Efficient estimation of average match output under conditionally exogenous matching

Bryan S. Graham*

23 September 2013

Abstract

Consider two heterogenous populations of agents who, when matched, jointly produce an output. For example, teachers and classrooms of students together produce achievement, husbands and wives jointly raise children, and assembly plants and their managers produce parts, cars, etc. Let $W \in \mathbb{W} = \{w_1, \ldots, w_J\}$ and $X \in \mathbb{X} = \{x_1, \ldots, x_K\}$ denote agent types in the two populations. Consider the following matching mechanism: take a random draw from the $W = w_j$ subgroup of the first population and match her with an independent random draw from the $X = x_k$ subgroup of the second population. Let $\beta(w_j, x_k)$, the *average match function* (AMF), denote the expected output associated with this match. I show that the AMF is identified when matching is conditionally exogenous and calculate its semiparametric efficiency bound. I propose an efficient estimator and use it, in an illustrative application, to study the relationship between assortative marriage by education and children's education in Brazil. The analysis suggests that, under the maintained assumptions, less parental sorting by education would raise educational attainment for children of poorly educated mothers substantially and modestly lower it for children of better educated mothers. On net, marital sorting increases inequality and lowers the population wide average educational attainment of children.

*Department of Economics, University of California - Berkeley, 530 Evans Hall #3380, Berkeley, CA 94720-3888, e-mail: bgraham@econ.berkeley.edu, web: http://emlab.berkeley.edu/bgraham/Preliminary versions of the research reported here was presented at the 2012 European Summer Meetings of the Econometric Society, the December 2012 SFB 884 Research Conference on the Evaluation of Political reforms at the University of Mannheim, and at seminars hosted by the University of California - Berkeley and the University of Southern California. I am grateful to Stephane Bonhomme, Richard Blundell, Konrad Menzel and James Powell for useful discussions. Financial support from the National Science Foundation (SES #0820361) is gratefully acknowledged. All the usual disclaimers apply. **This draft is preliminary and incomplete, please e-mail the author for the latest draft.**

Many economic outcomes are the product of collaboration across two (or more) heterogeneous agents. A school district matches teachers, with varying skills, background and other attributes, to classrooms composed of different types of students in order to "produce" achievement (e.g., Boyd, Lankford, Loeb and Wyckoff, 2013). Corporations match managers with plants in an attempt to maximize profit. Here both managers and plants may be heterogenous. For example only some plants may be unionized or have different histories of industrial action. Managers may vary in their experience, education and leadership skill. Other examples include insurers and hospitals, which form contractual agreements to provide health care (e.g., Ho, 2009) and marriages between men and women who subsequently raise children (Kremer, 1997).

In the standard empirical production function model a *single heterogenous agent* utilizes *homogenous inputs* to engage in production (e.g., Chamberlain, 1984; Griliches and Mairesse, 1996; Olley and Pakes, 1996). The canonical example is a population of firms with varying levels of total factor productivity which combine homogenous labor, capital and land to produce output. The basic program evaluation problem also falls into this framework; while agents may be heterogeneous in their response to a given treatment, the treatment itself is considered homogenous (e.g., Imbens, 2004; Heckman and Vytlacil 2007a,b).

I consider settings where production involves two heterogenous agents. This heterogeneity may have both observed and unobserved components. The goal is to determine how match output varies with the observables of the two paired agents. For example a superintendent may wish to understand how student achievement varies with teacher experience and class size. The question is not straightforward because *both* the distribution of unobserved teacher and student characteristics may vary systematically with experience and/or class size. Using knowledge of the mapping from agent characteristics into match output, a social planner can engineer reallocations. For example, a superintendent may choose to assign less experienced teachers to smaller classrooms. Graham, Imbens and Ridder (2007, forthcoming) also study reallocations, but under stronger assumptions and a less rich heterogeneity structure than allowed for here.

Section 1 introduces the notion of an *average match function* (AMF) and compares it with more familiar "single agent" estimands such as the average structural function (ASF) of Blundell and Powell (2005) (cf., Chamberlain, 1984; Wooldridge, 2005). I also introduce a social planning problem and show how the AMF is a key component of it.

Section 2 provides a (point) identification result for the AMF. A key requirement is that the status quo matching satisfies a conditional exogeneity assumption. This assumption may be viewed as a multi-agent analog of the familiar "selection on observables" or "unconfoundedness" assumption of the program evaluation literature (e.g., Heckman and Robb, 1985;

Imbens, 2004).

Section 3 presents the semiparametric efficiency bound for the AMF under the exogeneity assumption. Hirano and Porter (2009) show that efficient estimation of the conditional average treatment effect function (CATE) is needed to ensure that conditional empirical success (CES) treatment assignment rules are minimax optimal in large samples. In the present setting a feasible version of the social planner's problem would replace the unknown AMF with an efficient estimate (cf., Graham, Imbens and Ridder, 2007). A second reason for study semiparametric efficiency is that it may aid in experimental design (cf., Hahn, Hirano, Karlan, 2011).

Section 4 applies the proposed methods in a study of the relationship between parental schooling and child's education attainment using a large dataset from Brazil. A long research literature in sociology and economics studies the relationship between assortative marriage and intergenerational mobility (e.g., Kremer, 1997). If both parents' education strongly influences their children's educational attainment, and marriage is assortative by education, then inequality in the next generation will be higher than it would be in a counterfactual world of non-assortative or "random" marriage. While the direct policy implications of this analysis is probably best left for dystopian science fiction, the analysis does illuminate one potentially important source of inequality and low mobility.

1 Estimands

There are two heterogenous populations. Examples include women and men in a marriage market, workers and firms in a labor market, or teachers and classrooms of students in the elementary school setting. For concreteness I will adopt the last of these examples in what follows.

