := \{\varepsilon^s \in \mathcal{E} : \{(d^s(z,\varepsilon^s;\gamma), t^s(z,\varepsilon^s;\gamma))\}_{z\in\mathcal{Z}} = k\}.$$

Thus the probability that the response type is k can be expressed as

$$p_k^s = \int \mathbb{1}(\varepsilon \in \mathcal{E}^s(k;\gamma)) dF_\varepsilon^s(\varepsilon;\sigma). \tag{5.6}$$

We define a function $\tilde{p}_k^s : \Delta \to [0,1]$ as $\tilde{p}_k^s(\delta) = \int \mathbb{1}(\varepsilon \in \mathcal{E}^s(k;\gamma)) dF_\varepsilon^s(\varepsilon;\sigma)$. In what follows, we assume that $\tilde{p}_k^s$ is continuously differentiable with respect to δ for each $k \in \mathcal{K}$ and $s \in \{w, m\}$.

In summary, under the parametric selection model, the type probabilities p_k^s are no longer free parameters but are determined by δ through (5.6). This additional condition reduces the feasible set and can potentially tighten the bounds on the causal parameters. More formally, we have the following result.

Corollary 5.1. *Suppose that $\tilde{\Theta}_I \neq \emptyset$. Then, the lower bound and the upper bound of the ATE under*

F^* can be expressed as

$$\begin{aligned}
&\min_{(F,\pi)\in\tilde{\Theta}_I} \mathbb{E}_{F^*}\left[Y(d,d') - Y(0,0)\right] \\
=&\min_{(e,p,\delta)\in\mathbf{E}\times\mathbf{P}\times\Delta} ATE^*(e,p) \\
&\text{subject to (5.3), (5.4), and (5.6) for every } \tilde{d},\tilde{d}'\in\mathcal{D},\ z,z'\in\mathcal{Z},\ k,k'\in\mathcal{K},\ s,s'\in\{w,m\} \text{ with } s\neq s',
\end{aligned}$$

and

$$\begin{aligned}
&\max_{(F,\pi)\in\tilde{\Theta}_I} \mathbb{E}_{F^*}\left[Y(d,d') - Y(0,0)\right] \\
=&\max_{(e,p,\delta)\in\mathbf{E}\times\mathbf{P}\times\Delta} ATE^*(e,p) \\
&\text{subject to (5.3), (5.4), and (5.6) for every } \tilde{d},\tilde{d}'\in\mathcal{D},\ z,z'\in\mathcal{Z},\ k,k'\in\mathcal{K},\ s,s'\in\{w,m\} \text{ with } s\neq s'.
\end{aligned}$$

As in Theorem 5.1, Corollary 5.1 states that the optimization problem (5.1) can be transformed into a finite dimensional problem, even with the additional parametric assumptions on the selection model (5.5).

5.2.1 Computing the bounds

We now turn to the computational aspect of the bounds, that is, how the optimization problem in Corollary 5.1 can be solved in practice. In what follows, we consider the case in which there is a unique $\delta \in \Delta$ such that, for every $\tilde{d},\tilde{d}' \in \mathcal{D}$, $z,z' \in \mathcal{Z}$, $k,k' \in \mathcal{K}$, $s \in \{w,m\}$, (5.4) and (5.6) are satisfied. That is, δ is point identified from the selection model, which implies $p = (\{\tilde{p}_k^w(\delta)\}_{k\in\mathcal{K}}, \{\tilde{p}_k^m(\delta)\}_{k\in\mathcal{K}})$ can be also point identified. Then, given p, we have

$$\begin{aligned}
&\min_{(F,\pi)\in\tilde{\Theta}_I} \mathbb{E}_{F^*}\left[Y(d,d') - Y(0,0)\right] \\
=&\min_{e\in\mathbf{E}} ATE^*(e,p) \\
&\text{subject to (5.3) for every } \tilde{d},\tilde{d}'\in\mathcal{D},\ z,z'\in\mathcal{Z},\ k,k'\in\mathcal{K},\ s,s'\in\{w,m\} \text{ with } s\neq s'.
\end{aligned} \tag{5.7}$$

Note that the problem in the right hand side can be solved via linear programming. Partitioning the unobserved heterogeneity into response types and characterizing treatment-effect bounds as the value of a linear program is not new in the literature on causal models with instruments (e.g., Mogstad et al., 2018; Lafférs, 2019; Gu and Russell, 2023; Kamat et al., 2023; Bai et al., 2024; Ura

and Zhang, 2024).

6 Inference on the ATE bounds

In this section, we develop uniformly valid inference procedures for the lower and upper bounds of the ATE under the parametric assumptions on the selection model introduced in Section 5.2. Throughout, we maintain the assumption that the selection model parameter δ is point identified, in the sense that there exists a unique $\delta \in \Delta$ satisfying (5.4) and (5.6). Recall that, then, $p = (\{\tilde{p}_k^w(\delta)\}_{k\in\mathcal{K}}, \{\tilde{p}_k^m(\delta)\}_{k\in\mathcal{K}})$ can be also point identified.

Our inference procedure extends the framework of Gafarov, 2025 to the two-stage structure of our problem, as described below. In the first stage, we estimate the finite-dimensional parameter

$$\tau = (\delta, \{\mathbb{P}(D^w = d, D^m = d', Z^w = z, Z^m = z', Y = y)\}_{(d,d',z,z',y)\in\mathcal{D}^2\times\mathcal{Z}^2\times\mathcal{Y}}) \in \mathbb{R}^{\dim(\tau)},$$

which collects the selection model parameter δ and the joint distribution of treatments, instruments, and outcomes. The parameter δ is estimated based on the parametric selection model (5.4) and (5.6) and the joint probabilities are estimated from their sample analogs. Once τ is known, the constraints of the linear program (5.7) are fully determined, and the bounds on the ATE can be computed by solving a linear programming. In the second stage, given the first-stage estimate $\hat{\tau}$, we conduct inference for these bounds.

To formalize the inference problem, we introduce the following notation. Let $\mathcal{G}$ denote a class of distributions on the observed variables (Z^w, Z^m, D^w, D^m, Y), and let $\tilde{\Theta}_I(G)$ denote the identified set of (F, π) under $G \in \mathcal{G}$ with the parametric restrictions. For each $G \in \mathcal{G}$, let δ_G be the unique $\delta \in \Delta$ satisfying (5.4) and (5.6) under G, and write $\tau_G = (\delta_G, \{\mathbb{P}_G(D^w = d, D^m = d', Z^w = z, Z^m = z', Y = y)\}_{(d,d',z,z',y)\in\mathcal{D}^2\times\mathcal{Z}^2\times\mathcal{Y}}) \in \mathbb{R}^{\dim(\tau)}$ for the corresponding first-stage parameter. The lower bound of the ATE under G is $\underline{ATE^*}_G = \min_{(F,\pi)\in\tilde{\Theta}_I(G)} \mathbb{E}_{F^*}[Y(d,d') - Y(0,0)]$, and $\overline{ATE^*}_G$ is defined analogously by replacing the minimum with a maximum. Our goal is to construct confidence intervals for the identified set $[\underline{ATE^*}_G, \overline{ATE^*}_G]$ that are uniformly valid over $\mathcal{G}$. We guarantee the coverage probability simultaneously for all $G \in \mathcal{G}$ rather than only pointwise for a fixed distribution.

We now cast the problem in a form amenable to the inference methods of Gafarov, 2025. Let $\dim(\xi)$, L_{eq}, and L_{ineq} denote the number of decision variables, equality constraints, and inequality constraints of the linear program, respectively, and define $L = L_{eq} + L_{ineq}$. We construct an

$L \times (\dim(\xi)+1)$ matrix $Q(\tau) = (Q_1(\tau), Q_2(\tau), \ldots, Q_L(\tau))'$ as a function of τ such that $\underline{v}(Q(\tau_G)) = \underline{ATE}^*_G$ for every $G \in \mathcal{G}$, where, for any $L \times (\dim(\xi)+1)$ matrix Q, the quantity $\underline{v}(Q)$ denotes the optimal value of the linear program

$$\underline{v}(Q) = \min_{\xi \in \mathbb{R}^{\dim(\xi)}} (1, 0, \ldots, 0) \cdot \xi \quad \text{subject to} \quad \begin{cases} \sum_{\ell'=1}^{\dim(\xi)} \xi_{\ell'} Q_{\ell,\ell'} + Q_{\ell,(\dim(\xi)+1)} = 0, & 1 \le \ell \le L_{eq} \\ \sum_{\ell'=1}^{\dim(\xi)} \xi_{\ell'} Q_{\ell,\ell'} + Q_{\ell,(\dim(\xi)+1)} \le 0, & L_{eq}+1 \le \ell \le L. \end{cases} \tag{6.1}$$

The formal definition of $Q(\tau)$ is provided in the Appendix G.5. For each $1 \le \ell \le L$, the function $r_\ell(Q, \xi) := \sum_{\ell'=1}^{\dim(\xi)} \xi_{\ell'} Q_{\ell,\ell'} + Q_{\ell,(\dim(\xi)+1)}$ measures the slack in the ℓ-th constraint at a given point ξ. Then, the feasible set of the problem given Q can be written as

$$\Xi(Q) := \left\{ \xi \in \mathbb{R}^{\dim(\xi)} : \begin{array}{l} r_\ell(Q, \xi) = 0 \text{ for every } 1 \le \ell \le L_{eq} \\ r_\ell(Q, \xi) \le 0 \text{ for every } L_{eq}+1 \le \ell \le L \end{array} \right\}.$$

In what follows, we write $Q_G := Q(\tau_G)$ and $r(Q, \xi) := (r_1(Q, \xi), \ldots, r_L(Q, \xi))'$.

A central challenge for inference is that the linear program (6.1) may admit multiple optimal primal and dual solutions. Then, the value function $\underline{v}(Q)$ is not differentiable in Q at the true parameter value, which invalidates standard delta method for deriving the asymptotic distribution of the plug-in estimator $\underline{v}(\hat{Q}_G)$. To deal with this problem, we augment the objective function with a quadratic penalty term and consider the regularized problem

$$\underline{v}(\mu, Q) = \min_{\xi \in \Xi(Q)} (1, 0, \ldots, 0) \cdot \xi + \mu \|\xi\|^2,$$

where $\mu > 0$ is a regularization parameter. The strict convexity of the augmented objective guarantees that the regularized problem has a unique primal solution, which we denote by $\underline{\xi}(\mu, Q)$.

We now state the regularity conditions on the class of distributions $\mathcal{G}$.

Assumption 6.1 (Class of distributions). *(i) For every $G \in \mathcal{G}$, $\tilde{\Theta}_I(G)$ is nonempty.*

(ii) For every $G \in \mathcal{G}$, there exists a unique $\delta_G \in int(\Delta)$ such that, under δ_G and G, (5.4) and (5.6) hold for every $\tilde{d}, \tilde{d}' \in \mathcal{D}$, $z, z' \in \mathcal{Z}$, $k, k' \in \mathcal{K}$, $s \in \{w, m\}$.

(iii) $\inf_{G \in \mathcal{G}} \mathbb{P}_G(D^w = d, D^m = d', Z^w = z, Z^m = z') > 0$ for each $d, d' \in \mathcal{D}$ and $z, z' \in \mathcal{Z}$.

(iv) *For every $G \in \mathcal{G}$, there exists $\xi \in \Xi(Q_G)$ such that $r_\ell(Q_G, \xi) < 0$ for every $L_{eq} + 1 \leq \ell \leq L$.*

(v) *For any $G \in \mathcal{G}$ and any combination of constraints which consists of all L_{eq} equality constraints and $\dim(\xi) - L_{eq}$ inequality constraints, the corresponding submatrices of coefficients of the full set of constraints Q_G have singular values that are bounded from below by a positive constant uniformly over $\mathcal{G}$.*

(vi) *For any $G \in \mathcal{G}$, any $\xi \in \Xi(Q_G)$, and any combination of constraints which consists of all L_{eq} equality constraints and $(\dim(\xi) - L_{eq} + 1)$ inequality constraints, the norm of the corresponding subvector of $r(Q_G, \xi)$ is bounded from below by a positive constant uniformly over $\mathcal{G}$.*

Condition (i) ensures that the model is correctly specified for every distribution G in $\mathcal{G}$. Condition (ii) requires that the selection model parameter δ_G is uniquely determined and lies in the interior of the parameter space. Condition (iii) ensures that every combination of the two treatments and the two instruments must occur with positive probability, uniformly over $\mathcal{G}$. Condition (iv) is Slater's condition, which requires the existence of a strictly feasible point in the linear program. Together with the quadratic regularization, this guarantees strong duality. Conditions (v) and (vi) are technical conditions that ensure the uniqueness of the dual solution $\underline{\lambda}(\mu, Q)$ for every $Q = Q_G$ with $G \in \mathcal{G}$, which is crucial for the validity of the asymptotic linear representation provided below.

Let $\hat{\tau}_G$ denote the estimator of τ_G. Then we make the following assumptions.

Assumption 6.2 (First-stage estimator). *(i)* *For every $\epsilon > 0$, $\lim_{n\to\infty} \sup_{G\in\mathcal{G}} \mathbb{P}_G(\|\hat{\tau}_G - \tau_G\| \geq \epsilon) = 0$.*

(ii) *$\lim_{n\to\infty} \sup_{G\in\mathcal{G}} \rho(\sqrt{n}(\hat{\tau}_G - \tau_G), N(0, \Sigma_G)) = 0$, where ρ denotes the Lévy–Prokhorov metric.*

(iii) *$\sup_{G\in\mathcal{G}} \|\Sigma_G\| < \infty$.*

(iv) *$\lim_{n\to\infty} \sup_{G\in\mathcal{G}} \|\hat{\Sigma}_G - \Sigma_G\| = 0$.*

Condition (i) requires uniform consistency of the first-stage estimator over $\mathcal{G}$. Condition (ii) ensures uniform asymptotic normality at the $\sqrt{n}$ rate. Here, the convergence is measured in the Lévy–Prokhorov metric, which metrizes weak convergence of probability measures (see, for instance, Van der Vaart and Wellner, 1996) . Condition (iii) requires that the asymptotic covariance matrix Σ_G is uniformly bounded over $\mathcal{G}$. Condition (iv) requires a uniformly consistent estimator $\hat{\Sigma}_G$ of the asymptotic covariance matrix. These are high-level conditions that can be verified for specific parametric models and estimation methods.

Under Assumptions 6.1 and 6.2, the regularized value estimator admits the following asymptotic linear representation:

$$\sqrt{n}(\underline{v}(\mu_n, \hat{Q}_G) - \underline{v}(\mu_n, Q_G)) = \sqrt{n}\underline{\lambda}(\mu_n, Q_G)' r(\hat{Q}_G, \underline{\xi}(\mu_n, Q_G)) + O_{\mathcal{G}}\left(\frac{1}{\mu_n\sqrt{n}}\right),$$

where $\hat{Q}_G := Q(\hat{\tau}_G)$ and the remainder term is a uniformly tight sequence over $G \in \mathcal{G}$ scaled by $1/(\mu_n\sqrt{n})$. The leading term is driven by the constraint residuals evaluated at the estimated parameters, weighted by the dual solution of the regularized problem. The remainder vanishes provided $\mu_n \to 0$ and $\mu_n\sqrt{n} \to \infty$.

The asymptotic variance of the regularized value estimator is estimated by

$$\hat{\underline{\eta}}^2(\mu_n, G) := \hat{\underline{\lambda}}_G' \hat{\underline{R}}_G \hat{\Sigma}_G \hat{\underline{R}}_G' \hat{\underline{\lambda}}_G,$$

where $\hat{\underline{\lambda}}_G := \underline{\lambda}(\mu_n, \hat{Q}_G)$, $\hat{\underline{R}}_G := R(\hat{\tau}_G, \underline{\xi}(\mu_n, \hat{Q}_G))$, and $R(\tau, \xi) := \nabla_\tau (r_1(Q(\tau), \xi), \ldots, r_L(Q(\tau), \xi))'$.

Since $\underline{v}(\mu_n, \hat{Q}_G)$ is upwardly biased due to the regularization term, we consider the following bias-corrected estimator, $\underline{v}^{out}(\mu_n, \hat{Q}_G) := \underline{v}(\mu_n, \hat{Q}_G) - \mu_n \|\underline{\xi}^*(\hat{Q}_G)\|^2$, where $\underline{\xi}^*(Q)$ is given by

$$\underline{\xi}^*_\ell(Q) := \max\{\underline{\xi}^+_\ell(Q), \underline{\xi}^-_\ell(Q)\}$$

with

$$\underline{\xi}^\pm_\ell(Q) := \| \min_{\xi \in \Xi(Q), \xi_1 \le \underline{v}(Q) + \mu_n} \{\pm \xi_\ell\} \|.$$

For each component ℓ, the auxiliary linear programs defining $\underline{\xi}^\pm_\ell(Q)$ optimize ξ_ℓ in the positive and negative directions over the set of feasible solutions whose objective value is within μ_n of the optimum $\underline{v}(Q)$. Taking the componentwise maximum yields a vector $\underline{\xi}^*(Q)$ satisfying $\|\underline{\xi}^*(Q)\| \ge \|\underline{\xi}\|$ for any primal solution $\underline{\xi}$ of the original problem (6.1).

We define $\hat{\underline{\eta}}^+ = \max\{\hat{\underline{\eta}}(\mu_n, G), \eta_0\}$, where η_0 is a small positive constant. Analogously, we define $\overline{v}^{out}(\mu_n, Q)$ and $\hat{\overline{\eta}}^+$. Then, we can construct confidence intervals for the identified set $[\underline{ATE^*}_G, \overline{ATE^*}_G]$ using the bias-corrected estimators and their estimated standard errors, with critical values from the standard normal distribution.

Proposition 6.1. *Suppose that $\mu_n \to 0$ and $\mu_n\sqrt{n} \to \infty$. Then, under Assumptions 6.1 and 6.2,*

$$\liminf_{n\to\infty} \inf_{G\in\mathcal{G}} \mathbb{P}_G \left(\left[\underline{ATE^*}_G, \overline{ATE^*}_G\right] \subset \left[\underline{v}^{out}(\mu_n, \hat{Q}_G) - c_{1-\frac{\alpha}{2}} \frac{\hat{\eta}^+}{\sqrt{n}}, \overline{v}^{out}(\mu_n, \hat{Q}_G) + c_{1-\frac{\alpha}{2}} \frac{\hat{\eta}^+}{\sqrt{n}}\right] \right) \geq 1-\alpha,$$

where $c_{1-\frac{\alpha}{2}}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

The confidence interval in Proposition 6.1 covers the entire identified set $[\underline{ATE^*}_G, \overline{ATE^*}_G]$. The bias correction ensures conservativeness of the bounds, while the regularization enables the use of standard normal critical values $c_{1-\alpha/2}$ rather than more complex inference procedures that would be required to handle the non-differentiability of the original value function.

Remark 6.1. The inference procedure developed above relies on the parametric restrictions introduced in Section 5.2 and the assumption that δ is point identified. If no parametric restrictions are imposed, the optimization problem is no longer a linear program. In this case, we can apply McCormick relaxation as in Ura and Zhang, 2024 to obtain a linear programming relaxation, to which existing inference frameworks (e.g., Fang et al., 2023; Gafarov, 2025) directly apply. If the parametric restrictions are maintained but δ is not assumed to be point identified, a joint confidence set for (τ, ATE^*) can be constructed by directly applying, for instance, Fang et al., 2023. □

7 Empirical application

We use the framework developed above to study the causal effect of parental college attendance on the educational investment that a couple makes on their child. This investment is not a child outcome itself such as the child's own educational attainment or income. A large literature on human capital formation shows, however, that parental investments in children are a key channel through which human capital is transmitted across generations (e.g., Cunha and Heckman, 2007; Heckman and Mosso, 2014; Attanasio et al., 2022). Therefore, taking the educational investment itself as the outcome variable Y is well grounded. Studying the causal effect is complicated by the endogeneity of educational choice and matching. A parent who places weight on their own schooling and on that of their partner is plausibly inclined to invest more in the child's education as well. Hence, both each parent's college attendance and the formation of the couple are correlated with unobserved determinants of the investment.

Table 2: Sample composition.

(D^w, D^m)	Number of households	Fraction
$(0,0)$	1,770	0.527
$(0,1)$	824	0.245
$(1,0)$	163	0.048
$(1,1)$	604	0.180

Notes. The sample consists of 3,361 couples.

7.1 Data

We use Japanese household-level panel data, pooling the Japanese Panel Survey of Consumers (JPSC) and the Japan Household Panel Survey (JHPS). The unit of observation is a married couple. We restrict the sample to 3,361 households in which both the mother and the father were born between 1959 and 1988.

The treatments D^w and D^m equal one if the mother and the father, respectively, attended a four-year college. The outcome Y is a binary indicator of high investment in the child's education. For each age of the first child, we compare the household's education expenditure with the median among households whose child is of the same age in the same calendar year. We set $Y = 1$ when the household lies above this median for a majority of the child's observed ages. Each of the instruments Z^w and Z^m consists of two variables. The first is the education of that parent's own parents, measured by the sum of their years of schooling. For the mother's instrument Z^w these are the child's maternal grandparents, and for the father's instrument Z^m the child's paternal grandparents. The second is the local availability of higher education, measured by the density of colleges in the parent's birthplace prefecture when the parent was seventeen. We binarize each of these variables at its median into a low and a high value.

Table 2 summarizes the composition of the sample. Fathers are more likely to have attended college (0.42) than mothers (0.23). Education is strongly positively assorted within couples. A college-educated woman has a college-educated husband with probability $\mathbb{P}_F(D^m = 1 \mid D^w = 1) = 0.79$, against $\mathbb{P}_F(D^m = 1 \mid D^w = 0) = 0.32$ for a woman who did not attend college. Symmetrically, a college-educated man has a college-educated wife with probability $\mathbb{P}_F(D^w = 1 \mid D^m = 1) = 0.42$, against $\mathbb{P}_F(D^w = 1 \mid D^m = 0) = 0.08$. Couples with a college-educated mother and a non-college father, $(D^w, D^m) = (1,0)$, are correspondingly rare.

7.2 Specification

We specify the selection model as the sequential threshold-crossing model of Section 5.2 with bivariate-normal errors. Following Example 5.1, for each $s \in \{w, m\}$, the potential treatment and matching are

$$D^s(z) = \mathbb{1}(\varepsilon_1^s \leq h_1^s(z;\gamma)), \qquad T^s(z) = \mathbb{1}(\varepsilon_2^s \leq h_2^s(D^s(z), z;\gamma)),$$

and we parametrize the thresholds as linear indices in the instruments,

$$h_1^s(z;\gamma) = \gamma_{10}^s + \gamma_{11}^s z_1 + \gamma_{12}^s z_2, \qquad h_2^s(d,z;\gamma) = \gamma_{20}^s(1-d) + \gamma_{21}^s\, d + \gamma_{22}^s z_1 + \gamma_{23}^s z_2.$$

The shock $\varepsilon^s = (\varepsilon_1^s, \varepsilon_2^s)$ follows the bivariate normal distribution with mean zero and unit variances. In the main specification, we set the correlation $\mathrm{Corr}(\varepsilon_1^s, \varepsilon_2^s)$ to zero, and we assess the sensitivity of the results by estimating the bounds over a range of values of this correlation. The first-stage parameters τ collect the threshold coefficients γ and the distribution of the observables $\mathbb{P}(D^w, D^m, Z^w, Z^m, Y)$, which we estimate by the maximum likelihood method.

In addition, on the expected potential outcome, we consider the following three possible restrictions. The first restriction is the monotone treatment response

$$e(1, d' \mid k, k') \geq e(0, d' \mid k, k') \quad \text{and} \quad e(d, 1 \mid k, k') \geq e(d, 0 \mid k, k'), \quad \text{for every } d, d' \in \mathcal{D},\ k, k' \in \mathcal{K}.$$

Holding the response type fixed, each parent's own college attendance weakly raises the child's expected outcome, separately for the mother and the father. This is the componentwise, two-sided counterpart of the monotone treatment response of Manski, 1997.