Let $W \in \mathbb{W} = \{w_1, \ldots, w_J\}$ denote a teacher's observable "type". The support points of W may encode, for example, different unique combinations of years of teaching experience, levels of education, race and gender. Teachers of the same type may differ in terms of the unobserved characteristic U. The dimension of U is unrestricted. Teachers are a heterogenous population. While the econometrician does not observe U, she does observe the vector of proxies, R; which may have both discrete and continuous components. If U is teacher "ability", then R might include a licensure test score. The properties of R will be formally outlined in Section 2. All diversity in the population of teachers is captured by the vector $(W, R', U')'$. I index a random draw from the population of teachers by the subscript i, such that $(W_i, R_i', U_i')'$ corresponds to the i^{th} random draw. A generic random draw is denoted

by $(W, R', U')'$ (i.e., subscripts omitted).

Let $X \in \mathbb{X} = \{x_1, \ldots, x_K\}$ denote a classroom's observable type. The K types of classroom could enumerate different unique combinations of classroom size and/or student gender/ethnic mix. Unobserved classroom characteristics are given by V, with S the corresponding vector of proxies.

I index a random draw from the population of classrooms by the superscript h, such that $\left(X^h, S^h, V^h\right)$ equals measured and unmeasured characteristics of the h^{th} random draw. The sub- and super-script notation emphasizes the two-population aspect of the problem.

Teachers and classrooms of students are matched through some process, restrictions on this process will be imposed in Section 2. Together they jointly produce an output, say, student achievement. Associated with each teacher-classroom pair is a *potential* or *conjectural* output (Holland, 1986; Manski, 2007). Let $Y_i(h)$ denote the potential outcome when the i^{th} teacher is matched with the h^{th} classroom. Consider two classrooms h and h' with $X^h = X^{h'}$. A key feature of the current problem is that, in general,

$$Y_i(h) \neq Y_i(h'),$$

because even classrooms of the same observed type are heterogenous in terms of the unobserved vector V.[1] To clarify this point, consider the production function representation of $Y_i(h)$:

$$Y_i(h) = g\left(W_i, X^h, U_i, V^h\right). \tag{1}$$

Equation (1) is a production function with two *heterogenous* inputs. This set-up contrasts with the standard single agent production problem (e.g., Chamberlain, 1984, Griliches and Mairesse, 1996; Olley and Pakes, 1996). In that problem heterogenous agents (e.g., firms with varying levels of productivity) employ homogenous inputs (e.g., "capital"). To connect the current problem with this single agent analog, assume that all classroom heterogeneity is captured by X^h, such that V^h is degenerate at $V^h = \underline{v}$. Degeneracy of V^h implies that for a given teacher (i.e., firm) all classrooms homogenous in $X^h = x$ (i.e., the input level) yield identical output. This yields the simplification of (1):

$$Y_i(x) = g(W_i, x, U_i, \underline{v}), \ x \in \mathbb{X}, \tag{2}$$

which coincides with a standard potential outcomes representation of a production function (e.g., Manski, 2007; Chapter 7).

[1] This is a violation of the homogenous treatment or stable unit treatment value assumption (SUTVA) routinely made in the program evaluation literature.

A comparison of (1) and (2) clarifies that the heterogenous nature of the "input", X, distinguishes the present problem from the textbook production function one.

In the single agent problem interest typically centers on the average structural function (ASF) (cf., Chamberlain, 1984; Blundell and Powell, 2003; Wooldrige, 2005; Imbens and Newey, 2009)

$$\mathbb{E}\left[Y_i\left(x\right)\right] = \mathbb{E}\left[g\left(W_i, x, U_i, \underline{v}\right)\right] = \sum_{j=1}^{J} \int g\left(w_j, x, u, \underline{v}\right) f_{W,U}\left(w_j, u\right) \mathrm{d}u. \tag{3}$$

Equation (3) coincides with the expected output when a random draw from the population of teachers is assigned to a classroom of type $X^h = x$, when there is no within-classroom-type heterogeneity (i.e, if V^h is degenerate at $V^h = \underline{v}$).

When classrooms are heterogenous, as is presumed here, we can modify (3) by additionally integrating over the conditional distribution of V:

$$\gamma\left(x\right) = \sum_{j=1}^{J} \int g\left(w_j, x, u, v\right) f_{W,U}\left(w_j, u\right) f_{V|X}\left(v|\, x\right) \mathrm{d}u\mathrm{d}v. \tag{4}$$

Estimand (4) gives the expected outcome when a random draw from the population of teachers is matched to an independent random draw from the *subpopulation* of classrooms with $X^h = x$. In the context of the teacher-classroom example it provides a measure of how student achievement changes with classroom type that controls, in an average way, for observed and unobserved teacher characteristics. This control is engineered by integrating the production function with respect to the joint (marginal) distribution of W and U; as in a partial mean (Newey, 1994a).

Consider the difference

$$\gamma\left(x\right) - \gamma\left(x'\right).$$

This difference gives the expected achievement gap across classrooms of types x versus x', *when teacher assignment to classrooms is random.* It is *not* a "causal effect" of x; in the present setting X is non-manipulable. One manifestation of this perspective is that V is averaged out in (4) using its distribution conditional of $X = x$. It may be that predominantly minority classrooms, for example, systematically differ from other classrooms in terms of V (e.g., unmeasured parental characteristics).

I will call $\gamma\left(x\right)$ the average no sorting outcome (ANSO). It gives the average outcome for classroom type $X = x$, when teachers are randomly assigned to classrooms (i.e., when there is no sorting).

While $\gamma(x)$ provides a coherent measure of how outcomes vary across classroom types which controls for variation in teacher quality, it does not provide a measure of how how outcomes vary across different *combinations* of teacher and classroom types.

A social planner may wish to change the joint distribution of W and X in order to maximize average outcomes. To do so she requires (partial) knowledge of the structural mapping from $W = w$ and $X = x$ into outcomes. For example a school district may employ a set of teachers with varying characteristics encoded into W. These teachers must be assigned to a set of classrooms of varying composition, encoded into X. If the district wishes to maximize, say, average student achievement, it must know the mapping from different (w, x) combinations into achievement, Y. One approach to characterizing this mapping would be to average $g(w, x, u, v)$ with respect to the joint distribution of U, V:

$$\int\int g(w, x, u, v) f_{U,V}(u, v) \, \mathrm{d}u\mathrm{d}v. \tag{5}$$

This would correspond to the average structural function (ASF) as defined in Blundell and Powell (2003) with the relevant population now being that of *matches* (i.e., existing pairings of teachers and students).[2] If both w and x are manipulable, then (5) coincides with the expected outcome when a random draw from the population of matches is assigned the input combination (w, x). Here, however, manipulations of W and/or X for a *given* match are not of interest. Instead the focus is on manipulations of the joint distribution of (W, X) that leave the marginal distributions of W and X intact: *reallocations*. Equation (5) is not helpful for evaluating reallocations, indeed the heterogeneity distribution averaged over in (5), $F_{U,V}$, is the consequence of a specific allocation and would not, in general, be invariant to reallocations of teachers to classrooms.

For these reasons I introduce a different estimand. Consider the following thought experiment. A social planner takes a random draw from the subpopulation of type $W_i = w$ teachers. She takes an independent random draw from the subpopulation of type $X^h = x$ classrooms. The expected outcome associated with pairing together these two draws is

$$\beta(w, x) = \int\int g(w, x, u, v) f_{U|W}(u|\, w) f_{V|X}(v|\, x) \, \mathrm{d}u\mathrm{d}v. \tag{6}$$

I call (6) the *average match function* (AMF). Note that no presumption of independence between W and U or X and V is made. It may be that the distribution of teacher ability, U, varies systematically with observed years of teacher experience, W. To repeat, such dependence does not cause problems in the present setting because manipulation of W is

[2]Note the relevant population for the ASF defined in equation (3) is teachers.

not of interest. Reallocations leave the joint distribution of W and U unchanged.

The difference

$$\beta(w,x)-\beta(w,x')$$

gives the expected change in output when a type $W_i = w$ teacher is assigned to a type $X^h = x$ instead of a type $X^h = x'$ classroom (where both the teacher and classrooms are independent random draws from the appropriate subpopulation).

The discrete analog of a cross partial derivative (cf., Topkis, 1998), for $w > w'$ and $x > x'$,

$$\beta(w,x)-\beta(w,x')-[\beta(w',x)-\beta(w',x')]$$

is a local measure of complementarity between w and x (cf., Graham, 2011). The complementarity properties of $\beta(w,x)$ are of central interest.

Note that the average no sorting outcome (ANSO) is connected to the average match function (AMF):

$$\gamma(x)=\sum_{j=1}^{J}\beta(w_j,x)\,\rho_j,$$

with $\rho_j = \int f_{W,U}(w_j,u)\,\mathrm{d}u$ the marginal frequency of type j teachers.

1.1 Social planner's problem

Consider a social planner. The social planner knows $\beta(w,x)$ for all $(w,x)\in\mathbb{W}\times\mathbb{X}$ (perhaps up to sampling uncertainty). She also knows the marginal distributions of teacher and classroom type, respectively $\rho=(\rho_1,\ldots,\rho_J)'$ for $\rho_j=\Pr(W_i=w_j)$ and $\lambda=(\lambda_1,\ldots,\lambda_K)'$ for $\lambda_k=\Pr\left(X^h=x_k\right)$ (again perhaps up to sampling uncertainty). She does not observe $(R_i',U_i')'$ or $\left(S^h,V^h\right)'$ or is unable/unwilling to act on this knowledge if she does. Put differently, the planner is constrained to consider doubly randomized allocations (Graham, 2008; 2011). Each (i,h) pairing is composed of two independent random draws from the relevant subpopulations of teacher and classroom types.

Let $\pi_{jk}=\Pr(W=w_j,X=x_k)$ for $j=1,\ldots J$ and $k=1,\ldots,K$. The planner's problem is to choose an $\pi=(\pi_{11},\ldots,\pi_{1K},\ldots,\pi_{J1},\ldots,\pi_{JK})'$ that maximizes expected output

$$\theta(\pi)=\sum_{j=1}^{J}\sum_{k=1}^{K}\beta(w_j,x_k)\,\pi_{jk} \tag{7}$$

Table 1: The structure of feasible assignments

Teachers/Classrooms	x_1	$\cdots$	x_{K-1}	x_K	$f_W(w)$
w_1	π_{11}	$\cdots$	π_{1K-1}	$\rho_1 - \sum_{k=1}^{K-1} \pi_{1k}$	ρ_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
w_{J-1}	π_{J-11}	$\cdots$	π_{J-1K-1}	$\rho_{J-1} - \sum_{k=1}^{K-1} \pi_{J-1k}$	ρ_{J-1}
w_J	$\lambda_1 - \sum_{j=1}^{J-1} \pi_{j1}$	$\cdots$	$\lambda_{K-1} - \sum_{j=1}^{J-1} \pi_{jK-1}$	$1 - \sum_{j=1}^{J-1} \rho_j - \sum_{k=1}^{K-1} \lambda_k + \sum_{j=1}^{J-1} \sum_{k=1}^{K-1} \pi_{jk}$	ρ_J
$f_X(x)$	λ_1		λ_{K-1}	λ_K	

Notes: The $j = 1, \ldots, J$ types of teachers are enumerated in the first column, with the marginal frequency of each type given in the last column. The $k = 1, \ldots, K$ types of classrooms are enumerated in the first row, with the marginal frequency of each type given in the last row. The joint distribution of teachers and classrooms is characterized by the interior probabilities. The feasibility constraints are used to reduce the parameterization of an assignment to $(J-1)(K-1)$ probabilities.

subject to the $J+K$ feasibility constraints:

$$\begin{aligned} \sum_{k=1}^{K} \pi_{jk} &= \rho_j, \ j=1,\ldots,J \\ \sum_{j=1}^{J} \pi_{jk} &= \lambda_k, \ k=1,\ldots,K. \end{aligned} \tag{8}$$

See Graham, Imbens and Ridder (2007). Since $\sum_{j=1}^{J}\sum_{k=1}^{K}\pi_{jk}=1$ one constraint is redundant. Table 1 depicts the structure of a feasible assignment. By substituting out the feasibility constraints, an assignment can be represented in terms of $(J-1)(K-1)$ probabilities.