Second, we consider additive separability of $e(d, d' \mid k, k')$ across the two sides:

$$e(d, d' \mid k, k') = e^w(d \mid k) + e^m(d' \mid k'), \quad \text{for every } d, d' \in \mathcal{D}, k, k' \in \mathcal{K},$$

where $e^w(d \mid k)$ is the maternal part that depends only on the mother's own treatment and response type, and $e^m(d' \mid k')$ is the paternal part that depends only on the father's own treatment and response type. Note that this does not exclude the possibility that a parent who places weight on their partner's education tends to invest more in the child's education as well, because the response type K^s includes the matching response type.

The third restriction is the conditional ignorability of matching response. Then $e(d, d' \mid k, k')$

depends on the response types k and k' only through the treatment responses they induce, $D^w(\cdot)$ and $D^m(\cdot)$. Equivalently, $e(d, d' \mid k, k') = e(d, d' \mid \tilde{k}, \tilde{k}')$ whenever k and $\tilde{k}$ induce the same maternal treatment response and k' and $\tilde{k}'$ induce the same paternal treatment response. Note that this restriction is implied by the exogeneity of the matching shock ε_2^s. Even with these three additional restrictions, given the first-stage parameters τ, the bounds on the average treatment effect are obtained as the solutions of linear programs, exactly as in the analysis above.

7.3 Results

Table 3 reports the parameter estimates of the selection model in the main specification. In the treatment equation, the coefficients on both instruments are positive for women and men, which means that the instruments raise their college attendance. Both coefficients are significant at the 95% level for women and men. In the matching equation, the coefficients on the instruments are again positive, which means that the instruments raise the chance of matching with a college-educated partner even conditional on their own college attendance. For both women and men, both coefficients are significant at the 95% level. This suggests that the instruments are relevant for both the treatment choice and the matching choice.

First, Table 4 reports the LATE among complier-complier pairs, $LATE(d, d')$, defined in Theorem 4.1. The estimated effect of both parents attending college relative to neither, $LATE(1, 1)$, is 0.562, with a 95% confidence interval of $[0.467, 0.658]$. The effect of the father attending college, $LATE(0, 1)$, is 0.284, with a confidence interval of $[0.208, 0.360]$. The effect of the mother attending college, $LATE(1, 0)$, is 0.278, with a confidence interval of $[0.182, 0.374]$.

Table 5 reports the estimated bounds on the average treatment effects, together with their 95% confidence intervals, under the additional restrictions discussed above. Without additional restrictions, the bounds are wide, and every estimated bound covers both negative and positive values. Imposing monotone treatment response guarantees that each effect is at least zero. Note that the estimated lower bound of each effect nevertheless remains at zero rather than strictly positive. However, under additive separability and monotone treatment response together, the effect of both parents having attended college relative to neither, $\mathbb{E}_F[Y(1, 1) - Y(0, 0)]$, has an estimated bound of $[0.19, 0.53]$ and a 95% confidence interval of $[0.11, 0.64]$. For $\mathbb{E}_F[Y(0, 1) - Y(0, 0)]$ and $\mathbb{E}_F[Y(1, 0) - Y(0, 0)]$, the estimated bounds lie above zero, although their 95% confidence intervals contain zero.

Table 3: First-stage parameter estimates on the selection model.

	Mother (δ_w)	Father (δ_m)
Treatment equation: h_1^s		
Intercept: γ_{10}^s	-1.670	-1.106
	(0.060)	(0.053)
	$[-1.787,\ -1.552]$	$[-1.210,\ -1.003]$
Grandparents' education: γ_{11}^s	0.455	0.483
	(0.026)	(0.026)
	$[0.404,\ 0.507]$	$[0.432,\ 0.533]$
Density of colleges: γ_{12}^s	0.125	0.165
	(0.043)	(0.039)
	$[0.040,\ 0.209]$	$[0.088,\ 0.242]$
Matching equation: h_2^s		
Intercept $(D^s = 0)$: γ_{20}^s	-0.816	-1.718
	(0.056)	(0.063)
	$[-0.924,\ -0.707]$	$[-1.842,\ -1.594]$
Intercept $(D^s = 1)$: γ_{21}^s	0.334	-0.662
	(0.081)	(0.073)
	$[0.175,\ 0.493]$	$[-0.804,\ -0.520]$
Grandparents' education: γ_{22}^s	0.193	0.201
	(0.028)	(0.028)
	$[0.139,\ 0.247]$	$[0.145,\ 0.256]$
Density of colleges: γ_{23}^s	0.163	0.138
	(0.040)	(0.046)
	$[0.085,\ 0.241]$	$[0.048,\ 0.227]$

Notes. Maximum-likelihood estimates of the sequential threshold-crossing selection model in the main specification, with $\mathrm{Corr}(\varepsilon_1^s, \varepsilon_2^s)$ fixed at zero; standard errors in parentheses and 95% confidence intervals in brackets. $n = 3{,}361$. The matching-equation intercept is γ_{20}^s when $D^s = 0$ and γ_{21}^s when $D^s = 1$.

Under exogenous matching, the restrictions become substantially more informative. Under exogenous matching and monotone treatment response alone, all three effects are positive at the 95% level, with confidence intervals of $[0.15, 0.64]$ for $\mathbb{E}_F[Y(1,1) - Y(0,0)]$, $[0.04, 0.52]$ for $\mathbb{E}_F[Y(0,1) - Y(0,0)]$, and $[0.03, 0.57]$ for $\mathbb{E}_F[Y(1,0) - Y(0,0)]$. Adding additive separability tightens these bounds further.

Note that the correlation between ε_1^s and ε_2^s is not separately identified, and we fix it at zero in the main specification. Table 6 reports the bounds under additive separability and monotone treatment response when this correlation is instead fixed at several values on each side. The bounds are insensitive to these values, which indicates that the results are robust to the choice of correlation.

We also estimate the bounds with observed covariates, whose theoretical foundation is developed

Table 4: LATE among complier-complier pairs.

$LATE(1,1)$	$LATE(1,0)$	$LATE(0,1)$
0.562	0.278	0.284
(0.049)	(0.049)	(0.039)
[0.467, 0.658]	[0.182, 0.374]	[0.208, 0.360]

Notes. Estimates of the LATE among complier-complier pairs. Standard errors in parentheses and 95% confidence intervals in brackets. $n = 3{,}361$.

Table 5: Bound estimates and 95% confidence intervals on the ATE.

Additional restrictions	$\mathbb{E}_F[Y(1,1) - Y(0,0)]$	$\mathbb{E}_F[Y(0,1) - Y(0,0)]$	$\mathbb{E}_F[Y(1,0) - Y(0,0)]$
None	[−0.590, 0.675]	[−0.601, 0.643]	[−0.752, 0.689]
	[−0.618, 0.689]	[−0.620, 0.656]	[−0.766, 0.700]
MTR	[0.000, 0.675]	[0.000, 0.614]	[0.000, 0.660]
	[0.000, 0.688]	[0.000, 0.629]	[0.000, 0.673]
AS	[−0.244, 0.532]	[−0.027, 0.428]	[−0.265, 0.284]
	[−0.357, 0.638]	[−0.238, 0.526]	[−0.407, 0.810]
MTR +AS	[0.190, 0.532]	[0.077, 0.428]	[0.047, 0.277]
	[0.109, 0.636]	[0.000, 0.533]	[0.000, 0.472]
EXM	[−0.398, 0.598]	[−0.276, 0.488]	[−0.508, 0.561]
	[−0.438, 0.635]	[−0.318, 0.543]	[−0.545, 0.592]
EXM + MTR	[0.205, 0.598]	[0.087, 0.463]	[0.065, 0.531]
	[0.153, 0.638]	[0.044, 0.524]	[0.026, 0.566]
EXM + AS	[−0.179, 0.508]	[−0.070, 0.348]	[−0.217, 0.230]
	[−0.239, 0.595]	[−0.116, 0.423]	[−0.308, 0.344]
EXM + MTR + AS	[0.199, 0.508]	[0.112, 0.348]	[0.085, 0.230]
	[0.148, 0.595]	[0.054, 0.423]	[0.033, 0.344]

Notes. Each cell reports the point estimates of the bounds (top) and the 95% confidence interval (bottom). MTR is the monotone treatment response restriction, AS the additive separability of $e(d, d' \mid k, k')$, and EXM the exogenous matching restriction.

in Appendix F. The first covariate is an indicator for whether each parent was born in a major urban prefecture.[4] A second specification adds the academic test score of the parent's birthplace prefecture, which controls for the local level of academic achievement.[5] Table 7 reports the bounds under these specifications. Across both specifications, the effect of both parents attending college relative to neither remains positive and significant at the 95% level, and the overall pattern of the bounds matches the main specification.

Finally, we compare our estimates with two standard instrumental-variable LATE estimands, reported in Table 8. The first is the one-sided LATE, the estimand of the one-sided TSLS. Consider

[4] A major urban prefecture is one in the Tokyo metropolitan area or the Osaka metropolitan area.

[5] The prefecture test score is the average score on the 2007 National Assessment of Academic Ability of the National Institute for Educational Policy Research.

Table 6: Sensitivity of the ATE bounds to $\text{Corr}(\varepsilon_1^w, \varepsilon_2^w)$ and $\text{Corr}(\varepsilon_1^m, \varepsilon_2^m)$.

$\big(\text{Corr}(\varepsilon_1^w, \varepsilon_2^w),\ \text{Corr}(\varepsilon_1^m, \varepsilon_2^m)\big)$	$\mathbb{E}_F[Y(1,1) - Y(0,0)]$	$\mathbb{E}_F[Y(0,1) - Y(0,0)]$	$\mathbb{E}_F[Y(1,0) - Y(0,0)]$
$(-0.9, -0.9)$	–	–	–
$(-0.6, -0.6)$	[0.190, 0.460]	[0.113, 0.356]	[0.039, 0.171]
$(-0.3, -0.3)$	[0.189, 0.524]	[0.083, 0.422]	[0.047, 0.256]
$(0, 0)$	[0.190, 0.532]	[0.077, 0.428]	[0.047, 0.277]
$(0.3, 0.3)$	[0.191, 0.540]	[0.077, 0.429]	[0.046, 0.278]
$(0.6, 0.6)$	[0.194, 0.542]	[0.078, 0.430]	[0.046, 0.279]
$(0.9, 0.9)$	[0.202, 0.546]	[0.089, 0.426]	[0.047, 0.277]

Notes. Point estimates of the bounds under additive separability and monotone treatment response, with the correlation $\text{Corr}(\varepsilon_1^s, \varepsilon_2^s)$ fixed at the value shown on each side. The row $(0, 0)$ is the main specification. At $\text{Corr}(\varepsilon_1^w, \varepsilon_2^w) = \text{Corr}(\varepsilon_1^m, \varepsilon_2^m) = -0.9$, the linear program is infeasible, and the bounds are not reported $(-)$. This suggests that the parametric selection model is misspecified at that correlation.

Table 7: ATE bounds with covariates.

Restrictions	$\mathbb{E}_F[Y(1,1) - Y(0,0)]$	$\mathbb{E}_F[Y(0,1) - Y(0,0)]$	$\mathbb{E}_F[Y(1,0) - Y(0,0)]$
Urban origin			
AS + MTR	[0.182, 0.537]	[0.050, 0.439]	[0.045, 0.332]
	[0.081, 0.819]	[0.000, 0.528]	[0.000, 0.471]
EXM + AS + MTR	[0.201, 0.510]	[0.111, 0.345]	[0.084, 0.234]
	[0.133, 0.636]	[0.031, 0.444]	[0.014, 0.392]
Urban origin + test score			
AS + MTR	[0.156, 0.549]	[0.024, 0.418]	[0.034, 0.375]
	[0.059, 0.774]	[0.000, 0.543]	[0.000, 0.504]
EXM + AS + MTR	[0.178, 0.464]	[0.083, 0.214]	[0.086, 0.258]
	[0.090, 0.597]	[0.00, 0.577]	[0.000, 0.406]

Notes. Point estimates of the bounds (top) and 95% confidence intervals (bottom), under additive separability and monotone treatment response. EXM denotes the exogenous matching restriction. $n = 3{,}361$.

the maternal effect. It takes the household as the unit and treats the mother's college attendance as a single binary treatment with the selection model $D^w(z) = \mathbb{1}(\varepsilon_1^w \leq h_1^w(z;\gamma))$. The potential outcome is written as $Y(d^w) := Y(d^w, D^m)$, a function of the mother's treatment d^w that holds the father's treatment at its realized value D^m. Then, the estimated maternal effect is 0.508. Because the mother's and the father's college attendance are correlated through assortative matching, this estimand includes part of the father's effect. Therefore, it overstates the mother's own effect relative to our $LATE(1,0) = 0.278$. The paternal effect is defined symmetrically, with the roles of the mother and the father interchanged. Then, the estimated paternal effect is 0.377, which likewise overstates the father's own effect relative to our $LATE(0,1) = 0.284$.

The second is the two-sided LATE. It also takes the household as the unit. Each parent chooses

Table 8: Standard instrumental variable methods.

One-sided LATE		
Maternal effect	Paternal effect	
0.508	0.377	
(0.056)	(0.044)	
[0.399, 0.617]	[0.291, 0.464]	
Two-sided LATE		
$LATE(1,0)$	$LATE(0,1)$	$LATE(1,1)$
0.306	0.298	0.604
(0.056)	(0.042)	(0.056)
[0.197, 0.415]	[0.215, 0.381]	[0.495, 0.713]

Notes. The one-sided LATE is the estimand of the one-sided TSLS, which uses each parent's own instrument and treatment. The two-sided LATE uses both parents' instruments and is point-identified under joint independence. Standard errors in parentheses and 95% confidence intervals in brackets. $n = 3{,}361$.

their treatment status based on their own instrument, $D^w(z^w) = \mathbb{1}(\varepsilon_1^w \leq h_1^w(z^w;\gamma))$ and $D^m(z^m) = \mathbb{1}(\varepsilon_1^m \leq h_1^m(z^m;\gamma))$. Under the joint independence $(\varepsilon_1^w, \varepsilon_1^m) \perp\!\!\!\perp (Z^w, Z^m)$, the LATE among complier-complier pairs $\mathbb{E}_F[Y(d,d') - Y(0,0) \mid D^w(0) \neq D^w(1),\ D^m(0) \neq D^m(1)]$ can be point-identified. We obtain $LATE(1,1) = 0.604$, $LATE(1,0) = 0.306$, and $LATE(0,1) = 0.298$. Note that our LATE estimands are expected be overestimated since they are identified under conditional ignorability of the matching response. However, the two-sided LATE under joint independence is even larger at every pair, 0.604 versus our 0.562 for $LATE(1,1)$, 0.306 versus 0.278 for $LATE(1,0)$, and 0.298 versus 0.284 for $LATE(0,1)$.

8 Conclusion

This paper studies a heterogeneous treatment effect model in an endogenous one-to-one two-sided matching market. In the model, each agent selects their own treatment status and that of whom they match, after which the outcome of each matched pair is realized. A leading example is the causal effect of parental education on a child, where the schooling of each parent and the household formation are endogenous.

To identify causal parameters in this environment, we rely on instruments from both sides that shift the agents' choices of treatment and matching without directly affecting their child outcomes. When women and men match assortatively, it is difficult to justify the standard instrumental vari-

ables methods with treatments and instruments from both sides. In our setup, we provide point and partial identification results for various causal parameters, such as the average treatment effect. The key identification assumption is matching on observables, which requires that pairs are randomly matched conditional on observed characteristics and rules out assortative matching on unobservables. We also propose inference procedures for these causal parameters. The confidence intervals are uniformly valid over the identified set. Finally, we apply the framework to study the effect of parental education on the educational investment that a couple makes in their child, using Japanese household-level data.

References

Acerenza, S., Martinez-Iriarte, J., Sánchez-Becerra, A., & Spini, P. E. (2025). Bounds for within-household encouragement designs with interference. *arXiv preprint arXiv:2503.14314*.

Attanasio, O., Cattan, S., & Meghir, C. (2022). Early childhood development, human capital, and poverty. *Annual Review of Economics*, *14*, 853–892.

Bai, Y., Huang, S., Moon, S., Santos, A., Shaikh, A. M., & Vytlacil, E. J. (2024). Inference for treatment effects conditional on generalized principal strata using instrumental variables. *arXiv preprint arXiv:2411.05220*.

Balat, J. F., & Han, S. (2023). Multiple treatments with strategic substitutes. *Journal of Econometrics*, *234*(2), 732–757.

Balke, A., & Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, *92*(439), 1171–1176.

Björklund, A., & Salvanes, K. G. (2011). Education and family background: Mechanisms and policies. In *Handbook of the economics of education* (pp. 201–247, Vol. 3). Elsevier.

Black, S. E., & Devereux, P. J. (2011). Recent developments in intergenerational mobility. *Handbook of Labor Economics*, *4*, 1487–1541.

Black, S. E., Devereux, P. J., & Salvanes, K. G. (2005). Why the apple doesn't fall far: Understanding intergenerational transmission of human capital. *American Economic Review*, *95*(1), 437–449.

Breierova, L., & Duflo, E. (2004). The impact of education on fertility and child mortality: Do fathers really matter less than mothers? (Tech. rep.). National Bureau of Economic Research.

Cameron, S. V., & Heckman, J. J. (1998). Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of american males. *Journal of Political Economy*, *106*(2), 262–333.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling (tech. rep.). National Bureau of Economic Research.

Carneiro, P., Heckman, J. J., & Vytlacil, E. J. (2011). Estimating marginal returns to education. *American Economic Review*, *101*(6), 2754–2781.

Carneiro, P., Meghir, C., & Parey, M. (2013). Maternal education, home environments, and the development of children and adolescents. *Journal of the European Economic Association*, *11*(S1), 123–160.

Chevalier, A., Harmon, C., O'Sullivan, V., & Walker, I. (2013). The impact of parental income and education on the schooling of their children. *IZA Journal of Labor Economics*, *2*(1), 8.

Chiappori, P.-A. (2020). The theory and empirics of the marriage market. *Annual Review of Economics*, *12*(1), 547–578.

Chiappori, P.-A., Ciscato, E., & Guerriero, C. (2024). Analyzing matching patterns in marriage: Theory and application to Italian data. *Quantitative Economics*, *15*(3), 737–781.

Chiappori, P.-A., Costa Dias, M., & Meghir, C. (2018). The marriage market, labor supply, and education choice. *Journal of Political Economy*, *126*(S1), S26–S72.

Chiappori, P.-A., Salanié, B., & Weiss, Y. (2017). Partner choice, investment in children, and the marital college premium. *American Economic Review*, *107*(8), 2109–2167.

Choo, E., & Siow, A. (2006). Who marries whom and why. *Journal of Political Economy*, *114*(1), 175–201.

Chou, S.-Y., Liu, J.-T., Grossman, M., & Joyce, T. (2010). Parental education and child health: Evidence from a natural experiment in Taiwan. *American Economic Journal: Applied Economics*, *2*(1), 33–61.

Cunha, F., & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, *97*(2), 31–47.

Currie, J., & Moretti, E. (2003). Mother's education and the intergenerational transmission of human capital: Evidence from college openings. *The Quarterly Journal of Economics*, *118*(4), 1495–1532.

Fang, Z., Santos, A., Shaikh, A. M., & Torgovitsky, A. (2023). Inference for large-scale linear systems with known coefficients. *Econometrica*, *91*(4), 1337–1382.

Gafarov, B. (2025). Simple subvector inference on sharp identified set in affine models. *Journal of Econometrics*, *249*, 105952.

Galichon, A., Kominers, S. D., & Weber, S. (2019). Costly concessions: An empirical framework for matching with imperfectly transferable utility. *Journal of Political Economy*, *127*(6), 2875–2925.

Galichon, A., & Salanié, B. (2022). Cupid's invisible hand: Social surplus and identification in matching models. *The Review of Economic Studies*, *89*(5), 2600–2629.

Gu, J., & Russell, T. M. (2023). Partial identification in nonseparable binary response models with endogenous regressors. *Journal of Econometrics*, *235*(2), 528–562.

Gualdani, C., & Sinha, S. (2023). Partial identification in matching models for the marriage market. *Journal of Political Economy*, *131*(5), 1109–1171.

Han, S., & Yang, S. (2024). A computational approach to identification of treatment effects for policy evaluation. *Journal of Econometrics*, *240*(1), 105680.

Heckman, J. J., & Mosso, S. (2014). The economics of human development and social mobility. *Annual Review of Economics*, *6*, 689–733.

Heckman, J. J., & Pinto, R. (2018). Unordered monotonicity. *Econometrica*, *86*(1), 1–35.

Heckman, J. J., Urzua, S., & Vytlacil, E. (2008). Instrumental variables in models with multiple outcomes: The general unordered case. *Annales d'Economie et de Statistique*, 151–174.

Heckman, J. J., & Vytlacil, E. J. (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.

Heckman, J. J., & Vytlacil, E. J. (2001a). Instrumental variables, selection models, and tight bounds on the average treatment effect. In *Econometric evaluation of labour market policies* (pp. 1–15). Springer.

Heckman, J. J., & Vytlacil, E. J. (2001b). Policy-relevant treatment effects. *American Economic Review*, *91*(2), 107–111.

Heckman, J. J., & Vytlacil, E. J. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, *73*(3), 669–738.

Holmlund, H., Lindahl, M., & Plug, E. (2011). The causal effect of parents' schooling on children's schooling: A comparison of estimation methods. *Journal of Economic Literature*, *49*(3), 615–651.

Hoshino, T., & Yanagi, T. (2023). Treatment effect models with strategic interaction in treatment decisions. *Journal of Econometrics*, *236*(2), 105495.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475.

Kamat, V., Norris, S., & Pecenco, M. (2023). Identification in multiple treatment models under discrete variation. *arXiv preprint arXiv:2307.06174*.

Lafférs, L. (2019). Bounding average treatment effects using linear programming. *Empirical economics*, *57*(3), 727–767.

Lee, S., & Salanié, B. (2018). Identifying effects of multivalued treatments. *Econometrica*, *86*(6), 1939–1963.

Lee, S., & Salanié, B. (2020). Treatment effects with targeting instruments. *arXiv preprint arXiv:2007.10432*.

Lindeboom, M., Llena-Nozal, A., & van Der Klaauw, B. (2009). Parental education and child health: Evidence from a schooling reform. *Journal of Health Economics*, *28*(1), 109–131.

Lundborg, P., Nilsson, A., & Rooth, D.-O. (2014). Parental education and offspring outcomes: Evidence from the Swedish compulsory school reform. *American Economic Journal: Applied Economics*, *6*(1), 253–278.

Manski, C. F. (1997). Monotone treatment response. *Econometrica*, *65*(6), 1311–1334.

McCrary, J., & Royer, H. (2011). The effect of female education on fertility and infant health: Evidence from school entry policies using exact date of birth. *American Economic Review*, *101*(1), 158–195.

Mogstad, M., Santos, A., & Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, *86*(5), 1589–1619.

Mountjoy, J. (2022). Community colleges and upward mobility. *American Economic Review*, *112*(8), 2580–2630.

Navjeevan, M., Pinto, R., & Santos, A. (2026). Identification in instrumental variables models: The central role of Abadie's kappa. *Econometrica*, *94*(4), 1095–1133.

Oreopoulos, P., Page, M. E., & Stevens, A. H. (2006). The intergenerational effects of compulsory schooling. *Journal of Labor Economics*, *24*(4), 729–760.

Tsuda, T. (2022). Treatment effects with multidimensional unobserved heterogeneity: Identification of the marginal treatment effect. *arXiv preprint arXiv:2209.11444*.

Ura, T., & Zhang, L. (2024). Policy relevant treatment effects with multidimensional unobserved heterogeneity. *arXiv preprint arXiv:2403.13738*.

Van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes: With applications to statistics*. Springer.

Vazquez-Bare, G. (2023). Causal spillover effects using instrumental variables. *Journal of the American Statistical Association*, *118*(543), 1911–1922.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, *70*(1), 331–341.

A Simulation

This section assesses the finite-sample performance of the confidence intervals developed in Section 6. We generate data from a two-sided matching market whose primitives are known, and we report how often the confidence interval contains the true bounds $[\underline{ATE^*}_G, \overline{ATE^*}_G]$ across repeated samples.

A.1 Design

The selection follows the sequential threshold-crossing model of Example 5.1. For each $s \in \{w, m\}$, the treatment and matching response are

$$D^s(z) = \mathbb{1}(\varepsilon_1^s \le h_1^s(z)), \qquad T^s(z) = \mathbb{1}(\varepsilon_2^s \le h_2^s(D^s(z), z)),$$

with thresholds that are linear indices in the two-dimensional instrument $z = (z_1, z_2)$,

$$h_1^s(z) = \gamma_{10} + \gamma_{11} z_1 + \gamma_{12} z_2, \qquad h_2^s(d, z) = \gamma_{20}(1-d) + \gamma_{21}\, d + \gamma_{22} z_1 + \gamma_{23} z_2,$$

where the first subscript of γ labels the threshold equation and the second labels the coefficient. The second-stage intercept equals γ_{20} when $d = 0$ and γ_{21} when $d = 1$, and hence depends on the agent's own treatment d. We set the coefficients, common to the two sides, to

$$(\gamma_{10}, \gamma_{11}, \gamma_{12}, \gamma_{20}, \gamma_{21}, \gamma_{22}, \gamma_{23}) = (-2.25,\ 1.5,\ 0,\ -2.2,\ -1.7,\ 0,\ 1.3).$$

Since $\gamma_{21} > \gamma_{20}$, a treated agent is more inclined to match a treated partner. This generates positive assortative matching, that is, the woman's and the man's treatments are positively correlated. Each instrument $Z^s = (Z_1^s, Z_2^s)$ has two binary components and hence takes $2 \times 2 = 4$ support points. The unobserved shock $\varepsilon^s = (\varepsilon_1^s, \varepsilon_2^s)$ has independent standard normal components.

With the pair-specific shock ϕ^c in Remark 3.1, the potential outcome is

$$Y(d, d') = \pi\big(\phi^w(d, d'), \phi^m(d, d'), \phi^c\big) = \mathbb{1}\big(\phi^c \le \phi^w(d, d') + \phi^m(d, d')\big),$$

where ϕ^c is drawn independently across households from the uniform distribution on $[0, 1]$. The

potential factors are

$$\phi^w(d,d') = (0.5-\bar{\kappa})\,\Lambda(\upsilon_0+\upsilon_1\varepsilon_1^w+\upsilon_2\varepsilon_2^w)+\kappa^w(\varepsilon^w)\,\mathbb{1}(d=1),$$
$$\phi^m(d,d') = (0.5-\bar{\kappa})\,\Lambda(\upsilon_0+\upsilon_1\varepsilon_1^m+\upsilon_2\varepsilon_2^m)+\kappa^m(\varepsilon^m)\,\mathbb{1}(d'=1),$$

where Λ is the logistic function and the own-treatment effect of side s is

$$\kappa^s(\varepsilon^s) = \upsilon_3+\upsilon_4\,\Lambda(\upsilon_5\varepsilon_1^s+\upsilon_6\varepsilon_2^s)\in[\upsilon_3,\,\bar{\kappa}],\qquad \bar{\kappa}:=\upsilon_3+\upsilon_4.$$

Each potential factor depends only on the agent's own treatment, and the scale $0.5-\bar{\kappa}$ keeps $\phi^w(d,d')+\phi^m(d,d')\in[0,1]$.

We consider two designs, one in which the treatment effect is homogeneous across every combination of types and one in which it is heterogeneous. The location and scale of the index are common to the two designs, with $\upsilon_0=0$ and $\upsilon_1=\upsilon_2=0.3$, and the designs differ only in the own-treatment effect κ^s. Under the homogeneous design, we set $\upsilon_3=0.3$ and $\upsilon_4=0$, so that $\kappa^s(\varepsilon^s)=0.3$ is constant. Under the heterogeneous design, we set $\upsilon_3=0.2$, $\upsilon_4=0.2$, and $\upsilon_5=\upsilon_6=1$, so that $\kappa^s(\varepsilon^s)$ varies over $[0.2,\,0.4]$ with the shock ε^s. The two designs share the same average own-treatment effect $\mathbb{E}[\kappa^s(\varepsilon^s)]=0.3$, and hence the same true average treatment effects $\mathbb{E}_{F^*}[Y(1,0)-Y(0,0)]=0.3$ and $\mathbb{E}_{F^*}[Y(1,1)-Y(0,0)]=0.6$.

We consider two additional shape restrictions on the expected potential outcome $e(d,d'\mid k,k')$. The monotone treatment response requires the expected potential outcome to be nondecreasing in each treatment,

$$e(1,d'\mid k,k')\geq e(0,d'\mid k,k')\quad\text{and}\quad e(d,1\mid k,k')\geq e(d,0\mid k,k')\qquad\text{for every } d,d'\in\mathcal{D},\ k,k'\in\mathcal{K},$$

and the additive separability requires

$$e(d,d'\mid k,k') = e^w(d\mid k)+e^m(d'\mid k')\qquad\text{for every } d,d'\in\mathcal{D},\ k,k'\in\mathcal{K}.$$

Note that both restrictions hold in the present design. We report the four cases in which neither, either, or both of these restrictions are imposed in addition to the moment restrictions (5.3).

For each replication, we draw a sample of size n, estimate the first-stage parameter, and compute the confidence interval of Section 6 at the nominal level $1-\alpha=0.95$. We consider the sample sizes

$n \in \{3{,}000,\ 5{,}000,\ 10{,}000\}$. The regularization sequence is $\mu_n = c\,(\log n)/\sqrt{n}$ with $c = 10^{-6}$, which satisfies $\mu_n \to 0$ and $\mu_n\sqrt{n} \to \infty$.

A.2 Results

Table 9 reports, for each configuration, the fraction of replications whose confidence interval contains the population bounds $[\underline{ATE^*}_G, \overline{ATE^*}_G]$. The population bounds narrow as the shape restrictions are added, as the third column of the table shows. Across all configurations the coverage is close to or above the nominal level 0.95. Two features make the confidence interval conservative. To cover the population bounds, it controls the lower and upper endpoints separately and combines them through a Bonferroni union bound, and it applies the bias correction of Section 6. Coverage above 0.95 is therefore expected.

Table 9: Simulation results

Panel A. Homogeneous treatment effect

Additional restriction	(d, d')	$[\underline{ATE^*}_G, \overline{ATE^*}_G]$	$n = 3{,}000$	$n = 5{,}000$	$n = 10{,}000$
–	$(1,0)$	$[-0.575,\ 0.797]$	0.942	0.938	0.938
–	$(1,1)$	$[-0.395,\ 0.840]$	0.938	0.944	0.938
MTR	$(1,0)$	$[0.000,\ 0.761]$	0.976	0.978	0.970
MTR	$(1,1)$	$[0.021,\ 0.840]$	0.974	0.970	0.954
AS	$(1,0)$	$[0.016,\ 0.422]$	0.987	0.978	0.978
AS	$(1,1)$	$[0.156,\ 0.737]$	0.987	0.982	0.982
AS + MTR	$(1,0)$	$[0.086,\ 0.422]$	0.983	0.974	0.982
AS + MTR	$(1,1)$	$[0.238,\ 0.737]$	0.987	0.980	0.980

Panel B. Heterogeneous treatment effect

Additional restriction	(d, d')	$[\underline{ATE^*}_G, \overline{ATE^*}_G]$	$n = 3{,}000$	$n = 5{,}000$	$n = 10{,}000$
–	$(1,0)$	$[-0.555,\ 0.817]$	0.932	0.958	0.936
–	$(1,1)$	$[-0.409,\ 0.826]$	0.944	0.950	0.952
MTR	$(1,0)$	$[0.004,\ 0.753]$	0.948	0.964	0.970
MTR	$(1,1)$	$[0.020,\ 0.826]$	0.970	0.972	0.976
AS	$(1,0)$	$[0.004,\ 0.409]$	0.986	0.965	0.940
AS	$(1,1)$	$[0.142,\ 0.713]$	0.988	0.977	0.986
AS + MTR	$(1,0)$	$[0.085,\ 0.409]$	0.986	0.977	0.946
AS + MTR	$(1,1)$	$[0.238,\ 0.713]$	0.988	0.979	0.982

Notes. Each entry is the coverage frequency of the confidence interval, at the nominal level 0.95, computed over the feasible replications among 500 draws. The target (d, d') denotes $\mathbb{E}_{F^*}[Y(d, d') - Y(0, 0)]$. MTR denotes the monotone treatment response and AS denotes the additive separability of the expected potential outcome. A dash (–) indicates that only the moment restrictions of Section 5 are imposed. When AS and MTR are imposed together, the monotone treatment response is equivalent to $e^w(1 \mid k) \geq e^w(0 \mid k)$ and $e^m(1 \mid k') \geq e^m(0 \mid k')$ for every $k, k' \in \mathcal{K}$.

B On the joint independence condition

Standard instrumental variable methods for this setting include the two-sided two-stage least squares (e.g., Holmlund et al., 2011) and the causal models for multi-valued treatments with instruments (e.g., Heckman and Pinto, 2018; Lee and Salanié, 2018; Kamat et al., 2023; Bai et al., 2024), both of which treat the two parents' education as jointly endogenous and use the two parents' instruments together. The latter, in particular, requires the household-level joint independence $(D^w(\cdot), D^m(\cdot), Y(\cdot,\cdot)) \perp\!\!\!\perp (Z^w, Z^m)$ under F. This appendix shows that this condition is generically incompatible with the model of endogenous treatment selection and matching, and then applies the result to the numerical example of Section 2.

Taking the selection model $\{(D^s(z), T^s(z))\}_z$ as given, we have the following result.

Result B.1. *Suppose that F is an equilibrium matching that satisfies the joint independence of the instruments $(D^w(\cdot), D^m(\cdot), Y(\cdot,\cdot)) \perp\!\!\!\perp (Z^w, Z^m)$. Then*

$$\mathrm{Cov}_F\big(\mathbb{E}_F[D^w \mid Z^w],\, \mathbb{E}_F[T^w \mid Z^w]\big) = \mathrm{Cov}_F\big(\mathbb{E}_F[D^m \mid Z^m],\, \mathbb{E}_F[T^m \mid Z^m]\big). \tag{B.1}$$

Each side of (B.1) is the covariance, across values of the instrument, between the propensity scores of the own treatment D^s and of the partner's treatment T^s, that is, $\mathbb{E}_F[D^s \mid Z^s]$ and $\mathbb{E}_F[T^s \mid Z^s]$. It is positive when an instrument value that makes the agent more likely to be treated also makes her more likely to be matched with a treated partner. The left-hand side is a feature of the woman's selection and the right-hand side of the man's, and by Result B.1, the joint independence can hold only if the two coincide. This coincidence is non-generic, even under equilibrium matching. One leading case in which it holds is random matching, where the matching choice is independent of the agent's own treatment and instrument, so that $\mathbb{E}_F[T^s \mid Z^s]$ is constant and both covariances vanish. Except for such cases, the two covariances are different and the equality fails generically. Therefore, the standard instrumental variable method, which relies on the joint independence, is difficult to justify in the present setting. Our approach rests on the marginal independence (Assumption 3.2) and the matching on observables condition (Assumption 3.3), which are interpretable and always compatible with the model.

The argument above takes the mother's treatment response to be a function of her own instrument alone, $D^w = D^w(Z^w)$, with the father's instrument Z^m not entering it, and likewise for the father. This is plausible in the present setting. The mother chooses her own treatment before the

household is formed through matching, hence her schooling cannot respond to the instrument of a father she has not yet matched with. Applying the multi-valued treatment approach as it stands, by contrast, would let $D^w(\cdot)$ take Z^m as an argument. Allowing this is unrealistic for the same reason. It would also make the usual restrictions on treatment response types (e.g., monotonicity or a separable selection equation) hard to impose, since they would then have to govern how the mother's own treatment responds to the father's instrument.

We now apply Result B.1 to the numerical example of Section 2. The design sets $p^s(0)+p^s(1) = 1$ on each side, under which the market clears, that is, demand and supply coincide, $\mathbb{P}_F(D^w = d, T^w = d') = \mathbb{P}_F(D^m = d', T^m = d)$ for every d, d'. Each side of (B.1) is $(2q-1)\operatorname{Var}_F(p^s(Z^s)) = \frac{1}{4}(2q-1)(p^s(1)-p^s(0))^2$. At $q = \frac{1}{2}$, the matching is random, and both covariances coincide. For $q > \frac{1}{2}$, the two sides coincide only when the two instruments are equally strong, $(p^w(1)-p^w(0))^2 = (p^m(1)-p^m(0))^2$. Therefore, for every $q > \frac{1}{2}$, no equilibrium matching F satisfies the joint independence under the present design, where $(p^w(0), p^w(1)) = (0.2, 0.8)$ and $(p^m(0), p^m(1)) = (0.4, 0.6)$. Note that the equality of instrument strengths $(p^w(1)-p^w(0))^2 = (p^m(1)-p^m(0))^2$ is not generically satisfied even within the class of market-clearing designs (i.e., $p^s(0)+p^s(1) = 1$ on each side).

C Pair-specific shocks

Remark 3.1 allows the outcome to depend on a pair-specific shock ϕ^c in addition to the two sides' potential factors. This appendix shows that all of our identification results continue to hold in this case. Throughout, ϕ^c is a scalar shock, and hence π is a Borel function from $\mathbb{R}^3$ to $\mathcal{Y}$.

C.1 Theorem 4.1

First, we consider the LATE among complier-complier pairs under the assumption that matching response is conditionally ignorable. With the pair-specific shock, the LATE is now defined as

$$LATE(d, d') := \mathbb{E}_{\tilde{F}^*}\big[Y(d, d') - Y(0, 0) \mid D^w(1) \neq D^w(0),\, D^m(1) \neq D^m(0)\big],$$

where now $Y(d, d') = \pi(\phi^w(d, d'), \phi^m(d, d'), \phi^c)$.

The derivation of Theorem 4.1 relates this LATE parameter to the observed conditional means $e_{dd'}(z, z') = \mathbb{E}[Y \mid D^w = d, D^m = d', Z^w = z, Z^m = z']$ and the propensity scores $p^w(\cdot), p^m(\cdot)$. Because ϕ^c is independent of A, each step of that derivation goes through with $\pi(\phi^w(d, d'), \phi^m(d, d'))$ replaced by $\pi(\phi^w(d, d'), \phi^m(d, d'), \phi^c)$, while the observed means $e_{dd'}(z, z')$ and the propensity scores are unchanged. Therefore, $LATE(d, d')$ is identified by the same formula as in Theorem 4.1.

C.2 Theorem 5.1

Next, we consider partial identification of the average treatment effect under the independent coupling $\tilde{F}^*$, namely $\mathbb{E}_{\tilde{F}^*}[Y(d, d') - Y(0, 0)]$. In this setting, the identified set is now redefined as

$$\Theta_I := \big\{(F, F^c_\phi, \pi) : \text{Assumptions 3.1, 3.2, and 3.3 hold under } G \text{ and } (F, F^c_\phi, \pi)\big\}.$$

Then the lower and upper bounds of the ATE under $\tilde{F}^*$ can be expressed as

$$\min_{(F, F^c_\phi, \pi) \in \Theta_I} \mathbb{E}_{\tilde{F}^*}[Y(d, d') - Y(0, 0)] = \min_{(e,p) \in \mathbf{E} \times \mathbf{P}} ATE^*(e, p) \text{ subject to (5.3)–(5.4)},$$

$$\max_{(F, F^c_\phi, \pi) \in \Theta_I} \mathbb{E}_{\tilde{F}^*}[Y(d, d') - Y(0, 0)] = \max_{(e,p) \in \mathbf{E} \times \mathbf{P}} ATE^*(e, p) \text{ subject to (5.3)–(5.4)},$$

for every $\tilde{d}, \tilde{d}' \in \mathcal{D}$, $z, z' \in \mathcal{Z}$, and $s, s' \in \{w, m\}$ with $s \neq s'$, where now $e(d, d' \mid k, k')$ intuitively represents $\mathbb{E}_{\tilde{F}^*}[Y(d, d') \mid K^w = k, K^m = k']$.

The proof is almost the same as that of Theorem 5.1. The only step that changes is Step 3 of that proof (provided in Appendix G). Now, for a given feasible (e, p), one constructs $(F, F^c_\phi, \pi) \in \Theta_I$ that induces (e, p) instead of constructing (F, π). Because $\mathcal{K} \times \mathcal{D} \times \mathcal{D}$ is finite, fix a Borel injection $\alpha : \mathcal{K} \times \mathcal{D} \times \mathcal{D} \to \mathbb{R}$, and set $\phi^w(d, d') = \alpha(K^w, d, d')$ and $\phi^m(d, d') = \alpha(K^m, d, d')$. Let ϕ^c be uniform on $[0, 1]$, and define

$$\bar{\pi}(x, x') = \begin{cases} e(d, d' \mid k, k') & \text{if } \alpha^{-1}(x) = (k, d, d') \text{ and } \alpha^{-1}(x') = (k', d, d'), \\ 0 & \text{otherwise,} \end{cases}$$

$$\pi(x, x', c) = \mathbb{1}\{c \leq \bar{\pi}(x, x')\}.$$

Note that $\bar{\pi}$ is Borel measurable because α has finite range. Then, $Y(d, d') = \mathbb{1}\{\phi^c \leq e(d, d' \mid K^w, K^m)\}$ and $\mathbb{E}_{\tilde{F}^*}[Y(d, d') \mid k, k'] = e(d, d' \mid k, k')$. Setting the type distribution to p and taking F to be the corresponding equilibrium matching that satisfies matching on observables, as in Step 3 of the proof of Theorem 5.1, gives $(F, F^c_\phi, \pi) \in \Theta_I$ that induces (e, p). Therefore the bounds are sharp. The shock ϕ^c provides the randomness that turns the mean $\bar{\pi} = e$ into a binary outcome, and π takes the simple threshold-crossing form above.

D Foundation of selection models

This appendix provides the micro-foundation for the equilibrium matching (Assumption 3.3) under the two selection models used in the paper, namely the sequential threshold-crossing model (Remark 4.2 and Example 5.1) and the additive random utility model (Example 5.2). In each case, the systematic utilities are determined so that demand and supply in the marriage market coincide, as in a transferable-utility matching model in the spirit of Galichon and Salanié, 2022. For both models, we establish the existence and uniqueness of an aggregate equilibrium and construct the induced equilibrium matching. The two subsections treat the two models in turn.

D.1 Sequential threshold-crossing model

In the sequential threshold-crossing model of Remark 4.2 and Example 5.1, each agent first chooses her own treatment and then, given it, the treatment status of her matched partner.

D.1.1 Setup

We index objects by $(d, d') \in \mathcal{D}^2$, a household whose woman has treatment d and whose man has treatment d'. Households form in a matching market with transferable utility. The idiosyncratic surplus of a match between woman i with $D_i^w = d$ and man j with $D_j^m = d'$ is

$$\tilde{\Psi}_{ij} = \Psi_{dd'} + \Big(-C^w_{dd'}(Z_i^w) - \mathbb{1}(d=1)\varepsilon^w_{1,i} - \mathbb{1}(d'=1)\varepsilon^w_{2,i}\Big) + \Big(-C^m_{dd'}(Z_j^m) - \mathbb{1}(d'=1)\varepsilon^m_{1,j} - \mathbb{1}(d=1)\varepsilon^m_{2,j}\Big),$$

where $\Psi_{dd'} \in \mathbb{R}$ is the systematic total surplus of a (d, d') household, $C^s_{dd'}(Z^s)$ are instrument-dependent costs, and $\varepsilon^s = (\varepsilon^s_1, \varepsilon^s_2)$ are the taste shocks. Throughout this section, we consider that the distribution of $(Z^s, \boldsymbol{\phi}^s, \varepsilon^s)$ is exogenously given for each $s \in \{w, m\}$.

We describe the equilibrium through the matching mass μ and the surplus shares $V^w_{dd'}, V^m_{dd'}$, which split the total surplus, $V^w_{dd'} + V^m_{dd'} = \Psi_{dd'}$. We normalize the level of total surplus and of each side's shares by setting $\Psi_{00} = V^w_{00} = V^m_{00} = 0$, so that $V^s = (V^s_{01}, V^s_{10}, V^s_{11}) \in \mathbb{R}^3$. The matching mass is denoted likewise by $\mu = (\mu_{01}, \mu_{10}, \mu_{11}) \in \mathbb{R}^3$, the mass μ_{00} of households with $(0,0)$ being one minus their sum. Setting $\Psi_{00} = 0$ is without loss of generality, since only differences of Ψ enter the analysis. The remaining normalization $V^w_{00} = 0$ (hence $V^m_{00} = \Psi_{00} - V^w_{00} = 0$) fixes the one common shift $(V^w, V^m) \mapsto (V^w + c\mathbf{1}, V^m - c\mathbf{1})$ with $\mathbf{1} = (1,1,1,1)$ that leaves all choice probabilities unchanged.

The utility of a woman with instrument $Z^w = z$ from forming a household with (d, d') is

$$U^w_{dd'}(z) - \mathbb{1}(d = 1)\,\varepsilon^w_1 - \mathbb{1}(d' = 1)\,\varepsilon^w_2, \qquad U^w_{dd'}(z) := V^w_{dd'} - C^w_{dd'}(z),$$

whose systematic part $U^w_{dd'}(z)$ is the surplus share net of the instrument-dependent cost. The man's utility is analogous. We normalize the cost so that $C^s_{00}(z) = 0$, hence $U^s_{00}(z) = 0$, treating $U^s(z)$ as an element of $\mathbb{R}^3$ (the choice depends only on utility differences, which makes this without loss of generality).

Each woman first observes ε^w_1 and chooses D^w, and then observes ε^w_2 and chooses T^w. Solving backward, given $D^w = d$, she chooses $T^w = 1$ if and only if

$$\varepsilon^w_2 \le U^w_{d1}(z) - U^w_{d0}(z) =: \varsigma^w_d(z),$$

thus $\varsigma^w_d(z)$ is the second-stage threshold given own treatment d. For a utility vector $U = (U_{01}, U_{10}, U_{11}) \in \mathbb{R}^3$, the first-stage continuation value of choosing $D^w = d$ after observing $\varepsilon^w_1 = e$ is

$$\Omega^w_d(U; e) := \mathbb{E}_{\varepsilon^w_2}\left[\max\left\{U_{d0},\, U_{d1} - \varepsilon^w_2\right\} \,\middle|\, \varepsilon^w_1 = e\right],$$

and hence she chooses $D^w = 1$ if and only if $\Omega^w_1(U^w; \varepsilon^w_1) - \varepsilon^w_1 \ge \Omega^w_0(U^w; \varepsilon^w_1)$. The problem of each man is symmetric.

Then, we make the following assumptions on the shock $\varepsilon^s = (\varepsilon^s_1, \varepsilon^s_2)$.

Assumption D.1. *For each $s \in \{w, m\}$,*

(i) ε^s has a continuous density that is strictly positive on $\mathbb{R}^2$, and $\mathbb{E}\|\varepsilon^s\| < \infty$

(ii) $(\varepsilon^s, \boldsymbol{\phi}^s) \perp\!\!\!\perp Z^s$

(iii) $\varepsilon^s_2 = \Upsilon^s(\varepsilon^s_1) + \nu^s$, where $\nu^s \perp\!\!\!\perp \varepsilon^s_1$ and Υ^s is Lipschitz with constant $C^s_{\text{Lip}} < 1$.

Condition (i) imposes continuity of the shock distribution. Condition (ii) requires that the instrument is independent of the shock and the potential factors. Condition (iii) expresses the second stage shock as a location shift of the first stage shock that is a contraction. For example, when we consider a bivariate normal $(\varepsilon^s_1, \varepsilon^s_2)$, it requires that $|\sigma^s_{DT}|\,\sigma^s_T/\sigma^s_D < 1$, where σ^s_D, σ^s_T are the standard deviations of $\varepsilon^s_1, \varepsilon^s_2$ and σ^s_{DT} is their correlation. As shown below, it makes the first-stage

choice almost surely unique at every U, and it also makes the first-stage decision a threshold-crossing rule. Independence of ε_1^s and ε_2^s is the special case where Υ^s is a constant function.

D.1.2 Existence and uniqueness of the equilibrium

Let $\varrho_{dd'}^s(U)$ denote the probability that the optimal two-stage choice of a side-s agent with utility vector $U \in \mathbb{R}^3$ results in a household with (d, d'). The aggregate matching masses are

$$\mu_{dd'}^w(V^w) := \mathbb{E}_{Z^w}\left[\varrho_{dd'}^w\left(V^w - C^w(Z^w)\right)\right], \qquad \mu_{dd'}^m(V^m) := \mathbb{E}_{Z^m}\left[\varrho_{dd'}^m\left(V^m - C^m(Z^m)\right)\right].$$

We write $\mu^s(V^s) := \left(\mu_{01}^s(V^s), \mu_{10}^s(V^s), \mu_{11}^s(V^s)\right) \in \mathbb{R}^3$. Then we define the aggregate equilibrium as follows.