Graham (2011) shows that the difference between two doubly randomized allocations, π' and π is given by

$$\theta(\pi')-\theta(\pi)=\sum_{j=1}^{J-1}\sum_{k=1}^{K-1}\left(\pi'_{jk}-\pi_{jk}\right)\left(\beta(w_J,x_K)-\beta(w_J,x_k)-\left[\beta(w_j,x_K)-\beta(w_j,x_k)\right]\right). \tag{9}$$

Equation (9) indicates that the average outcome properties of an allocation depend critically on the complementarity properties of the average match function (AMF). Of particular interest is the difference between a candidate assignment π and the completely random matching $\pi^{\text{rdm}}_{jk}=\rho_j\lambda_k$ for all $j=1,\ldots,J$ and $k=1,\ldots,K$:

$$\theta(\pi')-\theta\left(\pi^{\text{rdm}}\right)=\sum_{j=1}^{J-1}\sum_{k=1}^{K-1}\left(\pi'_{jk}-\rho_j\lambda_k\right)\left(\beta(w_J,x_K)-\beta(w_J,x_k)-\left[\beta(w_j,x_K)-\beta(w_j,x_k)\right]\right). \tag{10}$$

Equation (9) suggests that outcome-maximizing assignments will tend to be assortative ($\pi'_{jk}>\rho_j\lambda_k$) in regions of complementarity ($\beta(w_J,x_K)-\beta(w_J,x_k)-[\beta(w_j,x_K)-\beta(w_j,x_k)]>0$)

and anti-assortative ($\pi'_{jk}<\rho_j\lambda_k$) in regions of substitutability ($\beta(w_J,x_K)-\beta(w_J,x_k)-[\beta(w_j,x_K)-\beta(w_j,x_k)]<0$).

[Elaborate on connection to the maximum score criterion function of Manski (1975, 1985); cf., Graham (2011)].

2 Identification

Prior to formulating conditions for identification, it is useful to consider the anatomy of the identification problem. For simplicity I will assume that the populations of teachers and classrooms are equally sized and that all units are matched in the status quo assignment. Let $h = m(i)$ equal the classroom assigned to teacher i under the status quo. *Observed* output is therefore given by

$$Y_i = g\left(W_i, X^{m(i)}, U_i, V^{m(i)}\right).$$

In what follows I will write $X_i = X^{m(i)}$ and $V_i = V^{m(i)}$ to simplify the notation (e.g., $Y_i = g(W_i, X_i, U_i, V_i)$). Put differently, the i subscript will be used to index both teachers and teacher-classroom matches (the latter in the status quo assignment).

Assumption 1. *(Random Sampling)* $\{Z_i\}_{i=1}^{\infty}$ *for* $Z_i = (X_i, W_i, R_i', S_i', Y_i)$ *is random sequence drawn from the status quo population of matches with distribution function* F*.*

Let $m(w,x) = \mathbb{E}[Y|\, W = w, X = x]$ denote the mean regression function of Y given $W = w$ and $X = x$. The difference between $m(w,x)$ and the AMF is given by

$$\begin{aligned} m(w,x) - \beta(w,x) &= \int g(w,x,u,v)\left\{f_{U,V|W,X}(u,v|\, w,x) - f_{U|W}(u|\, w)\, f_{V|X}(v|\, x)\right\} \mathrm{d}u\mathrm{d}v \\ &= \int g(w,x,u,v)\left\{f_{U,V|W,X}(u,v|\, w,x) - f_{U|W,X}(u|\, w,x)\, f_{V|W,X}(v|\, w,x)\right\} \mathrm{d}u\mathrm{d}v \\ &\quad + \int g(w,x,u,v)\left\{f_{U|W,X}(u|\, w,x)\, f_{V|W,X}(v|\, w,x)\right. \\ &\quad \left. - f_{U|W}(u|\, w)\, f_{V|X}(v|\, x)\right\} \mathrm{d}u\mathrm{d}v, \end{aligned} \tag{11}$$

which indicates that $m(w,x)$ may differ from $\beta(w,x)$ for two distinct reasons. First, there may be dependence between U and V within w-by-x cells. This would occur if, for example, high U teachers systematically match with high V classrooms. This is "bias" due to matching on unobservables. Second, X may help to predict U (given W) and/or W may help to predict V (given X). This is also a type of matching bias. It would arise, for example, if predominately white classrooms, say type $X = x$, tended to match with teachers with higher unobserved quality U, than predominately black classrooms, say type $X = x'$; even when both types of classrooms choose similar teachers in terms of the observed attribute W.

If the distribution of V is degenerate with mass point $\underline{v}$, then (11) specializes to

$$m(w,x) - \beta(w,x) = \int g(w,x,u,\underline{v})\left\{f_{U|W,X}(u|\, w,x) - f_{U|W}(u|\, w)\right\} \mathrm{d}u,$$

which is a traditional expression for "selection bias". In this case the divergence between $m(w,x) - \beta(w,x)$ arises because the X may covary with U, even conditional on W. The case with V degenerate was the one explored in Graham, Imbens and Ridder (forthcoming). If, additionally X is degenerate at $\underline{x}$, then

$$m(w,\underline{x}) - \beta(w,\underline{x}) = \int g(w,\underline{x},u,\underline{v}) \left\{ f_{U|W}(u|w) - f_{U|W}(u|w) \right\} \mathrm{d}u = 0,$$

underscoring that dependence between W and U is not of concern, since this joint distribution is unaffected by reallocations.