Definition D.1 (Aggregate equilibrium)**.** *A tuple (μ, V^w, V^m) is an aggregate equilibrium if, for $(d, d') \neq (0, 0)$,*

(i) $V_{dd'}^w + V_{dd'}^m = \Psi_{dd'}$, *and*

(ii) $\mu_{dd'} = \mu_{dd'}^w(V^w) = \mu_{dd'}^m(V^m)$.

Condition (i) is a feasibility condition, which requires that the sum of the (deterministic) women's and men's shares of a (d, d') household equals its total surplus $\Psi_{dd'}$. Condition (ii) is a market clearing condition, which requires that the surplus shares V^w, V^m are determined so that the mass of women who choose (d, d') equals the mass of men who form such a household. Thus, the surplus shares play the role of prices that clear the matching market.

For a utility vector $U \in \mathbb{R}^3$, let $H^w(U)$ be the ex-ante value of the woman's two-stage problem,

$$H^w(U) := \mathbb{E}_{\varepsilon_1^w}\left[\max\left\{\Omega_0^w(U; \varepsilon_1^w),\, \Omega_1^w(U; \varepsilon_1^w) - \varepsilon_1^w\right\}\right],$$

and define H^m analogously. Define, for $s \in \{w, m\}$, $\bar{H}^s(V^s) := \mathbb{E}_{Z^s}[H^s(V^s - C^s(Z^s))]$.

Then, we can show that market clearing is equivalent to minimizing the convex function $\Phi(V) := \bar{H}^w(V) + \bar{H}^m(\Psi - V)$ over the (woman's) surplus shares $V \in \mathbb{R}^3$. The function Φ is strictly convex and has a unique minimizer, at which the two sides' choice probabilities coincide. This yields the following result.

Proposition D.1. *Under Assumption D.1, there exists a unique aggregate equilibrium. Moreover, each agent's optimal two-stage policy is a sequential threshold-crossing rule of the form (4.1) in Remark 4.2.*

The first part follows because Φ has a unique minimizer, and the second part follows from the backward-induction solution of the two-stage problem under Assumption D.1. The normalization $V^w_{00} = 0$ removes the one common shift. Under this normalization, the equilibrium shares V are unique. The matching mass μ is unique regardless of this normalization. The function Φ is (minus) the regularized matching surplus in the sense of Galichon and Salanié, 2022: $\bar{H}^s$ is the convex conjugate of the side-s generalized entropy induced by the shock structure, and the equilibrium minimizes Φ. The only novelty relative to the cited framework is that the side-s entropy is generated by a nested (two-stage) rather than a one-shot discrete choice, which still yields a strictly convex function with the choice probability as its gradient.

D.1.3 Constructing F

We now construct an equilibrium matching from this policy at the unique aggregate equilibrium (μ, V^w, V^m) of Proposition D.1. Define the value functions $U^s_{dd'}(z) = V^s_{dd'} - C^s_{dd'}(z)$. Let $K^s(z)$ be the two-stage optimal policy mapping (Z^s, ε^s) into (D^s, T^s), which is consistent with the sequential threshold crossing decision above. Let F^w be the joint distribution of $(D^w, T^w, Z^w, K^w, \boldsymbol{\phi}^w)$ with $(D^w, T^w) = K^w(Z^w)$, and define F^m analogously. For each $(d, d') \in \mathcal{D}^2$, let $F_{A|dd'}$ be the independent coupling of $F^w_{A^w|D^w=d,\,T^w=d'}$ and $F^m_{A^m|D^m=d',\,T^m=d}$, and, setting $\mu_{00} := 1 - \mu_{01} - \mu_{10} - \mu_{11}$ for the remaining mass, define

$$F := \sum_{d,d' \in \mathcal{D}} \mu_{dd'}\, F_{A|dd'}.$$

Then, we have the following result.

Proposition D.2. *Under Assumption D.1, the distribution F constructed above satisfies the independence of instruments (i.e., Assumption 3.2) and is an equilibrium matching with the matching on observables condition (i.e., Assumption 3.3). Moreover the induced selection model (i.e., the marginal on K^w and K^m under F) is the sequential threshold-crossing model of the form (4.1) in Remark 4.2.*

Proposition D.2 gives a micro-foundation for the two restrictions on the distribution F, that is, F is an equilibrium matching (demand and supply in the marriage market coincide) and the

selection under F has the sequential threshold-crossing property. The unobservables $(\varepsilon_1^s, \varepsilon_2^s)$ are the primitive taste shocks of the structural model, whereas the thresholds are equilibrium objects. $h_1^s(z)$ and $h_2^s(z,d)$ are determined by the utilities $U_{dd'}^s(z) = V_{dd'}^s - C_{dd'}^s(z)$, in which the surplus shares V^s adjust so that the demand and supply coincide. The threshold-crossing selection is therefore the reduced form of a matching model with transferable utility, in which the market itself sets the thresholds.

D.2 Additive random utility model

In the additive random utility model of Example 5.2, each agent instead chooses the pair (D^s, T^s) in a single maximization problem. The analysis parallels the sequential case.

D.2.1 Setup

We index objects by $(d, d') \in \mathcal{D}^2$, a household whose woman has treatment d and whose man has treatment d'. As in the previous section, the deterministic total surplus $\Psi_{dd'} \in \mathbb{R}$ of a (d, d') household is split into the equilibrium shares $V_{dd'}^w$ and $V_{dd'}^m$ in a transferable-utility matching market, entering the idiosyncratic match surplus additively with the instrument-dependent costs $C_{dd'}^s(Z^s)$ and the taste shocks. The marginal on $(Z^s, \boldsymbol{\phi}^s, \varepsilon^s)$ is also exogenously given for each $s \in \{w, m\}$.

The only difference is that the shock structure is more general. Each person draws a four-dimensional shock $\varepsilon^s = (\varepsilon_{dd'}^s)_{(d,d')\in\mathcal{D}^2} \in \mathbb{R}^4$, that is, one taste shock per (d, d') pair rather than shocks for each stage $(\varepsilon_1^s, \varepsilon_2^s)$. Therefore, the idiosyncratic match surplus is

$$\tilde{\Psi}_{ij} = \Psi_{dd'} - C_{dd'}^w(Z_i^w) - C_{dd'}^m(Z_j^m) + \varepsilon_{i,dd'}^w + \varepsilon_{j,dd'}^m.$$

As in Section D.1, we normalize $\Psi_{00} = V_{00}^w = V_{00}^m = 0$, so that each $V^s = (V_{01}^s, V_{10}^s, V_{11}^s) \in \mathbb{R}^3$ and the matching mass $\mu = (\mu_{01}, \mu_{10}, \mu_{11}) \in \mathbb{R}^3$ with μ_{00} one minus their sum; as explained there, this is without loss of generality and fixes the one common shift that leaves all choice probabilities unchanged. We further normalize the cost so that $C_{00}^s(z) = 0$ since the argmax in (D.1)–(D.2) depends only on differences of $V_{ll'}^s - C_{ll'}^s(z)$.

Then, we make the following assumption on the shock vector.

Assumption D.2. *For each $s \in \{w, m\}$, the shock vector $\varepsilon^s = (\varepsilon_{dd'}^s)_{(d,d')\in\mathcal{D}^2}$ satisfies:*

(i) $\mathbb{E}\|\varepsilon^s\| < \infty$.

(ii) *For every $V^s \in \mathbb{R}^3$, the argmax of $\{V^s_{ll'} - C^s_{ll'}(Z^s) + \varepsilon^s_{ll'}\}_{(l,l')\in\mathcal{D}^2}$ is unique with probability one over the joint law of (Z^s, ε^s).*

(iii) *For any $V_1, V_2 \in \mathbb{R}^3$ with $V_1 \neq V_2$, the choice probabilities differ, that is, $\mu^s(V_1) \neq \mu^s(V_2)$.*

(iv) $(\varepsilon^s, \boldsymbol{\phi}^s) \perp\!\!\!\perp Z^s$.

Condition (i) requires that the shock has a finite mean. Condition (ii) requires that the utility-maximizing choice is almost surely unique, so that the choice probability is well-defined. Condition (iii) requires that distinct surplus shares induce distinct choice probabilities on $\mathbb{R}^3$. Condition (iv) requires that the instrument is independent of the shock and the potential factors.

D.2.2 Existence and uniqueness of the equilibrium

For $(d, d') \neq (0, 0)$, the side-s choice probability is the multinomial argmax probability

$$\mu^w_{dd'}(V^w) = \mathbb{P}\left((d, d') = \underset{(l,l')\in\mathcal{D}^2}{\operatorname{argmax}}\left\{V^w_{ll'} - C^w_{ll'}(Z^w) + \varepsilon^w_{ll'}\right\}\right), \tag{D.1}$$

$$\mu^m_{dd'}(V^m) = \mathbb{P}\left((d, d') = \underset{(l,l')\in\mathcal{D}^2}{\operatorname{argmax}}\left\{V^m_{ll'} - C^m_{ll'}(Z^m) + \varepsilon^m_{ll'}\right\}\right), \tag{D.2}$$

and we write $\mu^s(V^s) := \left(\mu^s_{01}(V^s), \mu^s_{10}(V^s), \mu^s_{11}(V^s)\right) \in \mathbb{R}^3$. We index every object—including the man's cost C^m and shock ε^m—by (d, d'), where d is the woman's and d' the man's treatment; a man in a household with (d, d') has own treatment d' and partner treatment d, i.e. $(D^m, T^m) = (d', d)$, and hence the two sides' choice problems are written symmetrically. An aggregate equilibrium is then exactly as in Definition D.1 of the previous section: a tuple (μ, V^w, V^m) with feasibility $V^w_{dd'} + V^m_{dd'} = \Psi_{dd'}$ and market clearing $\mu_{dd'} = \mu^w_{dd'}(V^w) = \mu^m_{dd'}(V^m)$ for $(d, d') \neq (0, 0)$.

For each $s \in \{w, m\}$, define the social surplus function

$$\bar{H}^s(V^s) := \mathbb{E}\left[\max_{(l,l')\in\mathcal{D}^2}\left\{V^s_{ll'} - C^s_{ll'}(Z^s) + \varepsilon^s_{ll'}\right\}\right].$$

Then, we have the following existence and uniqueness result.

Proposition D.3. *Under Assumption D.2, there exists a unique aggregate equilibrium.*

This follows because Φ is strictly convex with a unique minimizer, at which the two sides' choice probabilities coincide. As in the sequential case, Φ is (minus) the regularized matching surplus in

the sense of Galichon and Salanié, 2022, with $\bar{H}^s$ the convex conjugate of the side-s generalized entropy induced by the shock structure. The difference from the cited framework is that each agent chooses the pair (own treatment, partner treatment), so that the own treatment is itself part of the discrete choice rather than a given type.

D.2.3 Constructing F

We now construct an equilibrium matching from the equilibrium policy, as in the case of sequential choice. Given the unique aggregate equilibrium (μ, V^w, V^m), with the utilities $U^s_{dd'}(z) := V^s_{dd'} - C^s_{dd'}(z)$, let each side's optimal (d, d') be

$$K^s(z) := \underset{(d,d')\in\mathcal{D}^2}{\operatorname{argmax}} \left\{ U^s_{dd'}(z) + \varepsilon^s_{dd'} \right\}.$$

A woman whose optimizer K^w is (d, d') has $(D^w, T^w) = (d, d')$, while a man whose optimizer K^m is (d, d') has $(D^m, T^m) = (d', d)$. Let F^w be the joint distribution of $(D^w, T^w, Z^w, K^w, \boldsymbol{\phi}^w)$, and let F^m be defined analogously. For each $(d, d') \in \mathcal{D}^2$, let $F_{A|dd'}$ be the independent coupling of $F^w_{A^w|D^w=d,\,T^w=d'}$ and $F^m_{A^m|D^m=d',\,T^m=d}$. Setting $\mu_{00} := 1 - \mu_{01} - \mu_{10} - \mu_{11}$ for the remaining mass, define

$$F := \sum_{d,d'\in\mathcal{D}} \mu_{dd'}\, F_{A|dd'}.$$

Proposition D.4. *Under Assumption D.2, the distribution F constructed above satisfies the independence of instruments (i.e., Assumption 3.2) and is an equilibrium matching with the matching on observables condition (i.e., Assumption 3.3). Moreover the induced selection model (i.e., the marginal on K^w and K^m under F) is the additive random utility model of Example 5.2.*

Proposition D.4 gives the analogous micro-foundation for the additive random utility model of Example 5.2. As in Proposition D.2, the unobservables ε^s are the primitive taste shocks, whereas the utilities $U^s_{dd'}(z) = V^s_{dd'} - C^s_{dd'}(z)$ are equilibrium objects, and hence the selection is the reduced form of a matching model with transferable utility.

We have developed the foundation under transferable utility. This is for concreteness rather than necessity: the same construction extends, in the spirit of Galichon et al., 2019, to the more general imperfectly transferable utility matching, which nests the fully transferable, nontransferable, and collective models as special cases. Under their regularity conditions, the aggregate equilibrium

remains unique, and the induced F satisfies the independence of instruments and the matching on observables condition exactly as above.

E Target parameters

In Section 5, we defined ATE^*, the ATE under random matching, as the primary causal parameter of interest and showed that its bounds can be computed via finite-dimensional optimization over $(e,p) \in \mathbf{E} \times \mathbf{P}$. The same approach applies to other causal parameters. Specifically, by replacing the objective function $ATE^*(e,p)$ in Theorem 5.1 with the objective functions listed below, we can compute the bounds for each of the following causal parameters while maintaining the same constraint structure. The constraints (5.3)–(5.4) remain unchanged because they characterize the relationship between the observed data distribution and the (e,p), independently of the choice of the causal parameter.

Below, we present four additional causal parameters. The first, ATE, measures the unconditional ATE under the equilibrium matching F, as opposed to the random matching F^* used in ATE^*. The next two, ATT and $ATNT$, measure the ATE under the equilibrium matching F, conditional on the observed treatment statuses. The fourth, $GATE^*$, measures the ATE for a subpopulation defined by specific treatment-matching response types under random matching F^*.

Average treatment effect under equilibrium matching

The ATE under the equilibrium matching F:

$$\mathbb{E}_F\left[Y(d,d') - Y(0,0)\right].$$

This can be expressed as

$$ATE(e,p) = \sum_{\tilde{d},\tilde{d}' \in \mathcal{D}} \sum_{z,z' \in \mathcal{Z}} \sum_{k \in \mathcal{K}(\tilde{d},\tilde{d}'|z)} \sum_{k' \in \mathcal{K}(\tilde{d}',\tilde{d}|z')} \mathbb{P}\left(D^w = \tilde{d}, D^m = \tilde{d}', Z^w = z, Z^m = z'\right) \times \frac{p_k^w p_{k'}^m \{e(d,d' \mid k,k') - e(0,0 \mid k,k')\}}{\mathbb{P}\left(D^w = \tilde{d}, D^m = \tilde{d}' \mid Z^w = z\right) \mathbb{P}\left(D^w = \tilde{d}, D^m = \tilde{d}' \mid Z^m = z'\right)}.$$

Average treatment effect on the treated

The ATE conditional on $(D^w, D^m) = (d,d')$ under the equilibrium matching F:

$$\mathbb{E}_F\left[Y(d,d') - Y(0,0) \mid D^w = d, D^m = d'\right].$$

This can be expressed as

$$ATT(e,p) = \sum_{z,z' \in \mathcal{Z}} \sum_{k \in \mathcal{K}(d,d'|z)} \sum_{k' \in \mathcal{K}(d',d|z')} \{e(d,d' \mid k,k') - e(0,0 \mid k,k')\} \times p_k^w p_{k'}^m \frac{\mathbb{P}\left(Z^w = z, Z^m = z' \mid D^w = d, D^m = d'\right)}{\mathbb{P}\left(D^w = d, D^m = d' \mid Z^w = z\right) \mathbb{P}\left(D^w = d, D^m = d' \mid Z^m = z'\right)}.$$

Average treatment effect on the non-treated

The ATE conditional on $(D^w, D^m) = (0,0)$ under the equilibrium matching F:

$$\mathbb{E}_F\left[Y(d,d') - Y(0,0) \mid D^w = 0, D^m = 0\right].$$

This can be expressed as

$$ATNT(e,p) = \sum_{z,z' \in \mathcal{Z}} \sum_{k \in \mathcal{K}(0,0|z)} \sum_{k' \in \mathcal{K}(0,0|z')} \{e(d,d' \mid k,k') - e(0,0 \mid k,k')\} \times p_k^w p_{k'}^m \frac{\mathbb{P}\left(Z^w = z, Z^m = z' \mid D^w = 0, D^m = 0\right)}{\mathbb{P}\left(D^w = 0, D^m = 0 \mid Z^w = z\right) \mathbb{P}\left(D^w = 0, D^m = 0 \mid Z^m = z'\right)}.$$

Local average treatment effect among complier-complier pairs

Fix two instrument values $z_0, z_1 \in \mathcal{Z}$. Let $\mathcal{K}^{\mathrm{CO}} := \{\, k \in \mathcal{K} : d_{z_1} \neq d_{z_0} \,\}$ be the set of response types under which treatment response switches as the instrument changes from z_0 to z_1 (i.e., compliers).

The LATE among complier-complier pairs under the independent coupling F^*:

$$\mathbb{E}_{F^*}\left[\, Y(d,d') - Y(0,0) \,\middle|\, K^w \in \mathcal{K}^{\mathrm{CO}},\, K^m \in \mathcal{K}^{\mathrm{CO}} \,\right].$$

This can be expressed as

$$LATE^*(e,p) = \sum_{k \in \mathcal{K}^{\mathrm{CO}}} \sum_{k' \in \mathcal{K}^{\mathrm{CO}}} \{e(d,d' \mid k,k') - e(0,0 \mid k,k')\} \frac{p_k^w}{\sum_{\tilde{k} \in \mathcal{K}^{\mathrm{CO}}} p_{\tilde{k}}^w} \frac{p_{k'}^m}{\sum_{\tilde{k}' \in \mathcal{K}^{\mathrm{CO}}} p_{\tilde{k}'}^m}.$$

The counterpart under the equilibrium matching F:

$$\mathbb{E}_F\left[\, Y(d,d') - Y(0,0) \,\middle|\, K^w \in \mathcal{K}^{\mathrm{CO}},\, K^m \in \mathcal{K}^{\mathrm{CO}} \,\right].$$

This can be expressed as the ratio $LATE(e,p) = N(e,p)/\Lambda(p)$, where

$$N(e,p) = \sum_{\tilde{d},\tilde{d}' \in \mathcal{D}} \sum_{z,z' \in \mathcal{Z}} \sum_{k \in \mathcal{K}(\tilde{d},\tilde{d}'|z) \cap \mathcal{K}^{\text{CO}}} \sum_{k' \in \mathcal{K}(\tilde{d}',\tilde{d}|z') \cap \mathcal{K}^{\text{CO}}} \mathbb{P}\left(D^w = \tilde{d}, D^m = \tilde{d}', Z^w = z, Z^m = z'\right) \times \frac{p_k^w p_{k'}^m \{e(d,d' \mid k,k') - e(0,0 \mid k,k')\}}{\mathbb{P}\left(D^w = \tilde{d}, D^m = \tilde{d}' \mid Z^w = z\right) \mathbb{P}\left(D^w = \tilde{d}, D^m = \tilde{d}' \mid Z^m = z'\right)},$$

and

$$\Lambda(p) = \sum_{\tilde{d},\tilde{d}' \in \mathcal{D}} \sum_{z,z' \in \mathcal{Z}} \sum_{k \in \mathcal{K}(\tilde{d},\tilde{d}'|z) \cap \mathcal{K}^{\text{CO}}} \sum_{k' \in \mathcal{K}(\tilde{d}',\tilde{d}|z') \cap \mathcal{K}^{\text{CO}}} \mathbb{P}\left(D^w = \tilde{d}, D^m = \tilde{d}', Z^w = z, Z^m = z'\right) \times \frac{p_k^w p_{k'}^m}{\mathbb{P}\left(D^w = \tilde{d}, D^m = \tilde{d}' \mid Z^w = z\right) \mathbb{P}\left(D^w = \tilde{d}, D^m = \tilde{d}' \mid Z^m = z'\right)}.$$

Generalized LATE

The ATE restricted to specific response types $K^w \in \mathcal{K}^w$ and $K^m \in \mathcal{K}^m$ under random matching F^*:

$$\mathbb{E}_{F^*}\left[Y(d,d') - Y(0,0) \mid K^w \in \mathcal{K}^w, K^m \in \mathcal{K}^m\right].$$

This can be expressed as

$$GLATE^*(e,p) = \sum_{k \in \mathcal{K}^w} \sum_{k' \in \mathcal{K}^m} \{e(d,d' \mid k,k') - e(0,0 \mid k,k')\} \frac{p_k^w}{\sum_{\tilde{k} \in \mathcal{K}^w} p_{\tilde{k}}^w} \frac{p_{k'}^m}{\sum_{\tilde{k}' \in \mathcal{K}^m} p_{\tilde{k}'}^m}.$$

F Covariates

This section discusses models with observed covariates. We consider point identification of the conditional LATE and partial identification of the conditional ATE. Section F.1 considers a model in which each person does not choose the covariates of their matched partner, although their treatment-matching response may depend on their own covariates and their potential factors may be correlated with them. Section F.2 extends this setup to allow each person to choose their partner's covariates.

F.1 When each person does not choose the covariates of whom they match

F.1.1 Setup

We incorporate discrete covariates $X^w, X^m \in \mathcal{X}$ of both sides. Formally, we use $D^s(x,z)$ to denote the potential treatment status of side s when covariates are $X^s = x$ and instruments are $Z^s = z$, and similarly define $T^s(x,z)$. The treatment-matching response type is expressed as $K^s = \{(D^s(x,z), T^s(x,z))\}_{x\in\mathcal{X}, z\in\mathcal{Z}}$. Each woman and man is characterized by a random vector

$$A^w = (X^w, Z^w, D^w, T^w, K^w, \boldsymbol{\phi}^w) \in \mathcal{A}$$
$$A^m = (X^m, Z^m, D^m, T^m, K^m, \boldsymbol{\phi}^m) \in \mathcal{A}.$$

This specification implicitly assumes that each person does not choose the covariates of whom she or he matches.

We replace the instrument exogeneity condition in Assumption 3.2 with $(K^w, \boldsymbol{\phi}^w) \perp\!\!\!\perp Z^w$ conditional on X^w and $(K^m, \boldsymbol{\phi}^m) \perp\!\!\!\perp Z^m$ conditional on X^m under F. We also replace the matching on observables condition in Assumption 3.3 as

$$\begin{aligned}&F_{A|D^w=d,D^m=d',Z^w=z,Z^m=z',X^w=x,X^m=x'}\\ =&F_{A^w|D^w=d,T^w=d',Z^w=z,X^w=x} \times F_{A^m|D^m=d',T^m=d,Z^m=z',X^m=x'}.\end{aligned}$$

The condition allows assortative matching on observables including covariates, but the matching is as if random once we condition on the observables. Note that we allow the potential factors $\boldsymbol{\phi}^s$ to be correlated with the covariates X^s.

Since the treatment-matching response now depends on both covariates and instruments, the type space is $\mathcal{K} \subseteq (\mathcal{D}^2)^{|\mathcal{X}||\mathcal{Z}|}$. We express each $k \in \mathcal{K}$ as $k = \{(d_{xz}, t_{xz})\}_{x\in\mathcal{X}, z\in\mathcal{Z}}$, where (d_{xz}, t_{xz}) represents the treatment and matching response when covariates are x and instruments are z. Let

k_{xz} denote the (x,z)-th coordinate of k for $(x,z) \in \mathcal{X} \times \mathcal{Z}$. We define $\mathcal{K}(d,t \mid x,z) := \{k \in \mathcal{K} : (d,t) = k_{xz}\}$.