In order to develop a constructive identification result for $\beta(w,x)$, and also $\gamma(w)$, I make two additional assumptions. The first restricts the structure of the status quo matching

Assumption 2. *(Conditionally Exogenous Matching) The joint distribution of* (W,X,R,S,U,V) *factors*

$$F_{W,X,R,S,U,V}(w,x,r,s,u,v) = F_{W,X,R,S}(w,x,r,s)\, F_{U|W,R}(u|w,r)\, F_{V|X,S}(v|x,s).$$

Note that

$$\begin{aligned} f_{U,V|W,X}(u,v|w,x) &= \int\int f_{U|W,R}(u|w,r)\, f_{V|X,S}(v|x,s)\, f_{R,S|W,X}(r,s|w,x)\, \mathrm{d}r\mathrm{d}s \\ &\neq f_{U|W}(u|w)\, f_{V|X}(v|x). \end{aligned} \tag{12}$$

Under Assumption 2 there may be matching on unobservables within $W = w$ by $X = x$ cells. However within $W = w$, $R = r$ by $X = x$, $S = s$ cells there is no matching on unobservables. I call this conditionally exogenous matching. Assumption 2 may hold for two reasons. First, it can hold by design. In that case the researcher chooses a feasible joint distribution for (W,X,R,S), but forms a $(W,X,R,S) = (w,x,r,s)$ match by taking a random draw from the subpopulation of teachers homogenous in $(W,R) = (w,r)$ and matching her with an independent random draw from the subpopulation of classrooms homogenous in $(X,S) = (x,s)$. This is a doubly randomized assignment scheme (cf., Graham, 2008, 2011). Note, as indicated by (12), this scheme does allow for matching on unobservables within w-by-x sells. As shown below, the presence of the proxies R and S allows the researcher to "undo" this sorting in order to recover the AMF.

Second Assumption 2 can be consistent with settings where agents match optimally. This was shown formally by Graham, Imbens and Ridder (forthcoming) in the context of a version of the empirical transferable utility matching model of Choo and Siow (2006a,b). A heuristic explanation of their result is as follows. If both U and V are unobserved by agents prior to

matching, then U and V will be conditionally independent given (W, X, R, S). Furthermore (X, S) will have no predictive power for U given (W, R). Candidate partners will use teacher characteristics, including the proxy R, to forecast unobserved teacher ability, U. Candidate classroom characteristics (X, S) have no such forecast value. For symmetric reasons (W, R) will have no predictive power for V given (X, S). The connection between agent's information sets and versions of input exogeneity is, of course, not unique to the present problem (e.g., Chamberlain, 1984).

Identification will also require a support condition. Let, in an abuse of notation,

$$\begin{aligned} p_w(r) &= \Pr(W = w | R = r) \\ p_x(s) &= \Pr(X = x | S = s) \\ p_{wx}(r, s) &= \Pr(W = w, X = x | R = r, S = s). \end{aligned}$$

In certain instances I will also use the notation $p_j(r) = \Pr(W = w_j | R = r)$, $p_k(s) = \Pr(X = x_k | S = s)$, and $p_{jk}(r, s) = \Pr(W = w_j, X = x_k | R = r, S = s)$ for $j = 1, \ldots, J$ and $k = 1, \ldots, K$.

Let

$$\mathbb{S}_{RS}(w, x) = \{r, s : p_w(r) p_x(s) > 0\}$$

denote the feasible joint support of R and S across the set of $W = w$ to $X = x$ matches. This set contains all logically possible combinations of $R = r$ and $S = r$ that might be observed in a $W = w$ to $X = x$ match.

Assumption 3. *(Overlap)* $p_{wx}(r, s) \geq \kappa > 0$ for all $(r, s) \in \mathbb{S}_{RS}(w, x)$.

As suggested by its label, Assumption 3, is related to the overlap assumption made in the program evaluation literature (e.g., Hahn, 1998; Imbens, 2004). It ensures that all logically possible combinations of $R = r$ and $S = s$ for a given (w, x) pair are in fact observed in the set of (w, x) status quo matches.

Consider the mean regression of Y given W, X, R, S

$$\mathbb{E}[Y | W = w, X = x, R = r, S = s] \stackrel{\text{def}}{\equiv} q(w, x, r, s). \tag{13}$$

Note that $q(w, x, r, s)$ is a structural object under (1) and Assumptions 1, 2 and 3. Specifically, the difference

$$q(w, x', r, s') - q(w, x, r, s),$$

gives the expected change in output when a teacher with characteristics $(W, R) = (w, r)$ is

assigned to a classroom with characteristics $(X,S)=(x,s)$ instead of one with $(X,S)=(x',s')$.

Let $\rho_w=\Pr(W_i=w)$ and $\lambda_x=\Pr\left(X^h=x\right)$. My main identification result is

Proposition 1. *(Identification) Under (1) and Assumptions 1, 2 and 3 $\beta(w,x)$ is identified by*

$$\beta(w,x)=\frac{1}{\rho_w\lambda_x}\int\int q(w,x,r,s)\,p_w(r)\,p_x(s)\,f_R(r)\,f_S(s)\,\mathrm{d}r\mathrm{d}s. \tag{14}$$

Proof. Observe that under Assumption 2 we have

$$\begin{aligned} q(w,x,r,s) &= \int\int g(w,x,u,v)\,f_{U,V|W,X,R,S}(u,v|\,w,x,r,s)\,\mathrm{d}u\mathrm{d}v \\ &= \int\int g(w,x,u,v)\,f_{U|W,R}(u|\,w,r)\,f_{V|X,S}(v|\,x,s)\,\mathrm{d}u\mathrm{d}v. \end{aligned}$$

From Bayes' rule $f(r|\,w)=\frac{p_w(r)f(r)}{\rho_w}$ and $f(s|\,x)=\frac{p_x(s)f(s)}{\lambda_x}$. This and the second equality above yields

$$\begin{aligned} &\frac{1}{\rho_w\lambda_x}\int\int q(w,x,r,s)\,p_w(r)\,p_x(s)\,f_R(r)\,f_S(s)\,\mathrm{d}r\mathrm{d}s \\ =\ &\int\int q(w,x,r,s)\,f_{R|W}(r|\,w)\,f_{S|X}(s|\,x)\,\mathrm{d}r\mathrm{d}s \\ =\ &\int\int\left[\int\int g(w,x,u,v)\,f_{U|W,R}(u|\,w,r)\,f_{V|X,S}(v|\,x,s)\,\mathrm{d}u\mathrm{d}v\right] \\ &\times f_{R|W}(r|\,w)\,f_{S|X}(s|\,x)\,\mathrm{d}r\mathrm{d}s \\ =\ &\int\int\int\int g(w,x,u,v)\,f_{U,R|W}(u,r|\,w)\,f_{V,S|X}(v,s|\,x)\,\mathrm{d}u\mathrm{d}r\mathrm{d}v\mathrm{d}s \\ =\ &\int\int g(w,x,u,v)\,f_{U|W}(u|\,w)\,f_{V|X}(v|\,x)\,\mathrm{d}u\mathrm{d}v \\ =\ &\beta(w,x). \end{aligned}$$

Note that for the object prior to the first equality above to be well-defined we require Assumption 3. Since all the components to the right of the equality in (14) are asymptotically revealed under random sampling (Assumption 1), the result follows. □

Corollary 1. *The average no sorting outcome, $\gamma(x)$, is identified by*

$$\gamma(x)=\frac{1}{\lambda_x}\sum_{j=1}^{J}\int\int q(w_j,x,r,s)\,p_j(r)\,p_x(s)\,f_R(r)\,f_S(s)\,\mathrm{d}r\mathrm{d}s.$$

3 Semiparametric efficiency bound

Theorems 1 and 2 characterize the semiparametric efficiency bounds for, respectively, $\beta(w,x)$ and $\gamma(w)$, under (1) and Assumptions 1, 2 and 3.

Variance bound for $\beta(w,x)$

Let $D_w(W) = D_w = 1$ if $W = w$ and zero otherwise. Let $E_x(X) = E_x = 1$ if $X = x$ and zero otherwise. Let $T_{wx}(W,X) = T_{wx} = 1$ if $W = w$ and $X = x$ and zero otherwise. Define the candidate efficient influence function

$$\phi_0(Z,\beta(w,x)) = \psi_0(Z,\beta(w,x)) + \psi_R(Z,\beta(w,x)) + \psi_S(Z,\beta(w,x)) \tag{15}$$

where

$$\begin{aligned}
\psi_0(Z,\beta(w,x)) &= \frac{f(R|W)f(S|X)}{f(R,S)}\frac{T_{wx}}{p_{wx}(R,S)}(Y - q(w,x,R,S)) \\
\psi_R(Z,\beta(w,x)) &= \frac{D_w}{\rho_w}(e_S(w,x,R) - \beta(w,x)) \\
\psi_S(Z,\beta(w,x)) &= \frac{E_x}{\lambda_x}(e_R(w,x,S) - \beta(w,x))
\end{aligned}$$

with

$$\begin{aligned}
e_S(w,x,r) &= \int q(w,x,r,s)f(s|x)\,\mathrm{d}s \\
e_R(w,x,s) &= \int q(w,x,r,s)f(r|w)\,\mathrm{d}r.
\end{aligned}$$

Define the candidate variance bound

$$\begin{aligned}
\mathcal{I}_0(\beta(w,x))^{-1} &= \mathbb{E}\left[\left\{\frac{f(R|W=w)f(S|X=x)}{f(R,S)}\right\}^2\frac{\sigma^2_{wx}(R,S)}{p_{wx}(R,S)}\right] \\
&+\frac{1}{\rho_w}\mathbb{E}\left[(e_S(w,x,R) - \beta(w,x))^2\middle|W=w\right] \\
&+\frac{1}{\lambda_x}\mathbb{E}\left[(e_R(w,x,S) - \beta(w,x))^2\middle|X=x\right] \\
&+2\frac{\pi_{wx}}{\rho_w\lambda_x}\mathbb{E}\left[(e_S(w,x,R) - \beta(w,x))(e_R(w,x,S) - \beta(w,x))|W=w,X=x\right]
\end{aligned} \tag{16}$$

with

$$\begin{aligned}\sigma_{wx}^2(r,s) &= \mathbb{V}(Y|\,W=w,X=x,R=r,S=s)\\ \pi_{wx} &= \Pr(W=w,X=x).\end{aligned}$$

Theorem 1. *The semiparametric efficiency bound for $\beta(w,x)$ in the problem defined by (1) and Assumptions 1, 2 and 3 is equal to $\mathcal{I}_0(\beta(w,x))$ with an efficient influence function of $\phi_0(Z,\beta(w,x))$.*

Proof. See Appendix A. □

Both the efficient influence function and the variance bound have straightforward interpretations. Consider first the influence function. Its first term, $\psi_0(Z,\beta(w,x))$, reflects the asymptotic penalty associated not knowing conditional distribution of Y given (W,X,R,S). The second and third terms, $\psi_R(Z,\beta(w,x))$ and $\psi_S(Z,\beta(w,x))$, reflect the contributions of uncertainty about, respectively, the conditional distributions of R given W and S given X. The interpretation of $\mathcal{I}_0(\beta(w,x))^{-1}$ is analogous, with its last term arising from covariance between $\psi_R(Z,\beta(w,x))$ and $\psi_S(Z,\beta(w,x))$.