F.1.2 Conditional LATE

Under the framework of Section 4, the identification results extend to the conditional case. Suppose additionally that (i) $D^s(x,1) \geq D^s(x,0)$ F-almost surely for each $x \in \mathcal{X}$ and $s \in \{w,m\}$ (i.e., conditional monotonicity), and (ii) $T^s(\cdot,\cdot) \perp\!\!\!\perp \boldsymbol{\phi}^s \mid D^s(\cdot,\cdot), X^s$ under F (i.e., ignorability of matching response conditional on treatment response and covariates). Then, for each $x, x' \in \mathcal{X}$, we can identify the conditional LATE among complier-complier pairs:

$$\begin{aligned} &CLATE^*(d,d' \mid x,x') \\ &:= \mathbb{E}_{F^*}[Y(d,d') - Y(0,0) \mid D^w(x,1) > D^w(x,0), D^m(x',1) > D^m(x',0), X^w = x, X^m = x'].\end{aligned}$$

The identification formula is analogous to that in Theorem 4.1, with all quantities conditioned on $(X^w, X^m) = (x,x')$. We can also construct an aggregated LATE by weighting $CLATE^*(d,d' \mid x,x')$ over the joint distribution of covariates:

$$LATE^{agg}(d,d') := \sum_{x \in \mathcal{X}} \sum_{x' \in \mathcal{X}} \mathbb{P}_{F^*}(X^w = x, X^m = x') \cdot CLATE^*(d,d' \mid x,x').$$

This provides a summary measure that averages the conditional treatment effects under the independent coupling.

F.1.3 Bounds of conditional ATE

Under the general framework of Section 5, we can also derive partial identification results for the conditional ATE, $\mathbb{E}_{F^*}[Y(d,d') - Y(0,0) \mid X^w = x, X^m = x']$, for each $x, x' \in \mathcal{X}$. We define the conditional treatment response as

$$e(d,d' \mid k,k',x,x') := \mathbb{E}_{F^*}[Y(d,d') \mid K^w = k, K^m = k', X^w = x, X^m = x']$$

for each $d, d' \in \mathcal{D}$, $k, k' \in \mathcal{K}$, and $x, x' \in \mathcal{X}$. When Y is binary, we have $0 \leq e(d,d' \mid k,k',x,x') \leq 1$. We use $\mathbf{E} := [0,1]^{|\mathcal{D}|^2|\mathcal{K}|^2|\mathcal{X}|^2}$ to denote the space of $\{e(d,d' \mid k,k',x,x')\}_{d,d' \in \mathcal{D},\, k,k' \in \mathcal{K},\, x,x' \in \mathcal{X}}$. Note that e now additionally conditions on covariates, since the outcome Y may depend on X^w and X^m

through $\boldsymbol{\phi}^w$ and $\boldsymbol{\phi}^m$. Next, let $p^s_{k|x} := \mathbb{P}_F(K^s = k \mid X^s = x)$ be the probability that the treatment-matching response type is k conditional on $X^s = x$, for $s \in \{w, m\}$. By conditional instrument exogeneity, $p^s_{k|x}$ does not depend on Z^s. We use $\mathbf{P} := [0,1]^{2|\mathcal{K}||\mathcal{X}|}$ to denote the space of response type probabilities.

Then, for each $\tilde{d}, \tilde{d}' \in \mathcal{D}$, $z, z' \in \mathcal{Z}$, and $x, x' \in \mathcal{X}$, the conditional expectation of the outcome can be expressed as

$$\begin{aligned}
&\mathbb{E}\left[Y \mid D^w = \tilde{d}, D^m = \tilde{d}', Z^w = z, Z^m = z', X^w = x, X^m = x'\right] \\
&= \sum_{k \in \mathcal{K}(\tilde{d}, \tilde{d}' \mid x, z)} \sum_{k' \in \mathcal{K}(\tilde{d}', \tilde{d} \mid x', z')} \frac{p^w_{k|x}\, p^m_{k'|x'}\, e(\tilde{d}, \tilde{d}' \mid k, k', x, x')}{\mathbb{P}(D^w = \tilde{d}, D^m = \tilde{d}' \mid Z^w = z, X^w = x)\, \mathbb{P}(D^w = \tilde{d}, D^m = \tilde{d}' \mid Z^m = z', X^m = x')}.
\end{aligned} \tag{F.1}$$

This is the analogue of (5.3) with all quantities conditioned on covariates. The equality holds by the conditional instrument exogeneity and matching on observables conditional on covariates.

Next, for each $s, s' \in \{w, m\}$ such that $s \neq s'$, $\tilde{d}, \tilde{d}' \in \mathcal{D}$, $z \in \mathcal{Z}$, and $x \in \mathcal{X}$, by the conditional exogeneity of instruments, we have

$$\mathbb{P}(D^s = \tilde{d}, D^{s'} = \tilde{d}' \mid Z^s = z, X^s = x) = \sum_{k \in \mathcal{K}(\tilde{d}, \tilde{d}' \mid x, z)} p^s_{k|x}. \tag{F.2}$$

This is the analogue of (5.4) conditioned on covariates. The equality follows from the conditional instrument exogeneity: since $K^s \perp\!\!\!\perp Z^s \mid X^s$, the probability $\mathbb{P}(D^s = \tilde{d}, D^{s'} = \tilde{d}' \mid Z^s = z, X^s = x)$ equals $\sum_{k \in \mathcal{K}(\tilde{d}, \tilde{d}' \mid x, z)} \mathbb{P}(K^s = k \mid X^s = x) = \sum_{k \in \mathcal{K}(\tilde{d}, \tilde{d}' \mid x, z)} p^s_{k|x}$.

Under these constraints, the conditional ATE, $\mathbb{E}_{F^*}[Y(d, d') - Y(0, 0) \mid X^w = x, X^m = x']$, can be expressed as a function of $(e, p) \in \mathbf{E} \times \mathbf{P}$:

$$CATE(e, p \mid x, x') = \sum_{k \in \mathcal{K}} \sum_{k' \in \mathcal{K}} p^w_{k|x}\, p^m_{k'|x'} \{e(d, d' \mid k, k', x, x') - e(0, 0 \mid k, k', x, x')\}. \tag{F.3}$$

This is the analogue of the objective function $ATE^*(e, p)$ in Section 5, now conditional on $X^w = x$ and $X^m = x'$. Then, as in Theorem 5.1, the lower bound of the conditional ATE is given by

$$\begin{aligned}
&\min_{(e,p) \in \mathbf{E} \times \mathbf{P}} CATE(e, p \mid x, x') \\
&\text{subject to (F.1)–(F.2) for every } \tilde{d}, \tilde{d}' \in \mathcal{D},\ z, z' \in \mathcal{Z},\ s, s' \in \{w, m\} \text{ with } s \neq s'.
\end{aligned}$$

We can also construct an aggregated ATE by weighting the conditional ATE over the joint distribution of covariates.

F.2 When each person chooses the covariates of whom she or he matches

We now consider the setup in which each person chooses the values of covariates of their matched partner. Throughout this subsubsection, $\tilde{X}^s$ denote the covariates of whom she or he matches, respectively. The potential values of covariates of whom she or he matches when $X^s = x$ and $Z^s = z$ are denoted by $\tilde{X}^s(x, z)$. The treatment-matching response type is then given by

$$K^s = \{(D^s(x,z), T^s(x,z), \tilde{X}^s(x,z))\}_{x \in \mathcal{X},\, z \in \mathcal{Z}},$$

and each woman and man is characterized by a random vector

$$A^w = (X^w, Z^w, D^w, T^w, \tilde{X}^w, K^w, \boldsymbol{\phi}^w) \in \mathcal{A},$$
$$A^m = (X^m, Z^m, D^m, T^m, \tilde{X}^m, K^m, \boldsymbol{\phi}^m) \in \mathcal{A}.$$

The distribution of interest, F, is required to be an equilibrium matching in the sense that the following two conditions hold F-almost surely:

(i) $(D^s, T^s, \tilde{X}^s) = (D^s(X^s, Z^s), T^s(X^s, Z^s), \tilde{X}^s(X^s, Z^s))$ for each $s \in \{w, m\}$;

(ii) $(D^w, D^m, X^w, X^m) = (T^m, T^w, \tilde{X}^m, \tilde{X}^w)$.

The matching on observables condition is now imposed as

$$\begin{aligned} &F_{A|D^w=d,\, D^m=d',\, Z^w=z,\, Z^m=z',\, X^w=x,\, X^m=x'} \\ &= F_{A^w|D^w=d,\, T^w=d',\, \tilde{X}^w=x',\, Z^w=z,\, X^w=x} \times F_{A^m|D^m=d',\, T^m=d,\, \tilde{X}^m=x,\, Z^m=z',\, X^m=x'} \end{aligned}$$

for each $(d, d', z, z', x, x') \in \mathcal{D}^2 \times \mathcal{Z}^2 \times \mathcal{X}^2$. The interpretation is similar to the one in the previous subsection. The matching is as if random once we condition on the observables, which now include covariates of whom each person matches with.

The conditional LATE is accordingly redefined as

$$\begin{aligned} &CLATE^*(d, d' \mid x, x') \\ &:= \mathbb{E}_{F^*}[Y(d, d') - Y(0,0) \mid D^w(x,1) > D^w(x,0),\ D^m(x',1) > D^m(x',0), \\ &\qquad\qquad X^w = x,\ X^m = x',\ \tilde{X}^w = x',\ \tilde{X}^m = x], \end{aligned}$$

and the conditional ATE as

$$\mathbb{E}_{F^*}[Y(d, d') - Y(0,0) \mid X^w = x,\ X^m = x',\ \tilde{X}^w = x',\ \tilde{X}^m = x]$$

for each $x, x' \in \mathcal{X}$. Under this setup, the identification results for the conditional LATE and the partial identification results for the conditional ATE established in the previous subsubsection still apply, because the additional elements $\tilde{X}^s$ do not affect the constraints (F.1)–(F.2) or the objective function (F.3).

G Mathematical proofs

G.1 Proofs for Section 2

We consider the selection model $D^s = \mathbb{1}(\varepsilon_1^s \le h_1^s(Z^s))$ with binary instruments $Z^s \in \{0,1\}$ for $s \in \{w,m\}$, the potential outcome $Y(d,d')$, and the realized outcome $Y = Y(D^w, D^m)$.

Proposition G.1 (LATE among complier-complier parents under joint independence)**.** *Suppose that (i) the thresholds satisfy relevance $h_1^s(0) < h_1^s(1)$ for each $s \in \{w,m\}$; (ii) $\mathbb{P}(D^w = d, D^m = d', Z^w = z, Z^m = z') > 0$ for every $d,d',z,z' \in \{0,1\}$; and (iii) the joint independence $(Z^w, Z^m) \perp\!\!\!\perp (\varepsilon_1^w, \varepsilon_1^m, Y(\cdot,\cdot))$ holds. Then*

$$LATE_{JI}(1,0) = \frac{\sum_{(z,z')\in\{0,1\}^2} (-1)^{z+z'+1}\, \mathbb{E}[\,Y\,\mathbb{1}\{D^m = 0\} \mid Z^w = z, Z^m = z'\,]}{\sum_{(z,z')\in\{0,1\}^2} (-1)^{z+z'}\, \mathbb{P}(D^w = 1, D^m = 1 \mid Z^w = z, Z^m = z')},$$

$$LATE_{JI}(0,1) = \frac{\sum_{(z,z')\in\{0,1\}^2} (-1)^{z+z'+1}\, \mathbb{E}[\,Y\,\mathbb{1}\{D^w = 0\} \mid Z^w = z, Z^m = z'\,]}{\sum_{(z,z')\in\{0,1\}^2} (-1)^{z+z'}\, \mathbb{P}(D^w = 1, D^m = 1 \mid Z^w = z, Z^m = z')}.$$

Proof. Let $D^s(z) = \mathbb{1}(\varepsilon_1^s \le h_1^s(z))$. Then, by condition (iii), for every $d,d',z,z' \in \{0,1\}$,

$$\mathbb{P}(D^w = d, D^m = d' \mid Z^w = z, Z^m = z') = \mathbb{P}(D^w(z) = d,\, D^m(z') = d'), \tag{G.1}$$

$$\mathbb{E}[\,Y\,\mathbb{1}\{D^w = d, D^m = d'\} \mid Z^w = z, Z^m = z'\,] = \mathbb{E}[\,Y(d,d')\,\mathbb{1}\{D^w(z) = d,\, D^m(z') = d'\}\,]. \tag{G.2}$$

For the denominator of $LATE_{JI}(1,0)$, by (G.1), we have

$$\begin{aligned}&\sum_{(z,z')\in\{0,1\}^2} (-1)^{z+z'}\, \mathbb{P}(D^w(z) = 1, D^m(z') = 1) \\ &= \mathbb{E}\left[\Big(\sum_z (-1)^z \mathbb{1}\{\varepsilon_1^w \le h_1^w(z)\}\Big)\Big(\sum_{z'} (-1)^{z'} \mathbb{1}\{\varepsilon_1^m \le h_1^m(z')\}\Big)\right].\end{aligned}$$

Since $\sum_z (-1)^z \mathbb{1}\{\varepsilon_1^s \le h_1^s(z)\} = -\mathbb{1}\{h_1^s(0) < \varepsilon_1^s \le h_1^s(1)\}$, the denominator equals $\mathbb{P}(\mathbb{1}\{h_1^w(0) < \varepsilon_1^w \le h_1^w(1)\}\mathbb{1}\{h_1^m(0) < \varepsilon_1^m \le h_1^m(1)\})$, which is the share of complier-complier parents. Note that this is positive by condition (i).

For the numerator of $LATE_{JI}(1,0)$,

$$
\begin{aligned}
&\sum_{(z,z')} (-1)^{z+z'+1}\, \mathbb{E}[\, Y\, \mathbb{1}\{D^m = 0\} \mid Z^w = z, Z^m = z'\,] \\
&= -\sum_{(z,z')} (-1)^{z+z'}\, \mathbb{E}[\, Y\, \mathbb{1}\{D^m = 0, D^w = 0\} + Y\, \mathbb{1}\{D^m = 0, D^w = 1\} \mid Z^w = z, Z^m = z'\,] \\
&= -\sum_{(z,z')} (-1)^{z+z'}\, \mathbb{E}[\, Y(1,0)\, \mathbb{1}\{D^w(z) = 1,\ D^m(z') = 0\} + Y(0,0)\, \mathbb{1}\{D^w(z) = 0,\ D^m(z') = 0\}\,] \\
&= -\mathbb{E}\Big[\, Y(1,0) \sum_{(z,z')} (-1)^{z+z'}\, \mathbb{1}\{D^w(z) = 1\}\, \mathbb{1}\{D^m(z') = 0\} \\
&\qquad\qquad + Y(0,0) \sum_{(z,z')} (-1)^{z+z'}\, \mathbb{1}\{D^w(z) = 0\}\, \mathbb{1}\{D^m(z') = 0\}\,\Big] \\
&= -\mathbb{E}\big[\, -Y(1,0)\, \mathbb{1}\{h_1^w(0) < \varepsilon_1^w \le h_1^w(1),\ h_1^m(0) < \varepsilon_1^m \le h_1^m(1)\} \\
&\qquad\qquad + Y(0,0)\, \mathbb{1}\{h_1^w(0) < \varepsilon_1^w \le h_1^w(1),\ h_1^m(0) < \varepsilon_1^m \le h_1^m(1)\}\,\big] \\
&= \mathbb{E}\big[\, \{Y(1,0) - Y(0,0)\}\, \mathbb{1}\{h_1^w(0) < \varepsilon_1^w \le h_1^w(1),\ h_1^m(0) < \varepsilon_1^m \le h_1^m(1)\}\,\big],
\end{aligned}
$$

where the second equality follows from (G.2), and the fourth equality follows by the same factorization as in the denominator above.

Combining the numerator and the denominator, we obtain $LATE_{JI}(1,0) = \mathbb{E}[\, Y(1,0) - Y(0,0) \mid h_1^w(0) < \varepsilon_1^w \le h_1^w(1),\ h_1^m(0) < \varepsilon_1^m \le h_1^m(1)\,]$. The identification of $LATE_{JI}(0,1)$ follows by the same argument, interchanging the roles of the two sides.

□

G.2 Proofs for Section 3

Proof of Lemma 3.1. Fix $d, d' \in \mathcal{D}$ and $z, z' \in \mathcal{Z}$. First, since F is a matching,

$$
F_{A|D^w=d,\, D^m=d',\, Z^w=z,\, Z^m=z'} = F_{A|D^w=d, T^w=d', Z^w=z, D^m=d', T^m=d, Z^m=z'}. \tag{G.3}
$$

Second, the event $\{D^w = d,\, D^m = d',\, T^w = d',\, T^m = d,\, Z^w = z,\, Z^m = z'\}$ is the intersection of the woman-side event $\{D^w = d,\, T^w = d',\, Z^w = z\}$ and the man-side event $\{D^m = d',\, T^m =$

d, $Z^m = z'\}$, involving only A^w and only A^m, respectively. Hence

$$\begin{aligned} &F^*_{A|D^w=d,T^w=d',Z^w=z,D^m=d',T^m=d,Z^m=z'} \\ &\quad = F^*_{A^w|D^w=d,\,T^w=d',\,Z^w=z} \times F^*_{A^m|D^m=d',\,T^m=d,\,Z^m=z'} \\ &\quad = F_{A^w|D^w=d,\,T^w=d',\,Z^w=z} \times F_{A^m|D^m=d',\,T^m=d,\,Z^m=z'}, \end{aligned} \tag{G.4}$$

where the first equality holds because A^w and A^m are independent under F^*, and the second because F^* has the same marginals as F. By (G.3), (G.4), and Definition 3.3, the two conditions are equivalent for every (d, d', z, z'), which is the claim. □

G.3 Proofs for Section 4

Proof of Theorem 4.1. We have the following restrictions (i)–(iv).

(i)

$$
\begin{aligned}
&\mathbb{E}[Y \mid D^w = 1, D^m = 1, Z^w = 1, Z^m = 1] \\
&= \mathbb{E}_{F^*}[\pi(\phi^w(1,1), \phi^m(1,1)) \mid D^w = 1, D^m = 1, T^w = 1, T^m = 1, Z^w = 1, Z^m = 1] \\
&= \mathbb{E}_{F^*}[\pi(\phi^w(1,1), \phi^m(1,1)) \mid D^w(1) = 1, D^m(1) = 1, Z^w = 1, Z^m = 1] \\
&= \mathbb{E}_{F^*}[\pi(\phi^w(1,1), \phi^m(1,1)) \mid D^w(1) = 1, D^m(1) = 1] \\
&= \Big\{ \mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) = 1, D^m(0) = 1] \cdot \mathbb{P}_{F^*}(D^w(0) = 1, D^m(0) = 1) \\
&\quad + \mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) = 1, D^m(0) < D^m(1)] \cdot \mathbb{P}_{F^*}(D^w(0) = 1, D^m(0) < D^m(1)) \\
&\quad + \mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) < D^w(1), D^m(0) = 1] \cdot \mathbb{P}_{F^*}(D^w(0) < D^w(1), D^m(0) = 1) \\
&\quad + \mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) < D^w(1), D^m(0) < D^m(1)] \\
&\qquad \cdot \mathbb{P}_{F^*}(D^w(0) < D^w(1), D^m(0) < D^m(1)) \Big\} \times \frac{1}{\mathbb{P}_{F^*}(D^w(1) = 1, D^m(1) = 1)} \\
&= \Big\{ \mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) = 1, D^m(0) = 1] \cdot p^w(0) \cdot p^m(0) \\
&\quad + \mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) = 1, D^m(0) < D^m(1)] \cdot p^w(0) \cdot \{p^m(1) - p^m(0)\} \\
&\quad + \mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) < D^w(1), D^m(0) = 1] \cdot \{p^w(1) - p^w(0)\} \cdot p^m(0) \\
&\quad + \mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) < D^w(1), D^m(0) < D^m(1)] \\
&\qquad \cdot \{p^w(1) - p^w(0)\} \cdot \{p^m(1) - p^m(0)\} \Big\} \times \frac{1}{p^w(1) \cdot p^m(1)}
\end{aligned}
$$

where the first equality follows from Assumptions 3.1 and 3.3, the second one from Assumption 4.1, the third one from Assumption 3.2, the fourth one from Assumption 4.3 and properties of conditional expectations, and the last one from Assumption 3.2 and that F^* is the independent coupling of F^w and F^m.

Similarly, we can derive the following restrictions.

(ii)

$$
\begin{aligned}
&\mathbb{E}[Y \mid D^w = 1, D^m = 1, Z^w = 1, Z^m = 0] \\
&= \Big\{ \mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) = 1, D^m(0) = 1] \cdot p^w(0) \cdot p^m(0) \\
&\quad + \mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) < D^w(1), D^m(0) = 1] \cdot \{p^w(1) - p^w(0)\} \cdot p^m(0) \Big\} \\
&\quad \times \frac{1}{p^w(1) \cdot p^m(0)}
\end{aligned}
$$

(iii)

$$
\begin{aligned}
&\mathbb{E}[Y \mid D^w = 1, D^m = 1, Z^w = 0, Z^m = 1] \\
&= \Big\{ \mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) = 1, D^m(0) = 1] \cdot p^w(0) \cdot p^m(0) \\
&\quad + \mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) = 1, D^m(0) < D^m(1)] \cdot p^w(0) \cdot \{p^m(1) - p^m(0)\} \Big\} \\
&\quad \times \frac{1}{p^w(0) \cdot p^m(1)}
\end{aligned}
$$

(iv)

$$
\begin{aligned}
&\mathbb{E}[Y \mid D^w = 1, D^m = 1, Z^w = 0, Z^m = 0] \\
&= \mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) = 1, D^m(0) = 1]
\end{aligned}
$$

Notice that $p^w(z) = \mathbb{P}(D^w = 1 \mid Z^w = z)$ and $p^m(z') = \mathbb{P}(D^m = 1 \mid Z^m = z')$ for each $z, z' \in \{0, 1\}$. Thus, we have four restrictions (i)–(iv) on four unknown objects, that is,

$$
\begin{aligned}
&\mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) = 1, D^m(0) = 1] \\
&\mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) = 1, D^m(0) < D^m(1)] \\
&\mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) < D^w(1), D^m(0) = 1] \\
&\mathbb{E}_{F^*}[Y(1,1) \mid D^w(0) < D^w(1), D^m(0) < D^m(1)].
\end{aligned}
$$

In an analogous manner, we can derive four restrictions on four unknown objects, that is,

$$\mathbb{E}_{F^*}[Y(d,d') \mid D^w(1-d) = d, D^m(1-d') = d']$$
$$\mathbb{E}_{F^*}[Y(d,d') \mid D^w(1-d) = d, D^m(0) < D^m(1)]$$
$$\mathbb{E}_{F^*}[Y(d,d') \mid D^w(0) < D^w(1), D^m(1-d') = d']$$
$$\mathbb{E}_{F^*}[Y(d,d') \mid D^w(0) < D^w(1), D^m(0) < D^m(1)]$$

for each $d, d' \in \mathcal{D}$.