Variance bound for $\gamma(w)$

Define the candidate efficient influence function

$$\phi_0(Z,\gamma(w)) = \psi_0(Z,\gamma(w))+\psi_X(Z,\gamma(w))+\psi_R(Z,\gamma(w))+\psi_S(Z,\gamma(w))$$

where

$$\begin{aligned}\psi_0(Z,\gamma(w)) &= \frac{f(R|\,W)}{f(R|\,X,S)}\frac{D_w}{p_w(X,R,S)}(Y-m(X,R,S))\\ \psi_X(Z,\gamma(w)) &= \beta(w,X)-\gamma(w)\\ \psi_R(Z,\gamma(w)) &= \frac{D_w}{\rho_w}\{\overline{e}_{XS}(w,R)-\gamma(w)\}\\ \psi_S(Z,\gamma(w)) &= e_R(w,R,S)-\beta(w,X)\end{aligned}$$

with

$$
\begin{aligned}
p_w(X,R,S) &= \Pr(W=w|\,X=x,R=r,S=s)\\
m(x,r,s) &= \mathbb{E}[Y|\,X=x,R=r,S=s]\\
\overline{e}_{XS}(w,r) &= \mathbb{E}[e_S(w,X,r)] = \mathbb{E}[q(w,X,r,S)]\\
\overline{e}_{WR}(x,s) &= \mathbb{E}[e_R(W,x,s)] = \mathbb{E}[q(W,x,R,s)].
\end{aligned}
$$

Define the candidate variance bound

$$
\begin{aligned}
\mathcal{I}_0(\gamma(w))^{-1} &= E\left[\left\{\frac{f(R|\,W=w)}{f(R|\,X,S)}\right\}\frac{\sigma^2(X,R,S)}{p_w(X,R,S)}\right]\\
&\quad+\mathbb{E}\left[(\beta(w,X)-\gamma(w))^2\right]\\
&\quad+\frac{1}{\rho_w}\mathbb{E}\left[(\overline{e}_{XS}(w,R)-\gamma(w))^2\big|\,W=w\right]\\
&\quad+\mathbb{E}\left[(e_R(w,R,S)-\beta(w,X))^2\right]\\
&\quad+2\mathbb{E}\left[(\overline{e}_{XS}(w,R)-\gamma(w))(\beta(w,X)-\gamma(w))|\,W=w\right]\\
&\quad+2\mathbb{E}\left[(e_R(w,R,S)-\beta(w,X))(\overline{e}_{XS}(w,R)-\gamma(w))|\,W=w\right]
\end{aligned}
$$

where

$$
\sigma^2(X,R,S) = \mathbb{V}(Y|\,X=x,R=r,S=s).
$$

Theorem 2. *The semiparametric efficiency bound for $\gamma(w)$ in the problem defined by (1) and Assumptions 1, 2 and 3 is equal to $\mathcal{I}_0(\gamma(w))$ with an efficient influence function of $\phi_0(Z,\gamma(w))$.*

Proof. See Appendix B. □

Here the first term in $\phi_0(Z,\gamma(w))$ is due to uncertainty in Y given (W,X,R,S). The second from uncertainty in distribution of X and the third and fourth from uncertainty in, respectively, the distributions of R given W and S given X.

Connections to the program evaluation literature

Semiparametric efficiency for average treatment effect estimate under exogenous treatment assignment was studied by Hahn (1998), Hirano, Imbens and Ridder (2003), Chen, Hong and Tarozzi (2009), Graham (2011b) and others. One peculiar feature of this problem is that knowledge of the propensity score – the conditional distribution of treatment assignment

given covariates – lowers the variance bound for the average treatment effect on the treated (ATT), but not for the average treatment effect (ATE). In the present setting knowledge of the conditional probabilities of teacher type (W) given covariates (R), and classroom type (X) given covariates (S), should lower the variance bound for $\beta(w,x)$. This is because $\beta(w,x)$ is an average of $q(w,x,r,s)$ with respect to the density $f(r|w)f(r|x)$. When $p_w(r)$ and $p_w(x)$ are unknown the efficient estimate of the distribution of R given $W=w$ is its empirical distribution in the $W_i=w$ subsample. Similarly the efficient estimate of the distribution of S given $X=x$ is its empirical distribution in the $X^h=x$ subsample. When the propensity score is known, these distribution function estimates are no longer efficient. For example, the efficient estimate of the distribution of R given $W=w$ places weight

$$\omega_i = \frac{p_w(R_i)}{\sum_{j=1}^N p_w(R_j)}$$

on each observation (cf., Graham, Pinto and Egel, 2013). Since the form of the variance bound indicates that knowledge of $F_{R|W}$ and $F_{S|X}$ is valuable, knowledge of the "propensity scores" $p_w(R)$ and $p_x(S)$ will lower the variance bound.

Efficient estimation

Beyond the support condition, (1) and Assumptions 1, 2 and 3 do not restrict the joint distribution of the observed data $Z=(W,X,R',S',Y)$. In the terminology of Newey (1990, 1994b), $\beta(w,x)$ is an unrestricted parameter. Therefore an analog estimator based on Proposition 1 will have an asymptotic variance which coincides with the efficiency bound under suitable regularity conditions.

The form of this analog estimator is as follows. Let $\hat{\rho}_w = N^{-1}\sum_{i=1}^N \mathbf{1}(W_i=w)$ and $\hat{\lambda}_x = N^{-1}\sum_{h=1}^N \mathbf{1}\left(X^h=x\right)$, and $\hat{q}(w,x,r,s)$, $\hat{p}_w(r)$ and $\hat{p}_x(s)$ be consistent nonparametric estimates of $q(w,x,r,s)$, $p_w(r)$ and $p_x(s)$, then estimate $\beta(w,x)$ by

$$\hat{\beta}_V(w,x) = \frac{1}{\hat{\rho}_w\hat{\lambda}_x}\frac{1}{N^2}\sum_{i=1}^N\sum_{h=1}^N \hat{q}\left(w,x,R_i,S^h\right)\hat{p}_w(R_i)\,\hat{p}_x\left(S^h\right). \tag{17}$$

Estimator (17) is a V-statistic with an estimated kernel. Distribution theory for related estimators was developed by Honoré and Powell (1994, 2005) and Aradillas-Lopez, Honoré and Powell (2007). This estimator is also related to the correlated matching rule estimator introduced by Graham, Imbens and Ridder (forthcoming; cf., Theorem 7.3). This estimator is used for the empirical application below.