Then, some tedious but elementary calculations give us explicit forms of these unknown objects, and

$$LATE^*(d,d') = \mathbb{E}_{F^*}[Y(d,d') - Y(0,0) \mid D^w(0) < D^w(1), D^m(0) < D^m(1)]$$

can be expressed as the ones presented in the statement of the theorem. □

Proof of Theorem 4.2. It suffices to show

$$F_{\boldsymbol{\phi}^w, \boldsymbol{\phi}^m \mid D^w(0) \neq D^w(1), D^m(0) \neq D^m(1)} = F^*_{\boldsymbol{\phi}^w, \boldsymbol{\phi}^m \mid D^w(0) \neq D^w(1), D^m(0) \neq D^m(1)}.$$

First, consider the right-hand side. Since $F^* = F^w \times F^m$, we have

$$F^*_{\boldsymbol{\phi}^w, \boldsymbol{\phi}^m \mid D^w(0) \neq D^w(1), D^m(0) \neq D^m(1)} = F^w_{\boldsymbol{\phi}^w \mid D^w(0) \neq D^w(1)} \times F^m_{\boldsymbol{\phi}^m \mid D^m(0) \neq D^m(1)}.$$

Next consider the left-hand side. By the law of total probability,

$$\begin{aligned} & F_{\boldsymbol{\phi}^w, \boldsymbol{\phi}^m \mid D^w(0) \neq D^w(1), D^m(0) \neq D^m(1)} \\ = & \sum_{z, z' \in \mathcal{Z}} \mathbb{P}_F(Z^w = z, Z^m = z' \mid D^w(0) \neq D^w(1), D^m(0) \neq D^m(1)) \\ & \quad \times F_{\boldsymbol{\phi}^w, \boldsymbol{\phi}^m \mid D^w(0) \neq D^w(1), D^m(0) \neq D^m(1), Z^w = z, Z^m = z'}. \end{aligned}$$

Notice that, conditional on $D^w(0) \neq D^w(1)$ and $D^m(0) \neq D^m(1)$, it holds that $T^w = D^m = Z^m$ and $T^m = D^w = Z^w$ under F. Then, since F satisfies matching on observables (Assumption 3.3), conditional ignorability of the matching response (Assumption 4.1) and independence of instruments

(Assumption 3.2), we have

$$\begin{aligned}
&F_{\boldsymbol{\phi}^w,\boldsymbol{\phi}^m \mid D^w(0)\neq D^w(1),\, D^m(0)\neq D^m(1),\, Z^w=z,\, Z^m=z'} \\
=&F^w_{\boldsymbol{\phi}^w \mid D^w(0)\neq D^w(1),\, D^w=z,\, T^w=z',\, Z^w=z} \times F^m_{\boldsymbol{\phi}^m \mid D^m(0)\neq D^m(1),\, D^m=z',\, T^m=z,\, Z^m=z'} \\
=&F^w_{\boldsymbol{\phi}^w \mid D^w(0)\neq D^w(1)} \times F^m_{\boldsymbol{\phi}^m \mid D^m(0)\neq D^m(1)}
\end{aligned}$$

Therefore,

$$F_{\boldsymbol{\phi}^w,\boldsymbol{\phi}^m \mid D^w(0)\neq D^w(1),\, D^m(0)\neq D^m(1)} = F^w_{\boldsymbol{\phi}^w \mid D^w(0)\neq D^w(1)} \times F^m_{\boldsymbol{\phi}^m \mid D^m(0)\neq D^m(1)},$$

which concludes the proof. □

G.4 Proofs for Section 5

Proof of Theorem 5.1. We show the equivalence of the two minimization problems. The equivalence of the two maximization problems can be shown analogously.

Step 1. First, we consider the problem

$$\inf_{(e,p)\in\mathbf{E}\times\mathbf{P}} ATE^*(e,p) \text{ subject to (5.3)–(5.4) for every } \tilde{d},\tilde{d}'\in\mathcal{D},\ z,z'\in\mathcal{Z},\ k,k'\in\mathcal{K},\ s\in\{w,m\}.$$

The infimum is attained since the feasible set is compact and ATE^* is continuous.

Step 2. Take any $(F,\pi)\in\Theta_I$. As in the main text, let $e(d,d' \mid k,k') := \mathbb{E}_{F^*}[Y(d,d') \mid K^w = k, K^m = k'] \in [0,1]$ for each $d,d'\in\mathcal{D}$ and each $k,k'\in\mathcal{K}$. We also define $p^s_k = \mathbb{P}_F(K^s = k)\in[0,1]$ for each $k\in\mathcal{K}$ and $s\in\{w,m\}$.

Then, for each $\tilde{d},\tilde{d}'\in\mathcal{D}^2$, $z,z'\in\mathcal{Z}$, we have

$$\begin{aligned}
&\mathbb{E}_F\left[Y(\tilde{d},\tilde{d}') \mid D^w=\tilde{d}, D^m=\tilde{d}', Z^w=z, Z^m=z'\right] \\
&= \mathbb{E}_{F^*}\left[Y(\tilde{d},\tilde{d}') \mid D^w=\tilde{d}, D^m=\tilde{d}', T^w=\tilde{d}', T^m=\tilde{d}, Z^w=z, Z^m=z'\right] \\
&= \mathbb{E}_{F^*}\left[Y(\tilde{d},\tilde{d}') \mid K^w\in\mathcal{K}(\tilde{d},\tilde{d}'\mid z), K^m\in\mathcal{K}(\tilde{d}',\tilde{d}\mid z')\right] \\
&= \sum_{k\in\mathcal{K}(\tilde{d},\tilde{d}'|z)}\ \sum_{k'\in\mathcal{K}(\tilde{d}',\tilde{d}|z')} \frac{p^w_k p^m_{k'} e(\tilde{d},\tilde{d}'\mid k,k')}{\mathbb{P}\left(D^w=\tilde{d}, D^m=\tilde{d}'\mid Z^w=z\right)\mathbb{P}\left(D^w=\tilde{d}, D^m=\tilde{d}'\mid Z^m=z'\right)},
\end{aligned}$$

where the last equality follows from the definition of p^s_k and $e(\tilde{d},\tilde{d}'\mid k,k')$, together with the exogeneity of instruments and the independent coupling under F^*. Hence, (5.3) is satisfied. Also, by

construction, (5.4) is satisfied. Indeed,

$$\begin{aligned}
&\mathbb{P}\left(D^s = \tilde{d}, D^{s'} = \tilde{d}' \mid Z^s = z\right) \\
&= \mathbb{P}\left(D^s(z) = \tilde{d}, T^s(z) = \tilde{d}' \mid Z^s = z\right) \\
&= \mathbb{P}\left(D^s(z) = \tilde{d}, T^s(z) = \tilde{d}'\right) \\
&= \sum_{k \in \mathcal{K}(\tilde{d}, \tilde{d}' \mid z)} p_k^s
\end{aligned}$$

Since the treatment-matching response types K^w and K^m are independent under $F^* = F^w \times F^m$ with marginals p^w and p^m, the target can be expressed directly as

$$\begin{aligned}
&\mathbb{E}_{F^*}\left[Y(d,d') - Y(0,0)\right] \\
=&\sum_{k \in \mathcal{K}} \sum_{k' \in \mathcal{K}} p_k^w p_{k'}^m \left(\mathbb{E}_{F^*}[Y(d,d') \mid K^w = k, K^m = k'] - \mathbb{E}_{F^*}[Y(0,0) \mid K^w = k, K^m = k']\right) \\
=&\sum_{k \in \mathcal{K}} \sum_{k' \in \mathcal{K}} p_k^w p_{k'}^m \left(e(d,d' \mid k,k') - e(0,0 \mid k,k')\right) \\
=&ATE^*(e,p),
\end{aligned}$$

where the first equality uses $\mathbb{P}_{F^*}(K^w = k, K^m = k') = p_k^w p_{k'}^m$ and the second uses the definition $e(d,d' \mid k,k') = \mathbb{E}_{F^*}[Y(d,d') \mid K^w = k, K^m = k']$.

Therefore, it holds that

$$\begin{aligned}
&\inf_{(F,\pi) \in \Theta_I} \mathbb{E}_{F^*}\left[Y(d,d') - Y(0,0)\right] \\
\geq &\min_{(e,p) \in \mathbf{E} \times \mathbf{P}} ATE^*(e,p) \text{ subject to (5.3)–(5.4) for every } \tilde{d}, \tilde{d}' \in \mathcal{D},\ z, z' \in \mathcal{Z},\ k, k' \in \mathcal{K},\ s \in \{w, m\}.
\end{aligned}$$

Step 3. Take any $(e,p) \in \mathbf{E} \times \mathbf{P}$ such that (5.3)–(5.4) hold for every $\tilde{d}, \tilde{d}' \in \mathcal{D}$, $z, z' \in \mathcal{Z}$, $k, k' \in \mathcal{K}$, $s \in \{w, m\}$. We wish to construct $(F, \pi) \in \Theta_I$ such that $ATE^*(e,p) = \mathbb{E}_{F^*}[Y(d,d') - Y(0,0)]$.

For each $s \in \{w, m\}$, let $\tilde{F}_K^s$ be the distribution of K^s under which $K^s = k$ with probability p_k^s.

Identify $\mathcal{K}$ with $\{1, 2, \ldots, |\mathcal{K}|\} \subset \mathbb{R}$, and fix an injection

$$\alpha : \ \mathcal{K} \times \{0,1\}^{|\mathcal{K}|} \to \mathbb{R},$$

which exists since the domain is finite.

For $K^w \sim \tilde{F}^w_K$, let $\phi^w(d,d')$ be the real-valued random variable defined as

$$\phi^w(d,d') = K^w \in \mathbb{R},$$

which does not depend on (d,d'). Then we define $\tilde{F}^w_{K\boldsymbol{\phi}}$ as the distribution of $(K^w, \boldsymbol{\phi}^w)$, where $\boldsymbol{\phi}^w = \{\phi^w(d,d')\}_{(d,d')\in\mathcal{D}^2}$.

Let $F^{dd'}_{kk'}$ be the Bernoulli distribution under which $\int \mathbb{1}(\phi = 1) dF^{dd'}_{kk'}(\phi) = e(d,d' \mid k,k')$. Then, let $(K^m, \{\phi^{dd'}_{kk'}\}_{d,d'\in\mathcal{D},k,k'\in\mathcal{K}})$ be a random vector whose distribution is the product measure of $\tilde{F}^m_K$ and $F^{dd'}_{kk'}$ for $d,d' \in \mathcal{D}, k,k' \in \mathcal{K}$. Then the real-valued random variable $\phi^m(d,d')$ is defined as

$$\phi^m(d,d') = \alpha\Big(K^m,\ \big(\phi^{dd'}_{1,K^m}, \phi^{dd'}_{2,K^m}, \ldots, \phi^{dd'}_{|\mathcal{K}|,K^m}\big)\Big) \in \mathbb{R}.$$

Then we define $\tilde{F}^m_{K\boldsymbol{\phi}}$ as the distribution of $(K^m, \boldsymbol{\phi}^m)$, where $\boldsymbol{\phi}^m = \{\phi^m(d,d')\}_{(d,d')\in\mathcal{D}^2}$.

For each $s \in \{w,m\}$, we construct $\tilde{F}^s_{ZK\boldsymbol{\phi}}$ as the product measure of $F^s_{obs,Z}$ and $\tilde{F}^s_{K\boldsymbol{\phi}}$, where $F^s_{obs,Z}$ is the marginal distribution of Z^s under G. Next, for each $s \in \{w,m\}$, let $(Z^s, K^s, \boldsymbol{\phi}^s)$ be a random vector which follows $\tilde{F}^s_{ZK\boldsymbol{\phi}}$. Then we define D^s and T^s as $D^s = D^s(Z^s)$ and $T^s = T^s(Z^s)$, respectively. We use $\tilde{F}^s$ to denote the distribution of $(Z^s, D^s, T^s, K^s, \boldsymbol{\phi}^s)$. Finally, we construct F as the mixture of $\{\tilde{F}^*_{A|D^w=d,D^m=d',T^w=d',T^m=d,Z^w=z,Z^m=z'}\}_{d,d'\in\mathcal{D},z,z'\in\mathcal{Z}}$ with weights $\{\mathbb{P}(D^w=d, D^m=d', Z^w=z, Z^m=z')\}_{d,d'\in\mathcal{D},z,z'\in\mathcal{Z}}$, where $\tilde{F}^*$ is the product measure of $\tilde{F}^w$ and $\tilde{F}^m$.

We define $\pi : \mathbb{R}\times\mathbb{R} \to \mathcal{Y}$ by

$$\pi(x,x') = \begin{cases} b_x & \text{if } x \in \{1,2,\ldots,|\mathcal{K}|\},\ x' \in \alpha\big(\mathcal{K}\times\{0,1\}^{|\mathcal{K}|}\big),\ \alpha^{-1}(x') = (k', (b_1, b_2, \ldots, b_{|\mathcal{K}|})), \\ 0 & \text{otherwise.} \end{cases}$$

The function π is Borel measurable. By construction, for every $(d,d') \in \mathcal{D}^2$,

$$\pi\big(\phi^w(d,d'),\ \phi^m(d,d')\big) = \pi\Big(K^w,\ \alpha\big(K^m,\ (\phi^{dd'}_{k,K^m})_{k\in\mathcal{K}}\big)\Big) = \phi^{dd'}_{K^w,K^m},$$

which, conditional on $(K^w, K^m) = (k,k')$, is a Bernoulli random variable with mean $e(d,d' \mid k,k')$.

Then, since under F^* the types K^w, K^m are independent with marginals p^w, p^m and $\mathbb{E}_{F^*}[Y(d,d') \mid K^w = k, K^m = k'] = e(d,d' \mid k,k')$ by construction, we have

$$\mathbb{E}_{F^*}[Y(d,d') - Y(0,0)] = \sum_{k\in\mathcal{K}}\sum_{k'\in\mathcal{K}} p^w_k p^m_{k'} \big(e(d,d' \mid k,k') - e(0,0 \mid k,k')\big) = ATE^*(e,p).$$

Next we wish to show that Assumptions 3.1–3.3 are satisfied under G and (F, π) constructed above.

For Assumption 3.3, first, we will show the marginal on A^s under F is $\tilde{F}^s$ for $s \in \{w, m\}$. Indeed, for each $d, d' \in \mathcal{D}$, $z \in \mathcal{Z}$, $k \in \mathcal{K}$, $\boldsymbol{\phi} \in \Phi^{|\mathcal{D}^2|}$,

$$
\begin{aligned}
&\mathbb{P}_F(D^w = d, T^w = d', Z^w = z, K^w = k, \boldsymbol{\phi}^w \leq \boldsymbol{\phi}) \\
&= \sum_{\tilde{d}, \tilde{d}' \in \mathcal{D}} \sum_{\tilde{z}, \tilde{z}' \in \mathcal{Z}} \mathbb{P}(D^w = \tilde{d}, D^m = \tilde{d}', Z^w = \tilde{z}, Z^m = \tilde{z}') \\
&\qquad \times \mathbb{P}_{\tilde{F}^*_{A | D^w = \tilde{d}, D^m = \tilde{d}', T^w = \tilde{d}', T^m = \tilde{d}, Z^w = \tilde{z}, Z^m = \tilde{z}'}}(D^w = d, T^w = d', Z^w = z, K^w = k, \boldsymbol{\phi}^w \leq \boldsymbol{\phi}) \\
&= \sum_{z' \in \mathcal{Z}} \mathbb{P}(D^w = d, D^m = d', Z^w = z, Z^m = z') \\
&\qquad \times \mathbb{P}_{\tilde{F}^*_{A | D^w = d, D^m = d', T^w = d', T^m = d, Z^w = z, Z^m = z'}}(D^w = d, T^w = d', Z^w = z, K^w = k, \boldsymbol{\phi}^w \leq \boldsymbol{\phi}) \\
&= \mathbb{P}(D^w = d, D^m = d', Z^w = z) \times \mathbb{P}_{\tilde{F}^*_{A | D^w = d, T^w = d', Z^w = z,}}(K^w = k, \boldsymbol{\phi}^w \leq \boldsymbol{\phi}) \\
&= \mathbb{P}_{\tilde{F}^w}(D^w = d, T^w = d', Z^w = z) \times \mathbb{P}_{\tilde{F}^w_{A^w | D^w = d, T^w = d', Z^w = z,}}(K^w = k, \boldsymbol{\phi}^w \leq \boldsymbol{\phi}) \\
&= \mathbb{P}_{\tilde{F}^w}(D^w = d, T^w = d', Z^w = z, K^w = k, \boldsymbol{\phi}^w \leq \boldsymbol{\phi}),
\end{aligned}
$$

where the third equation follows from the fact that A^w and A^m are independent under $\tilde{F}^*$ and the fourth equation follows from the fact that $\tilde{F}^*$ is the product measure of $\tilde{F}^w$ and $\tilde{F}^m$, and that

$$
\begin{aligned}
\mathbb{P}(D^w = d, D^m = d', Z^w = z) &= \mathbb{P}(D^w = d, D^m = d' \mid Z^w = z) \mathbb{P}(Z^w = z) \\
&= \sum_{k \in \mathcal{K}(d, d' | z)} p^w_k \times \mathbb{P}_{\tilde{F}^w}(Z^w = z) \\
&= \mathbb{P}_{\tilde{F}^w}(D^w(z) = d, T^w(z) = d') \mathbb{P}_{\tilde{F}^w}(Z^w = z) \\
&= \mathbb{P}_{\tilde{F}^w}(D^w(z) = d, T^w(z) = d' \mid Z^w = z) \mathbb{P}_{\tilde{F}^w}(Z^w = z) \\
&= \mathbb{P}_{\tilde{F}^w}(D^w = d, T^w = d' \mid Z^w = z) \mathbb{P}_{\tilde{F}^w}(Z^w = z) \\
&= \mathbb{P}_{\tilde{F}^w}(D^w = d, T^w = d', Z^w = z).
\end{aligned}
$$

Therefore, we find that $F^* = \tilde{F}^*$, where F^* is defined as the independent coupling of the marginal distributions F^w and F^m of A^w and A^m, respectively, under F. This, in turn, implies that F is a

matching on observables since

$$F^*_{A|D^w=d,D^m=d',T^w=d',T^m=d,Z^w=z,Z^m=z'} = \tilde{F}^*_{A|D^w=d,D^m=d',T^w=d',T^m=d,Z^w=z,Z^m=z'}$$
$$= F_{A|D^w=d,D^m=d',T^w=d',T^m=d,Z^w=z,Z^m=z'}.$$

Next, for Assumption 3.2, recall that the marginal on A^s under F is $\tilde{F}^s$ for $s \in \{w, m\}$. Since $Z^s \perp\!\!\!\perp (K^s, \boldsymbol{\phi}^s)$ under $\tilde{F}^s$ by construction, Assumption 3.2 is satisfied under F.

Finally, we consider Assumption 3.1. First, by construction,

$$\mathbb{P}_F(D^w = d, D^m = d', Z^w = z, Z^m = z') = \mathbb{P}(D^w = d, D^m = d', Z^w = z, Z^m = z').$$

Second,

$$\begin{aligned}
&\mathbb{P}_F(\pi\left(\phi^w\left(D^w, D^m\right), \phi^m\left(D^w, D^m\right)\right) = 1 \mid D^w = d, D^m = d', Z^w = z, Z^m = z') \\
&= \mathbb{P}_{F^*}(\pi\left(\phi^w\left(D^w, D^m\right), \phi^m\left(D^w, D^m\right)\right) = 1 \mid D^w = d, D^m = d', T^w = d', T^m = d, Z^w = z, Z^m = z') \\
&= \mathbb{P}_{\tilde{F}^*}(\pi\left(\phi^w\left(D^w, D^m\right), \phi^m\left(D^w, D^m\right)\right) = 1 \mid K^w \in \mathcal{K}(d, d' \mid z), K^m \in \mathcal{K}(d', d \mid z')) \\
&= \sum_{k \in \mathcal{K}(d,d'|z)} \sum_{k' \in \mathcal{K}(d',d|z')} \frac{p^w_k p^m_{k'} \mathbb{P}_{\tilde{F}^*}(\pi\left(\phi^w\left(D^w, D^m\right), \phi^m\left(D^w, D^m\right)\right) = 1 \mid K^w = k, K^m = k')}{\mathbb{P}\left(D^w = d, D^m = d' \mid Z^w = z\right) \mathbb{P}\left(D^w = d, D^m = d' \mid Z^m = z'\right)} \\
&= \sum_{k \in \mathcal{K}(d,d'|z)} \sum_{k' \in \mathcal{K}(d',d|z')} \frac{p^w_k p^m_{k'} \mathbb{P}_{F^{dd'}_{kk'}}(\phi^{dd'}_{kk'} = 1)}{\mathbb{P}\left(D^w = d, D^m = d' \mid Z^w = z\right) \mathbb{P}\left(D^w = d, D^m = d' \mid Z^m = z'\right)} \\
&= \sum_{k \in \mathcal{K}(d,d'|z)} \sum_{k' \in \mathcal{K}(d',d|z')} \frac{p^w_k p^m_{k'} e(d, d' \mid k, k')}{\mathbb{P}\left(D^w = d, D^m = d' \mid Z^w = z\right) \mathbb{P}\left(D^w = d, D^m = d' \mid Z^m = z'\right)} \\
&= \mathbb{P}(Y = 1 \mid D^w = d, D^m = d', Z^w = z, Z^m = z')
\end{aligned}$$

Hence we conclude that Assumption 3.1 is satisfied. Therefore, $(F, \pi) \in \Theta_I$, which implies

$$\begin{aligned}
&\min_{(F,\pi)\in\Theta_I} \mathbb{E}_{F^*}\left[Y(d, d') - Y(0, 0)\right] \\
&\leq \min_{(e,p)\in\mathbf{E}\times\mathbf{P}} ATE^*(e, p) \text{ subject to (5.3)–(5.4) for every } \tilde{d}, \tilde{d}' \in \mathcal{D},\ z, z' \in \mathcal{Z},\ k, k' \in \mathcal{K},\ s \in \{w, m\}.
\end{aligned}$$

□

Proof of Corollary 5.1. Step 1. First, we consider the problem

$$\inf_{(e,p,\delta)\in\mathbf{E}\times\mathbf{P}\times\Delta} ATE^*(e, p) \text{ subject to (5.3), (5.4), and (5.6) for every } \tilde{d}, \tilde{d}' \in \mathcal{D},\ z, z' \in \mathcal{Z},\ k, k' \in \mathcal{K},\ s \in \{w, m\}.$$

The infimum is attained since the feasible set is compact and the function ATE^* is continuous.

Step 2. Take any $(F, \pi) \in \tilde{\Theta}_I$. Again, let $e(d, d' \mid k, k') := \mathbb{E}_{F^*}[Y(d, d') \mid K^w = k, K^m = k'] \in [0, 1]$ for each $d, d' \in \mathcal{D}$ and each $k, k' \in \mathcal{K}$. We also define $p_k^s = \mathbb{P}_F(K^s = k) \in [0, 1]$ for each $k \in \mathcal{K}$ and $s \in \{w, m\}$.

Then as in the proof of Theorem 5.1, we can show (5.3) and (5.4) are satisfied. We can also show $\mathbb{E}_{F^*}[Y(d, d') - Y(0, 0)]$ can be expressed as the right-hand side of (5.2). Finally, since there exists $\delta \in \Delta$ such that (5.5) holds under (F, π), for each $s \in \{w, m\}$,

$$\begin{aligned} p_k^s &= \mathbb{P}_F(K^s = k) \\ &= \int \mathbb{1}(\{(d^s(z, \varepsilon^s; \gamma), t^s(z, \varepsilon^s; \gamma))\}_{z \in \mathcal{Z}} = k) dF_\varepsilon^s(\varepsilon^s, \sigma) \\ &= \int \mathbb{1}(\varepsilon^s \in \mathcal{E}^s(k; \gamma)) dF_\varepsilon^s(\varepsilon^s, \sigma), \end{aligned}$$

and hence, (5.6) is satisfied. Thus,

$$\begin{aligned} &\min_{(F,\pi) \in \tilde{\Theta}_I} \mathbb{E}_{F^*}[Y(d, d') - Y(0, 0)] \\ &\geq \min_{(e,p,\delta) \in \mathbf{E} \times \mathbf{P} \times \Delta} ATE^*(e, p) \text{ subject to (5.3), (5.4), and (5.6) for every } \tilde{d}, \tilde{d}' \in \mathcal{D},\ z, z' \in \mathcal{Z},\ k, k' \in \mathcal{K},\ s \in \{w, m\}. \end{aligned}$$

Step 3. Take any $(e, p, \delta) \in \mathbf{E} \times \mathbf{P} \times \Delta$ such that (5.3), (5.4), and (5.6) hold for every $\tilde{d}, \tilde{d}' \in \mathcal{D}$, $z, z' \in \mathcal{Z}$, $k, k' \in \mathcal{K}$, $s \in \{w, m\}$. We wish to construct $(F, \pi) \in \tilde{\Theta}_I$ such that $ATE^*(e, p) = \mathbb{E}_{F^*}[Y(d, d') - Y(0, 0)]$.

For each $s \in \{w, m\}$, for $\varepsilon^s \sim F_\varepsilon^s(\cdot, \sigma)$, we define a random vector K^s as

$$K^s = \{(d^s(z, \varepsilon^s; \gamma), t^s(z, \varepsilon^s; \gamma))\}_{z \in \mathcal{Z}}.$$

Let $\tilde{F}_K^s$ be the distribution of K^s.

Next, given $\tilde{F}_K^s$, we construct (F, π) in exactly the same manner as in the proof of Theorem 5.1. Then, we can show that $\mathbb{E}_{F^*}[Y(d, d') - Y(0, 0)] = ATE^*(e, p)$ and $(F, \pi) \in \tilde{\Theta}_I$. Also, by construction, $\{(d^s(z, \varepsilon^s; \gamma), t^s(z, \varepsilon^s; \gamma))\}_{z \in \mathcal{Z}} \sim F_K^s$ for $\varepsilon^s \sim F_\varepsilon^s(\cdot; \sigma)$. Hence,

$$\begin{aligned} &\min_{(F,\pi) \in \tilde{\Theta}_I} \mathbb{E}_{F^*}[Y(d, d') - Y(0, 0)] \\ &\leq \min_{(e,p,\delta) \in \mathbf{E} \times \mathbf{P} \times \Delta} ATE^*(e, p) \text{ subject to (5.3), (5.4), and (5.6) for every } \tilde{d}, \tilde{d}' \in \mathcal{D},\ z, z' \in \mathcal{Z},\ k, k' \in \mathcal{K},\ s \in \{w, m\}. \end{aligned}$$

□

G.5 Proofs for Section 6

We focus on the ATE under random matching $\mathbb{E}_{F^*}[Y(d,d') - Y(0,0)]$. Other causal parameters can be handled similarly.

Let $\xi = (a, \{e(d,d' \mid k,k')\}_{k,k' \in \mathcal{K}}, \{e(0,0 \mid k,k')\}_{k,k' \in \mathcal{K}})$, where a denotes the target parameter (i.e., ATE^*), intuitively. We define a $(1 + 2|\mathcal{Z}|^2 + 4|\mathcal{K}|^2) \times (2 + 2|\mathcal{K}|^2)$ matrix $Q(\tau)$ as a function of $\tau = (\delta, \{\mathbb{P}(D^w = d, D^m = d', Z^w = z, Z^m = z', Y = y)\}_{(d,d',z,z',y) \in \mathcal{D}^2 \times \mathcal{Z}^2 \times \mathcal{Y}})$;

$$Q(\tau) = \begin{pmatrix} -1 & \mathbf{q}^1 & -\mathbf{q}^1 & 0 \\ \mathbf{0}_{|\mathcal{Z}|^2} & \mathbf{q}^2_{dd'} & \mathbf{O}_{|\mathcal{Z}|^2,|\mathcal{K}|^2} & \mathbf{q}^e_{dd'} \\ \mathbf{0}_{|\mathcal{Z}|^2} & \mathbf{O}_{|\mathcal{Z}|^2,|\mathcal{K}|^2} & \mathbf{q}^2_{00} & \mathbf{q}^e_{00} \\ \mathbf{0}_{|\mathcal{K}|^2} & \mathbf{I}_{|\mathcal{K}|^2} & \mathbf{O}_{|\mathcal{K}|^2,|\mathcal{K}|^2} & -\mathbf{1}_{|\mathcal{K}|^2} \\ \mathbf{0}_{|\mathcal{K}|^2} & \mathbf{O}_{|\mathcal{K}|^2,|\mathcal{K}|^2} & \mathbf{I}_{|\mathcal{K}|^2} & -\mathbf{1}_{|\mathcal{K}|^2} \\ \mathbf{0}_{|\mathcal{K}|^2} & -\mathbf{I}_{|\mathcal{K}|^2} & \mathbf{O}_{|\mathcal{K}|^2,|\mathcal{K}|^2} & \mathbf{0}_{|\mathcal{K}|^2} \\ \mathbf{0}_{|\mathcal{K}|^2} & \mathbf{O}_{|\mathcal{K}|^2,|\mathcal{K}|^2} & -\mathbf{I}_{|\mathcal{K}|^2} & \mathbf{0}_{|\mathcal{K}|^2} \end{pmatrix},$$

where

$$\mathbf{q}^1 = \{\tilde{p}^w_k(\delta)\tilde{p}^m_{k'}(\delta)\}_{k,k' \in \mathcal{K}}$$

is a row vector of dimension $1 \times |\mathcal{K}|^2$,

$$\mathbf{q}^2_{dd'} = \left\{ \frac{\mathbb{1}\big(k \in \mathcal{K}(d,d' \mid z), k' \in \mathcal{K}(d',d \mid z')\big)\, \tilde{p}^w_k(\delta)\tilde{p}^m_{k'}(\delta)\, \mathbb{P}(D^w = d, D^m = d', Z^w = z, Z^m = z')}{\mathbb{P}\left(D^w = d, D^m = d' \mid Z^w = z\right) \mathbb{P}\left(D^w = d, D^m = d' \mid Z^m = z'\right)} \right\}_{z,z' \in \mathcal{Z}, k,k' \in \mathcal{K}}$$

is a matrix of dimension $|\mathcal{Z}|^2 \times |\mathcal{K}|^2$,

$$\mathbf{q}^e_{dd'} = -\left\{\mathbb{P}(D^w = d, D^m = d', Z^w = z, Z^m = z', Y = 1)\right\}_{z,z' \in \mathcal{Z}}$$

is a column vector of dimension $|\mathcal{Z}|^2 \times 1$, $\mathbf{I}_{|\mathcal{K}|^2}$ denotes the identity matrix of dimension $|\mathcal{K}|^2 \times |\mathcal{K}|^2$, and $\mathbf{O}$ and $\mathbf{0}$ denote zero matrices and zero vectors of appropriate dimensions, respectively. Note that the total number of restrictions is $L = 1 + 2|\mathcal{Z}|^2 + 4|\mathcal{K}|^2$ and the number of variables is $\dim(\xi) = 1 + 2|\mathcal{K}|^2$.

Lemma G.1 (Lemma 7 in Gafarov, 2025)**.** *We consider the following perturbed problem:*

$$\min_{\xi\in\mathbb{R}^{d_\xi}} (1,0,\dots,0)\cdot\xi+\mu\|\xi\|^2 \text{ subject to } \begin{cases} r_\ell(Q_G+tq,\xi)=0, & 1\le\ell\le L_{eq} \\ r_\ell(Q_G+tq,\xi)\le 0, & L_{eq}+1\le\ell\le L. \end{cases}$$

There exist $C_1, C_2 > 0$ such that, for any $G \in \mathcal{G}$, $q \in \mathbb{R}^{L\times(d_\xi+1)}$ with norm $\|q\| < C_1$, any $\mu \in (0, 1/2]$, and any $t \in [0, 1]$,

$$\left\|\underline{\xi}_q(t)-\underline{\xi}\right\| \le C_2\frac{\|tq\|}{\mu},$$
$$\left\|\underline{\lambda}_q(t)-\underline{\lambda}\right\| \le C_2\|tq\|,$$
$$\left|\underline{v}_q(t)-\underline{v}-t\underline{\lambda}'(r_1(q,\underline{\xi}),\dots,r_L(q,\underline{\xi}))'\right| \le C_2\frac{\|tq\|^2}{\mu},$$

where (i) $\underline{v}$, $\underline{\xi}$, and $\underline{\lambda}$ are the value, the primal solution, and the dual solution of the original problem (i.e., when $q = 0$) and (ii) $\underline{v}_q(t)$, $\underline{\xi}_q(t)$, and $\underline{\lambda}_q(t)$ are the value, the primal solution, and the dual solution of the perturbed problem.

Lemma G.2. *Let $c = \min_{d,d'\in\mathcal{D},\, z,z'\in\mathcal{Z}} \inf_{G\in\mathcal{G}} \mathbb{P}_G(D^w = d, D^m = d', Z^w = z, Z^m = z')$, which is positive by Assumption 6.1(iii). Define*

$$\mathcal{T} := \Bigg\{(\delta,\{p_{dd'zz'y}\}_{d,d'\in\mathcal{D},\, z,z'\in\mathcal{Z},\, y\in\mathcal{Y}}) \in \Delta\times[0,1]^{|\mathcal{D}|^2|\mathcal{Z}|^2|\mathcal{Y}|} : \textstyle\sum_{d,d',z,z',y} p_{dd'zz'y} = 1, \\ \textstyle\sum_{y\in\mathcal{Y}} p_{dd'zz'y} \ \ge\ c \text{ for each } d,d'\in\mathcal{D},\, z,\, z'\in\mathcal{Z}\Bigg\},$$

which is compact and satisfies $\tau_G \in \mathcal{T}$ for every $G \in \mathcal{G}$. Let $\tilde{\mathcal{T}}$ be a compact neighborhood of $\mathcal{T}$, that is, a compact set with $\mathcal{T} \subset \mathrm{int}(\tilde{\mathcal{T}})$. Suppose that the functions $\{f_G\}_{G\in\mathcal{G}}$ from $\mathbb{R}^{\dim(\tau)}$ to $\mathbb{R}^k$ are continuously differentiable on an open set containing $\tilde{\mathcal{T}}$, of which Jacobians $\{\nabla f_G\}_{G\in\mathcal{G}}$ satisfy equicontinuity on $\tilde{\mathcal{T}}$ and $\sup_{G\in\mathcal{G}}\sup_{\tau\in\tilde{\mathcal{T}}}\|\nabla f_G(\tau)\| < \infty$. Then,

$$\lim_{n\to\infty}\sup_{G\in\mathcal{G}} \rho(\sqrt{n}(f_G(\hat{\tau}_G)-f_G(\tau_G)), N(0,\nabla f_G(\tau_G)\Sigma_G\nabla f_G(\tau_G)')) = 0.$$

Proof of Lemma G.2. Since $\mathcal{T}$ is compact and $\mathcal{T} \subset \mathrm{int}(\tilde{\mathcal{T}})$, there exists $\kappa > 0$ such that every τ with $\inf_{\tau'\in\mathcal{T}}\|\tau-\tau'\| \le \kappa$ belongs to $\tilde{\mathcal{T}}$. By Assumption 6.2(i), $\sup_{G\in\mathcal{G}}\mathbb{P}_G(\|\hat{\tau}_G-\tau_G\| > \kappa) \to 0$. On the event $\{\|\hat{\tau}_G-\tau_G\| \le \kappa\}$, the segment joining τ_G and $\hat{\tau}_G$ lies in $\tilde{\mathcal{T}}$, because $\tau_G \in \mathcal{T}$, and we

restrict attention to this event. Since f_G is continuously differentiable on an open set containing the segment, the fundamental theorem of calculus applied to $t \mapsto f_G(\tau_G + t(\hat{\tau}_G - \tau_G))$ gives

$$\sqrt{n}(f_G(\hat{\tau}_G) - f_G(\tau_G)) = \nabla f_G(\tau_G)\sqrt{n}(\hat{\tau}_G - \tau_G) + W_n(G),$$

where $W_n(G) := \left[\int_0^1 \left(\nabla f_G(\tau_G + t(\hat{\tau}_G - \tau_G)) - \nabla f_G(\tau_G)\right) dt\right] \sqrt{n}(\hat{\tau}_G - \tau_G)$.

Step 1. Since $\sup_{G\in\mathcal{G}} \rho(\sqrt{n}(\hat{\tau}_G - \tau_G), N(0, \Sigma_G)) \to 0$ (i.e., Assumption 6.2(ii)), and $\sup_G \|\Sigma_G\| < \infty$ (i.e., Assumption 6.2(iii)), the sequence $\{\sqrt{n}(\hat{\tau}_G - \tau_G)\}_{n,G}$ is uniformly tight. That is, for any $\eta > 0$, there exists a constant $M_\eta < \infty$ such that

$$\limsup_{n\to\infty} \sup_{G\in\mathcal{G}} \mathbb{P}_G(\|\sqrt{n}(\hat{\tau}_G - \tau_G)\| > M_\eta) < \eta.$$

This follows from the fact that $\{N(0, \Sigma_G)\}$ is uniformly tight and the Prohorov distance converges uniformly.

Step 2. Since $\{\nabla f_G\}_{G\in\mathcal{G}}$ is equicontinuous on the compact set $\tilde{\mathcal{T}}$, it is uniformly equicontinuous (equicontinuity version of Heine-Cantor theorem). Thus, for any $\epsilon > 0$ and $M > 0$, there exists a $\varsigma > 0$ such that $\|\tau_1 - \tau_2\| < \varsigma$ implies $\|\nabla f_G(\tau_1) - \nabla f_G(\tau_2)\| < \epsilon/M$ for all $G \in \mathcal{G}$ and all $\tau_1, \tau_2 \in \tilde{\mathcal{T}}$. We observe that

$$\begin{aligned}\mathbb{P}_G(\|W_n(G)\| > \epsilon) \;&\le \mathbb{P}_G\left(\|\sqrt{n}(\hat{\tau}_G - \tau_G)\| > M\right) \\ &\quad + \mathbb{P}_G\left(\sup_{t\in[0,1]} \|\nabla f_G(\tau_G + t(\hat{\tau}_G - \tau_G)) - \nabla f_G(\tau_G)\| > \epsilon/M\right).\end{aligned}$$

Note that every point of the segment is within $\|\hat{\tau}_G - \tau_G\| = \|\sqrt{n}(\hat{\tau}_G - \tau_G)\|/\sqrt{n}$ of τ_G. By Step 1, the first term on the right hand side can be made arbitrarily small by choosing sufficiently large M. For a fixed M, the second term vanishes uniformly over G as $n \to \infty$ because $\mathbb{P}_G(\|\hat{\tau}_G - \tau_G\| > \varsigma) \to 0$ (i.e., Assumption 6.2(i)). Thus, $\sup_{G\in\mathcal{G}} \mathbb{P}_G(\|W_n(G)\| > \epsilon) \to 0$ as $n \to \infty$.

Step 3. By the triangle inequality, we obtain

$$\begin{aligned}&\rho(\sqrt{n}(f_G(\hat{\tau}_G) - f_G(\tau_G)), N(0, \nabla f_G(\tau_G)\Sigma_G \nabla f_G(\tau_G)')) \\ &\le \rho(\nabla f_G(\tau_G)\sqrt{n}(\hat{\tau}_G - \tau_G) + W_n(G), \nabla f_G(\tau_G)\sqrt{n}(\hat{\tau}_G - \tau_G)) \\ &\quad + \rho(\nabla f_G(\tau_G)\sqrt{n}(\hat{\tau}_G - \tau_G), N(0, \nabla f_G(\tau_G)\Sigma_G \nabla f_G(\tau_G)')).\end{aligned}$$

The first term is bounded by $\inf_{\epsilon>0}\{\epsilon + \mathbb{P}_G(\|W_n(G)\| > \epsilon)\}$, which converges to 0 uniformly by Step

2. For the second term, since $\nabla f_G(\tau_G)$ is uniformly bounded over $G \in \mathcal{G}$ by assumption, there exists $C < \infty$ such that $\rho(\nabla f_G(\tau_G)\sqrt{n}(\hat{\tau}_G - \tau_G), N(0, \nabla f_G(\tau_G)\Sigma_G \nabla f_G(\tau_G)')) \leq \max(1, C)\rho(\sqrt{n}(\hat{\tau}_G - \tau_G), N(0, \Sigma_G))$, which implies that the second term also converges to 0 uniformly. □

Lemma G.3 (Theorem 1 in Gafarov, 2025)**.** *Let $\underline{v}^{out}(\mu, Q_G) := \underline{v}(\mu, Q_G) - \mu\|\underline{\xi}^*\|^2$, where $\underline{\xi}^*$ is any point in the set of all optimal solutions of (6.1). Then, $\underline{v}^{out}(\mu, Q_G) \leq \underline{v}(Q_G)$ for any $G \in \mathcal{G}$ and $\mu > 0$.*

Lemma G.4.

$$\sqrt{n}(\underline{v}(\mu_n, \hat{Q}_G) - \underline{v}(\mu_n, Q_G)) = \sqrt{n}\underline{\lambda}(\mu_n, Q_G)' r(\hat{Q}_G, \underline{\xi}(\mu_n, Q_G)) + O_{\mathcal{G}}(\frac{1}{\mu_n\sqrt{n}})$$

Proof of Lemma G.4. As in Lemma 10 in Gafarov, 2025, we can show that $\underline{\xi}(\mu_n, \hat{Q}_G)$ and $\underline{v}(\mu_n, \hat{Q}_G)$ are well defined with probability approaching one uniformly over $G \in \mathcal{G}$. Take $\tilde{\mathcal{T}}$ as a sufficiently small compact neighborhood of $\mathcal{T}$. Then, the function $Q^v(\cdot) := vec(Q(\cdot))$, which does not depend on G, is continuously differentiable on $\tilde{\mathcal{T}}$, and its Jacobian $\nabla_\tau Q^v$ is uniformly continuous and bounded on $\tilde{\mathcal{T}}$. Applying Lemma G.2 therefore yields

$$\lim_{n\to\infty} \sup_{G\in\mathcal{G}} \rho(\sqrt{n}(Q^v(\hat{\tau}_G) - Q^v(\tau_G)), N(0, \nabla_\tau Q^v(\tau_G)\Sigma_G \nabla_\tau Q^v(\tau_G)')) = 0.$$

Since $\nabla_\tau Q^v(\tau_G)\Sigma_G \nabla_\tau Q^v(\tau_G)'$ is uniformly bounded, $\sqrt{n}(Q^v(\hat{\tau}_G) - Q^v(\tau_G)) = O_{\mathcal{G}}(1)$. Also, it is easy to see $\lim_{n\to\infty} \sup_{G\in\mathcal{G}} \mathbb{P}_G(\|Q(\hat{\tau}_G) - Q(\tau_G)\| \geq \epsilon) = 0$ for every $\epsilon > 0$ since Q is uniformly continuous on $\tilde{\mathcal{T}}$. Then, by applying Lemma G.1 with $q = Q(\hat{\tau}_G) - Q(\tau_G) = \hat{Q}_G - Q_G$, we obtain

$$\begin{aligned}\sqrt{n}(\underline{v}(\mu_n, \hat{Q}_G) - \underline{v}(\mu_n, Q_G)) &= \sqrt{n}\underline{\lambda}(\mu_n, Q_G)'(r(\hat{Q}_G, \underline{\xi}(\mu_n, Q_G)) - r(Q_G, \underline{\xi}(\mu_n, Q_G))) + O_{\mathcal{G}}(\frac{1}{\mu_n\sqrt{n}}) \\ &= \sqrt{n}\underline{\lambda}(\mu_n, Q_G)' r(\hat{Q}_G, \underline{\xi}(\mu_n, Q_G)) + O_{\mathcal{G}}(\frac{1}{\mu_n\sqrt{n}}),\end{aligned}$$

where $\underline{\lambda}(\mu_n, Q_G)' r(Q_G, \underline{\xi}(\mu_n, Q_G)) = 0$ since the regularized problem is a convex optimization problem, $\Xi(Q_G)$ is nonempty, and Slater's condition holds (i.e., Assumption 6.1(iv)).

□

Lemma G.5. *Consider any sequence μ_n such that $\mu_n \to 0$ and $\mu_n\sqrt{n} \to \infty$. Then,*

$$\lim_{n\to\infty} \sup_{G\in\mathcal{G}} \rho\left(\sqrt{n}(\underline{v}(\mu_n, \hat{Q}_G) - \underline{v}(\mu_n, Q_G)), N\left(0, \underline{\eta}^2(\mu_n, G)\right)\right) = 0, \tag{G.5}$$

$$\underline{\xi}\left(\mu_n, \hat{Q}_G\right) = \underline{\xi}(\mu_n, Q_G) + O_{\mathcal{G}}\left(\frac{1}{\mu_n\sqrt{n}}\right), \tag{G.6}$$

$$\underline{\lambda}\left(\mu_n, \hat{Q}_G\right) = \underline{\lambda}(\mu_n, Q_G) + O_{\mathcal{G}}\left(\frac{1}{\sqrt{n}}\right), \tag{G.7}$$

$$\hat{\underline{\eta}}^2(\mu_n, G) = \underline{\eta}^2(\mu_n, G) + o_{\mathcal{G}}(1), \tag{G.8}$$

where $\underline{\eta}^2(\mu_n, G) := \underline{\lambda}_G' \underline{R}_G \Sigma_G \underline{R}_G' \underline{\lambda}_G$, $\underline{R}_G := R(\tau_G, \underline{\xi}(\mu_n, Q_G))$, *and* $\underline{\lambda}_G := \underline{\lambda}(\mu_n, Q_G)$.

Proof of Lemma G.5. (G.5) follows from Lemma G.2 and Lemma G.4. By applying Lemma G.1, we obtain (G.6) and (G.7). Note that $\sup_{G\in\mathcal{G}} \|\underline{R}_G\| < \infty$ by construction, $\sup_{G\in\mathcal{G}} \|\underline{\lambda}_G\| < \infty$ (see Lemma 4 in Gafarov, 2025), and $\sup_{G\in\mathcal{G}} \|\Sigma_G\| < \infty$ by Assumption 6.2(iii). Thus, we have (G.8) since (G.6), (G.7), and $\|\hat{\underline{R}}_G - \underline{R}_G\| = o_{\mathcal{G}}(1)$. □

Lemma G.6. *For any $\epsilon_1 > 0$, there exists $M > 0$ such that for any n,*

$$\sup_{G\in\mathcal{G}} \mathbb{P}_G(\sqrt{n}\|\underline{\xi}^*(\hat{Q}_G) - \underline{\xi}^*(Q_G)\| \geq M) \leq \epsilon_1.$$

Proof of Lemma G.6. Given that $\sqrt{n}(\hat{Q}_G - Q_G) = O_{\mathcal{G}}(1)$, established in the proof of Lemma G.4, the claim follows from the argument for Lemma 12 in Gafarov, 2025. □

Then, we are ready to prove Proposition 6.1.

Proof of Proposition 6.1. Pick any $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that $c_{1-\frac{\alpha}{2}}\eta_0 - 3\epsilon_2 > 0$. Write $\underline{\eta}^+ := \max\{\underline{\eta}(\mu_n, G), \eta_0\} \geq \eta_0$. Since $x \mapsto \max\{x, \eta_0\}$ is 1-Lipschitz, we have $|\hat{\underline{\eta}}^+ - \underline{\eta}^+| \leq |\hat{\underline{\eta}} - \underline{\eta}|$. Then, there exists N such that for all $n \geq N$,

$$\inf_{G\in\mathcal{G}} \mathbb{P}_G\left(\mu_n\sqrt{n}\left|\|\underline{\xi}^*(\hat{Q}_G)\|^2 - \|\underline{\xi}^*(Q_G)\|^2\right| \leq \epsilon_2\right) \geq 1 - \epsilon_1 \tag{G.9}$$

$$\inf_{G\in\mathcal{G}} \mathbb{P}_G\left(c_{1-\frac{\alpha}{2}}\left|\hat{\underline{\eta}}(\mu_n, G) - \underline{\eta}(\mu_n, G)\right| \leq \epsilon_2\right) \geq 1 - \epsilon_1 \tag{G.10}$$

$$\sup_{G\in\mathcal{G}} \rho\Big(\sqrt{n}\big(\underline{v}(\mu_n, \hat{Q}_G) - \underline{v}(\mu_n, Q_G)\big), N\big(0, \underline{\eta}^2(\mu_n, G)\big)\Big) \leq \epsilon_2, \tag{G.11}$$

where (G.9) follows from Lemma G.6 and the uniform boundedness of $\|\underline{\xi}^*(Q)\|$ over the Q under

consideration, which implies

$$\mu_n\sqrt{n}\left|\|\underline{\xi}^*(\hat{Q}_G)\|^2 - \|\underline{\xi}^*(Q_G)\|^2\right| \le \mu_n\Big(\sqrt{n}\,\|\underline{\xi}^*(\hat{Q}_G) - \underline{\xi}^*(Q_G)\|\Big)\cdot O(1) = \mu_n\, O_{\mathcal{G}}(1).$$

(G.10) follows from (G.8) of Lemma G.5 and the elementary inequality $|a-b|^2 \le |a^2-b^2|$ for $a, b \ge 0$, which gives $|\hat{\underline{\eta}} - \underline{\eta}| \le |\hat{\underline{\eta}}^2 - \underline{\eta}^2|^{1/2} = o_{\mathcal{G}}(1)$. (G.11) is (G.5) of Lemma G.5. Similarly, we have their counterparts for the upper bound.

Let

$$LB_{G,n} := \underline{v}^{out}(\mu_n, \hat{Q}_G) - c_{1-\frac{\alpha}{2}}\frac{\hat{\underline{\eta}}^+}{\sqrt{n}}$$
$$UB_{G,n} := \overline{v}^{out}(\mu_n, \hat{Q}_G) + c_{1-\frac{\alpha}{2}}\frac{\hat{\overline{\eta}}^+}{\sqrt{n}}.$$

Then,

$$\begin{aligned}
&\mathbb{P}_G\big(\big[\underline{ATE^*}_G, \overline{ATE^*}_G\big] \subset [LB_{G,n}, UB_{G,n}]\big)\\
&= 1 - \mathbb{P}_G(\overline{v}(Q_G) > UB_{G,n} \text{ or } \underline{v}(Q_G) < LB_{G,n})\\
&\ge 1 - \mathbb{P}_G(\underline{v}(Q_G) < LB_{G,n}) - \mathbb{P}_G(\overline{v}(Q_G) > UB_{G,n})
\end{aligned}$$

Now, for any $n \ge N$,

$$\begin{aligned}
&\mathbb{P}_G(\underline{v}(Q_G) < LB_{G,n})\\
&= \mathbb{P}_G(\underline{v}(Q_G) < \underline{v}(\mu_n, \hat{Q}_G) - \mu_n\|\underline{\xi}^*(\hat{Q}_G)\|^2 - c_{1-\frac{\alpha}{2}}\frac{\hat{\underline{\eta}}^+}{\sqrt{n}})\\
&\le \mathbb{P}_G\left(\underline{v}(\mu_n, Q_G) - \mu_n\|\underline{\xi}^*(Q_G)\|^2 < \underline{v}(\mu_n, \hat{Q}_G) - \mu_n\|\underline{\xi}^*(\hat{Q}_G)\|^2 - c_{1-\frac{\alpha}{2}}\frac{\hat{\underline{\eta}}^+}{\sqrt{n}}\right)\\
&\le \mathbb{P}_G\left(\sqrt{n}(\underline{v}(\mu_n, \hat{Q}_G) - \underline{v}(\mu_n, Q_G)) > c_{1-\frac{\alpha}{2}}\underline{\eta}^+ - 2\epsilon_2\right) + 2\epsilon_1\\
&\le \mathbb{P}_G\left(\underline{\eta}(\mu_n, G)\, X \ge c_{1-\frac{\alpha}{2}}\underline{\eta}^+ - 3\epsilon_2\right) + \epsilon_2 + 2\epsilon_1\\
&\le \left(1 - F_{\mathcal{N}(0,1)}\Big(c_{1-\frac{\alpha}{2}} - \frac{3\epsilon_2}{\eta_0}\Big) + \epsilon_2\right) + 2\epsilon_1,
\end{aligned}$$

where the first inequality follows from the fact that $\underline{v}(\mu_n, Q_G) - \mu_n\|\underline{\xi}^*(Q_G)\|^2 \le \underline{v}(Q_G)$, the second inequality follows from (G.9), (G.10), and $|\hat{\underline{\eta}}^+ - \underline{\eta}^+| \le |\hat{\underline{\eta}} - \underline{\eta}|$, and the third inequality follows from (G.11) and the definition of the Lévy–Prokhorov metric, which gives $\mathbb{P}_G\big(\sqrt{n}(\underline{v}(\mu_n, \hat{Q}_G) - \underline{v}(\mu_n, Q_G)) > t\big) \le \mathbb{P}_G\big(\underline{\eta}(\mu_n, G)\, X \ge t - \epsilon_2\big) + \epsilon_2$ for every $t \in \mathbb{R}$, where X is a standard normal

random variable independent of the data. The last inequality is obtained by considering the following three cases. If $\underline{\eta}(\mu_n, G) \geq \eta_0$, then $\underline{\eta}^+ = \underline{\eta}(\mu_n, G)$ and $\mathbb{P}_G\big(\underline{\eta}(\mu_n, G)\, X \geq c_{1-\frac{\alpha}{2}}\underline{\eta}(\mu_n, G) - 3\epsilon_2\big) = 1 - F_{\mathcal{N}(0,1)}\big(c_{1-\frac{\alpha}{2}} - 3\epsilon_2/\underline{\eta}(\mu_n, G)\big) \leq 1 - F_{\mathcal{N}(0,1)}\big(c_{1-\frac{\alpha}{2}} - 3\epsilon_2/\eta_0\big)$. If $0 < \underline{\eta}(\mu_n, G) < \eta_0$, then $\underline{\eta}^+ = \eta_0$ and $\mathbb{P}_G\big(\underline{\eta}(\mu_n, G)\, X \geq c_{1-\frac{\alpha}{2}}\eta_0 - 3\epsilon_2\big) = 1 - F_{\mathcal{N}(0,1)}\big((c_{1-\frac{\alpha}{2}}\eta_0 - 3\epsilon_2)/\underline{\eta}(\mu_n, G)\big) \leq 1 - F_{\mathcal{N}(0,1)}\big((c_{1-\frac{\alpha}{2}}\eta_0 - 3\epsilon_2)/\eta_0\big) = 1 - F_{\mathcal{N}(0,1)}\big(c_{1-\frac{\alpha}{2}} - 3\epsilon_2/\eta_0\big)$. If $\underline{\eta}(\mu_n, G) = 0$, then $\underline{\eta}^+ = \eta_0$, $\underline{\eta}(\mu_n, G)\, X = 0$, and $c_{1-\frac{\alpha}{2}}\eta_0 - 3\epsilon_2 > 0$, and hence $\mathbb{P}_G\big(\underline{\eta}(\mu_n, G)\, X \geq c_{1-\frac{\alpha}{2}}\eta_0 - 3\epsilon_2\big) = 0$. By a symmetric argument, we can also show that $\mathbb{P}_G(\overline{v}(Q_G) > UB_{G,n})$ has exactly the same bound.

Thus, for any $n \geq N$, we obtain

$$\begin{aligned}
&\inf_{G\in\mathcal{G}} \mathbb{P}_G\big(\big[\underline{ATE^*}_G, \overline{ATE^*}_G\big] \subset [LB_{G,n}, UB_{G,n}]\big) \\
&\geq 1 - 2\left[\left(1 - F_{\mathcal{N}(0,1)}\Big(c_{1-\frac{\alpha}{2}} - \frac{3\epsilon_2}{\eta_0}\Big) + \epsilon_2\right) + 2\epsilon_1\right].
\end{aligned}$$

Since $\epsilon_1 > 0$ and $\epsilon_2 \in (0, c_{1-\frac{\alpha}{2}}\eta_0/3)$ are arbitrary, we conclude that

$$\liminf_{n\to\infty} \inf_{G\in\mathcal{G}} \mathbb{P}_G\big(\big[\underline{ATE^*}_G, \overline{ATE^*}_G\big] \subset [LB_{G,n}, UB_{G,n}]\big) \geq 1 - 2(1 - F_{\mathcal{N}(0,1)}(c_{1-\frac{\alpha}{2}})) = 1 - \alpha.$$

□

G.6 Proofs for Appendix B

Proof of Result B.1. Under the joint independence, it holds that $\mathbb{E}_F[D^m \mid Z^w, Z^m] = \mathbb{E}_F[D^m \mid Z^m]$. By the definition of covariance,

$$\begin{aligned}
\mathrm{Cov}_F\big(\mathbb{E}_F[D^w \mid Z^w],\, D^m\big) &= \mathbb{E}_F\big[\mathbb{E}_F[D^w \mid Z^w]\, D^m\big] - \mathbb{E}_F\big[\mathbb{E}_F[D^w \mid Z^w]\big]\, \mathbb{E}_F[D^m] \\
&= \mathbb{E}_F\big[\mathbb{E}_F[D^w \mid Z^w]\, \mathbb{E}_F[D^m \mid Z^m]\big] - \mathbb{E}_F\big[\mathbb{E}_F[D^w \mid Z^w]\big]\, \mathbb{E}_F\big[\mathbb{E}_F[D^m \mid Z^m]\big] \\
&= \mathrm{Cov}_F\big(\mathbb{E}_F[D^w \mid Z^w],\, \mathbb{E}_F[D^m \mid Z^m]\big).
\end{aligned}$$

Similarly, $\mathrm{Cov}_F\big(\mathbb{E}_F[D^m \mid Z^m], D^w\big) = \mathrm{Cov}_F\big(\mathbb{E}_F[D^m \mid Z^m], \mathbb{E}_F[D^w \mid Z^w]\big)$. Combining these two equations, we have

$$\mathrm{Cov}_F\big(\mathbb{E}_F[D^w \mid Z^w],\, D^m\big) = \mathrm{Cov}_F\big(\mathbb{E}_F[D^m \mid Z^m],\, D^w\big). \tag{G.12}$$

Since F is a matching, $D^m = T^w$ and $D^w = T^m$. Then, the law of total covariance gives $\mathrm{Cov}_F\big(\mathbb{E}_F[D^w \mid Z^w], D^m\big) = \mathrm{Cov}_F\big(\mathbb{E}_F[D^w \mid Z^w], T^w\big) = \mathrm{Cov}_F\big(\mathbb{E}_F[D^w \mid Z^w], \mathbb{E}_F[T^w \mid Z^w]\big)$. Simi-

larly, $\mathrm{Cov}_F\big(\mathbb{E}_F[D^m \mid Z^m], D^w\big) = \mathrm{Cov}_F\big(\mathbb{E}_F[D^m \mid Z^m], \mathbb{E}_F[T^m \mid Z^m]\big)$, and hence (B.1) holds. □

G.7 Proofs for Appendix D

G.7.1 Proofs for Appendix D.1

Lemma G.7. *Under Assumption D.1, for each $s \in \{w, m\}$ the function $\bar{H}^s$ is finite, continuously differentiable, and strictly convex on $\mathbb{R}^3$, with $\nabla \bar{H}^s = \mu^s$.*

Proof of Lemma G.7. We focus on the case with $s = w$. The man's case is symmetric.

Step 1 (Convexity). For each fixed value e of ε_1^w, the integrand $U \mapsto \max\{U_{d0}, U_{d1} - \varepsilon_2^w\}$ is convex since this is a pointwise maximum of affine functions. Taking the conditional expectation given $\varepsilon_1^w = e$ preserves convexity. Thus, $\Omega_d^w(\cdot; e)$ is also convex in U. The map $\max\{\Omega_0^w, \Omega_1^w - e\}$ is convex and nondecreasing in (Ω_0^w, Ω_1^w), and the composition of a convex nondecreasing function with convex functions is convex. Taking the expectation over ε_1^w preserves convexity, and therefore H^w is convex. Hence $\bar{H}^w$ is convex.

Step 2 (Finiteness). Since first choosing her own treatment d and then the treatment status d' of the partner she matches with is a feasible policy, it attains the payoff $U_{dd'} - \mathbb{1}(d = 1)\varepsilon_1^w - \mathbb{1}(d' = 1)\varepsilon_2^w$, giving $H^w(U) \geq \max_{dd'} U_{dd'} - \mathbb{E}|\varepsilon_1^w| - \mathbb{E}|\varepsilon_2^w|$. The reverse bound $H^w(U) \leq \max_{dd'} U_{dd'} + \mathbb{E}|\varepsilon_1^w| + \mathbb{E}|\varepsilon_2^w|$ is immediate, and hence H^w is finite by Assumption D.1(i). Since $\mathcal{Z}$ is finite, it follows that $\bar{H}^w$ is finite.

Step 3 (Differentiability). Because ε_2^w is continuously distributed, by the same logic as the Daly–Zachary–Williams theorem, we can show that $\Omega_d^w(\cdot; e)$ is differentiable with $\partial\Omega_d^w(\cdot; e)/\partial U_{d1} = \mathbb{P}(\varepsilon_2^w \leq \varsigma_d^w \mid \varepsilon_1^w = e)$ and $\partial\Omega_d^w(\cdot; e)/\partial U_{d0} = \mathbb{P}(\varepsilon_2^w > \varsigma_d^w \mid \varepsilon_1^w = e)$, where $\varsigma_d^w = U_{d1} - U_{d0}$.

The first-stage choice is a threshold rule and is almost surely unique at every U. The agent chooses $D^w = 1$ if and only if $\Delta(\varepsilon_1^w) \geq 0$, where $\Delta(e) := B(e) - e$ and $B(e) := \Omega_1^w(U; e) - \Omega_0^w(U; e) = \mathbb{E}[\psi(\varepsilon_2^w) \mid \varepsilon_1^w = e]$ with the 1-Lipschitz $\psi(x) := \max\{U_{10}, U_{11} - x\} - \max\{U_{00}, U_{01} - x\}$. By Assumption D.1(iii), $B(e) = \mathbb{E}[\psi(\Upsilon^w(e) + \nu^w)]$ is Lipschitz with constant at most $C_{\mathrm{Lip}}^w < 1$, and hence Δ is strictly decreasing and continuous with $\Delta(-\infty) = +\infty$ and $\Delta(+\infty) = -\infty$. Thus Δ has a unique zero $\chi^w(U)$, so that $D^w = \mathbb{1}(\varepsilon_1^w \leq \chi^w(U))$ and the tie in the first-stage (i.e., $\{\varepsilon_1^w = \chi^w(U)\}$) occurs with probability zero.

Then, by applying the envelope theorem to the outer expectation, for every U,

$$\frac{\partial H^w}{\partial U_{d1}} = \mathbb{E}_{\varepsilon_1^w}[\mathbb{1}\{D^w = d\}\,\mathbb{P}(\varepsilon_2^w \leq \varsigma_d^w \mid \varepsilon_1^w)] = \mathbb{P}(D^w = d, T^w = 1) = \varrho_{d1}^w(U),$$

and likewise $\partial H^w/\partial U_{d0} = \varrho^w_{d0}(U)$. Therefore, H^w is differentiable at every U with $\nabla H^w(U) = \varrho^w(U)$. Differentiating under the finite expectation over Z^w gives that $\bar{H}^w$ is continuously differentiable with $\nabla \bar{H}^w = \mu^w$.

Step 4 (Strict convexity). Recall $\bar{H}^w(V) = \mathbb{E}_{Z^w}[H^w(V - C^w(Z^w))]$. Strict convexity in V is preserved by the shift $V \mapsto V - C^w(Z^w)$ and by the expectation over Z^w. Therefore, it suffices to show that H^w is strictly convex on $\mathbb{R}^3$. As H^w is convex with $\nabla H^w = \varrho^w$, and a convex differentiable function is strictly convex if and only if its gradient is injective, it suffices to show that ϱ^w is injective.

By Step 3, under a utility vector $U = (U_{01}, U_{10}, U_{11})$, the first-stage threshold is $\chi = \chi^w(U)$ (i.e., the unique zero of Δ) and the second-stage threshold is $\varsigma_1 = U_{11} - U_{10}$ if $D^w = 1$ and $\varsigma_0 = U_{01}$ if $D^w = 0$, so that

$$\varrho^w_{11} = \mathbb{P}(\varepsilon^w_1 \le \chi,\, \varepsilon^w_2 \le \varsigma_1), \qquad \varrho^w_{10} = \mathbb{P}(\varepsilon^w_1 \le \chi,\, \varepsilon^w_2 > \varsigma_1), \qquad \varrho^w_{01} = \mathbb{P}(\varepsilon^w_1 > \chi,\, \varepsilon^w_2 \le \varsigma_0).$$

Now take any utility vector $U' \neq U$. Let $\chi' = \chi^w(U')$, $\varsigma_1' = U'_{11} - U'_{10}$, and $\varsigma_0' = U'_{01}$. We wish to show that the two vectors cannot share all three thresholds. Suppose that $\varsigma_0 = \varsigma_0'$ and $\varsigma_1 = \varsigma_1'$. These two equalities fix $U'_{01} = U_{01}$ and $U'_{11} - U'_{10} = U_{11} - U_{10}$, leaving $U' = U + (0, c, c)$ for some $c \neq 0$. Adding c to U_{10} and U_{11} raises ψ, and hence B and Δ, defined in Step 3, by c. Writing Δ_U for this map at U, we have $\Delta_{U'} = \Delta_U + c$. Since $\Delta_U(\chi) = 0$, we get $\Delta_{U'}(\chi) = c \neq 0$. Therefore $\chi' \neq \chi$. Hence at least one of the three thresholds differs, and we consider the following exhaustive three cases. (i) If $\chi \neq \chi'$, then $\varrho^w_{11} + \varrho^w_{10} = \mathbb{P}(\varepsilon^w_1 \le \chi)$ differs, because ε^w_1 has a strictly positive density. (ii) If $\chi = \chi'$ but $\varsigma_1 \neq \varsigma_1'$, then $\varrho^w_{11} = \mathbb{P}(\varepsilon^w_1 \le \chi,\, \varepsilon^w_2 \le \varsigma_1)$ differs, because the density is strictly positive on $\{\varepsilon^w_1 \le \chi\}$. (iii) If $\chi = \chi'$ and $\varsigma_1 = \varsigma_1'$ but $\varsigma_0 \neq \varsigma_0'$, then $\varrho^w_{01} = \mathbb{P}(\varepsilon^w_1 > \chi,\, \varepsilon^w_2 \le \varsigma_0)$ differs. Thus, in every case $\varrho^w(U) \neq \varrho^w(U')$. Hence ϱ^w is injective, and we conclude that H^w and $\bar{H}^w$ are strictly convex on $\mathbb{R}^3$. □

Proof of Proposition D.1. Define $\Phi(V) := \bar{H}^w(V) + \bar{H}^m(\Psi - V)$ on $\mathbb{R}^3$. By Lemma G.7, $\bar{H}^w$ and $\bar{H}^m$ are continuously differentiable and strictly convex, and hence Φ is continuously differentiable and strictly convex on $\mathbb{R}^3$, with

$$\nabla\Phi(V) = \mu^w(V) - \mu^m(\Psi - V).$$

Thus, $\nabla\Phi(V) = 0$ if and only if $\mu^w(V) = \mu^m(\Psi - V)$, which is exactly the aggregate equilibrium condition with $V^m = \Psi - V$. Therefore, the aggregate equilibria correspond to the solutions of

$\nabla\Phi(V) = 0$.

Now, it remains to show that $\nabla\Phi(V) = 0$ has exactly one solution. By the lower bound in the proof of Lemma G.7 and the finiteness of $\mathcal{Z}$, there is a constant B with

$$\Phi(V) \geq \max_{dd'} V_{dd'} + \max_{dd'}\left(\Psi_{dd'} - V_{dd'}\right) - B \geq \left(\max_{dd'} V_{dd'} - \min_{dd'} V_{dd'}\right) + \min_{dd'} \Psi_{dd'} - B,$$

and since $V_{00} = 0$ gives $\max_{dd'} V_{dd'} - \min_{dd'} V_{dd'} \geq \|V\|_\infty$, it follows that $\Phi(V) \to +\infty$ as $\|V\| \to \infty$. Hence Φ attains a minimum on $\mathbb{R}^3$. By strict convexity, the minimizer is unique and is the unique solution of $\nabla\Phi(V) = 0$. □

Lemma G.8. *Under Assumption D.1, for each $s \in \{w, m\}$ and every utility vector U, the two-stage optimal choice follows the sequential threshold-crossing rule $D^s = \mathbb{1}(\varepsilon_1^s \leq \chi^s(U))$ and $T^s = \mathbb{1}(\varepsilon_2^s \leq \varsigma_{D^s}^s)$.*

Proof of Lemma G.8. It is immediate from the proof of Lemma G.7. □

Proof of Proposition D.2. By Lemma G.8, K^s is the threshold-crossing rule $D^s = \mathbb{1}(\varepsilon_1^s \leq \chi^s(U^s(Z^s)))$ and $T^s = \mathbb{1}(\varepsilon_2^s \leq \varsigma_{D^s}^s(Z^s))$, which is exactly the functional form (4.1) of Remark 4.2, with first-stage threshold $h_1^s(z) = \chi^s(U^s(z))$ and second-stage threshold $h_2^s(z, d) = \varsigma_d^s(z)$.

Now $(D^w, D^m) = (T^m, T^w)$ holds F-almost surely, and hence F is a matching (Definition 3.1). The equilibrium matching condition of Definition 3.2 holds because $(D^s, T^s) = K^s(Z^s)$ by construction. Since we couple the two sides independently within each pair (d, d'), F satisfies matching on observables (Definition 3.3). Finally, for each z, the two-stage optimal policy producing $(D^s(z), T^s(z))$ depends only on ε^s, which implies that the response type K^s is a deterministic function of ε^s. Thus, $(K^s, \boldsymbol{\phi}^s)$ is a function of $(\varepsilon^s, \boldsymbol{\phi}^s)$. Then, $(\varepsilon^s, \boldsymbol{\phi}^s) \perp\!\!\!\perp Z^s$ (Assumption D.1(ii)) gives $(K^s, \boldsymbol{\phi}^s) \perp\!\!\!\perp Z^s$ under F, which is the independence of instruments (Assumption 3.2). Therefore F satisfies Assumptions 3.2 and 3.3. □

G.7.2 Proofs for Appendix D.2

Lemma G.9 (Surplus function)**.** *Under Assumption D.2, for each $s \in \{w, m\}$ the function $\bar{H}^s$ is finite, continuously differentiable, and strictly convex on $\mathbb{R}^3$, with $\partial\bar{H}^s/\partial V_{dd'}^s = \mu_{dd'}^s(V^s)$ for $(d, d') \neq (0, 0)$.*

Proof of Lemma G.9. Step 1 (Convexity). For each realization of (Z^s, ε^s), the map $V^s \mapsto V^s_{ll'} - C^s_{ll'}(Z^s) + \varepsilon^s_{ll'}$ is affine in V^s for every (l, l'). Then, the maximum over (l, l') is a pointwise maximum of affine functions of V^s, and hence convex. Taking expectations preserves convexity, and hence $\bar{H}^s$ is convex on $\mathbb{R}^3$.

Step 2 (Finiteness). Fix any (d, d'). Then we have that $\bar{H}^s(V^s) \geq V^s_{dd'} - \max_{dd',z} |C^s_{dd'}(z)| - \mathbb{E}\|\varepsilon^s\|$, which is finite by Assumption D.2(i). Similarly, we obtain the upper bound $\bar{H}^s(V^s) \leq \max_{dd'} V^s_{dd'} + \max_{dd',z} |C^s_{dd'}(z)| + \mathbb{E}\|\varepsilon^s\|$.

Step 3 (Gradient). By Assumption D.2(ii), the argmax of $\{V^s_{ll'} - C^s_{ll'}(Z^s) + \varepsilon^s_{ll'}\}$ is almost surely unique, and hence the indicator of the maximizing (d, d') is well-defined. Then, the Daly–Zachary–Williams theorem yields differentiability of $\bar{H}^s$ with $\partial \bar{H}^s / \partial V^s_{dd'} = \mu^s_{dd'}(V^s)$ for $(d, d') \neq (0, 0)$. Continuity of μ^s in V^s from Assumption D.2(ii) gives that $\bar{H}^s$ is continuously differentiable on $\mathbb{R}^3$.

Step 4 (Strict convexity). Suppose $\bar{H}^s$ is not strictly convex on $\mathbb{R}^3$, so that it fails to be strictly convex on some segment $[V_1, V_2] \subseteq \mathbb{R}^3$ with $V_1 \neq V_2$. A convex function is strictly convex on a segment if and only if it is not affine on any nondegenerate subsegment. Hence $\bar{H}^s$ is affine on some nondegenerate subsegment $[V_a, V_b] \subseteq [V_1, V_2] \subseteq \mathbb{R}^3$ with $V_a \neq V_b$. On that subsegment, $\nabla \bar{H}^s = (\mu^s_{dd'})_{(d,d') \neq (0,0)}$ is constant. Hence $\mu^s(V_a) = \mu^s(V_b)$ with $V_a \neq V_b$ in $\mathbb{R}^3$, which contradicts Assumption D.2(iii). Therefore, we conclude that $\bar{H}^s$ is strictly convex on $\mathbb{R}^3$. □

Proof of Proposition D.3. Let $\Phi(V) := \bar{H}^w(V) + \bar{H}^m(\Psi - V)$ on $\mathbb{R}^3$. By Lemma G.9, Φ is also continuously differentiable and strictly convex with $\nabla \Phi(V) = \mu^w(V) - \mu^m(\Psi - V)$. Then, as in the proof of Proposition D.1, we can show that there exists a unique aggregate equilibrium. □

Proof of Proposition D.4. The argument is identical to the proof of Proposition D.2, with the single-stage optimizer $K^s(z) = \operatorname{argmax}_{(d,d') \in \mathcal{D}^2} \{U^s_{dd'}(z) + \varepsilon^s_{dd'}\}$ in place of the sequential policy. □