There may be other efficient estimators. For example, $\beta(w,x)$ has the inverse probability

weighting (IPW) representation (assuming $f(R,S)$ is bounded away from zero on the support of (R,S)):

$$\begin{aligned}
\mathbb{E}\left[\frac{f(R|w)f(S|x)}{f(R,S)}\frac{T_{wx}Y}{p_{wx}(R,S)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{f(R|w)f(S|x)}{f(R,S)}\frac{T_{wx}Y}{p_{wx}(R,S)}\middle| W,X,R,S\right]\right] \\
&= \mathbb{E}\left[\frac{f(R|w)f(S|x)}{f(R,S)}\frac{T_{wx}q(W,X,R,S)}{p_{wx}(R,S)}\right] \\
&= \mathbb{E}\left[\frac{f(R|w)f(S|x)}{f(R,S)}\frac{T_{wx}q(w,x,R,S)}{p_{wx}(R,S)}\right] \\
&= \mathbb{E}\left[\frac{f(R|w)f(S|x)}{f(R,S)}\frac{q(w,x,R,S)}{p_{wx}(R,S)}\mathbb{E}\left[T_{wx}|R,S\right]\right] \\
&= \mathbb{E}\left[\frac{f(R|w_j)f(S|x_k)}{f(R,S)}q(w_j,x_k,R,S)\right] \\
&= \int\int q(w,x,r,s)f(r|w)f(s|x)\,\mathrm{d}r\mathrm{d}s. \\
&= \beta(w,x).
\end{aligned}$$

This suggests the analog estimator

$$\hat{\beta}_{IPW}(w,x) = \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{f}(R_i|w)\hat{f}(S_i|x)}{\hat{f}(R_i,S_i)}\frac{T_{wx,i}Y_i}{\hat{p}_{wx}(R_i,S_i)}.$$

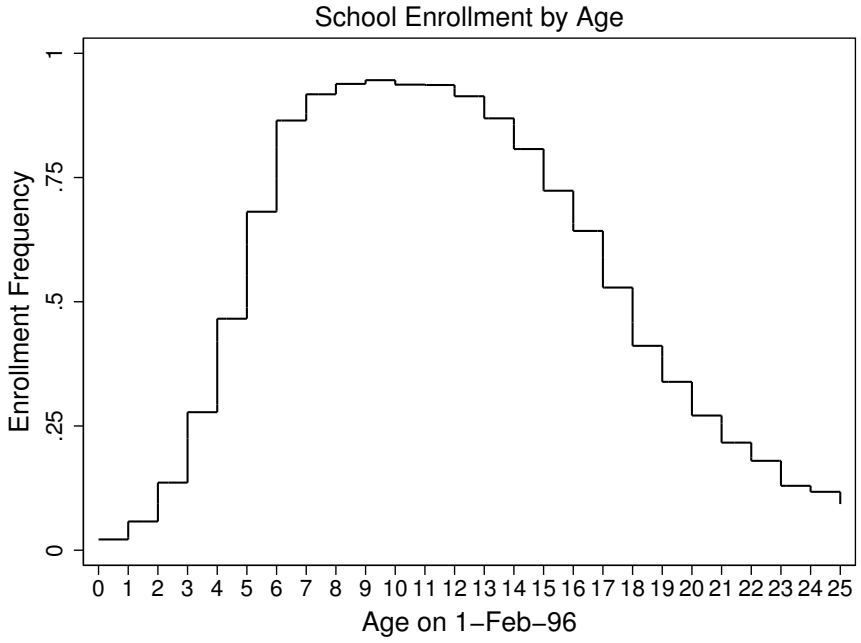
The presence of $\hat{f}(R_i,S_i)$ in the denominator of $\hat{\beta}_{IPW}(w,x)$ may require the imposition of additional regularity conditions relative to those required for $\hat{\beta}_V(w,x)$ (cf., Chen, Hong and Tarozzi, 2008).

[Elaborate more on the IPW representation; intuition etc.]

4 Marital sorting and inequality in Brazil

Brazil is well-known for its high level of inequality and low levels of measured intergenerational mobility (e.g., Lam, 1999; Behrman, Gaviria and Szekely, 2001). Inequality in the distribution of years of completed schooling "explains" a substantial portion of this inequality (e.g., Almeida dos Reis and Barros, 1991; Blom, Holm-Neilsen and Verner, 2001). Brazil is also known for its highly assortative distribution of marriages (e.g., Lam and Schoeni, 1993, 1994; Torche, 2010). The correlation in years of completed schooling across spouses exceeds

Figure 1: School attendance by age in Brazil



Source: PNAD 1996 and author's calculations. Age as of the start of the 1996 school year (February 1st). Figure computed using all 171,750 individuals between the ages of 0 and 25 in the 1996 PNAD (individual-level sampling weights used).

70 percent.[3] Here we explore the contribution of martial sorting by education, age and race to intergenerational mobility. Specifically we study the relationship between the distribution of parents' schooling and the distribution of educational attainment across co-resident adolescent children. Kremer (1997) undertakes a similar analysis using US data and parametric methods.

Following Lam (1999) our measure of child's educational attainment is the number of grades completed per year since age 7. For example, an individual who was age 13 in February 1st, 1996 would have completed a total of 6 grades of schooling, one per year, if she started school the February in which she was 7 and did not repeat any grades.[4] Table 2 lists the number of grades completed per year for the 27,547 youth in our estimation sample (see below). The Table also reports the percent of respondents who failed to complete any years of school by each age, as well as the percentage who have successfully completed all scheduled years

[3]Lam and Schoeni (1994) construct comparable samples of married couples from Brazil and the United States. The correlation in years of completed schooling across spouses is 0.77 in their Brazilian sample and 0.57 in the US sample. Their Brazil sample was from 1982, their US sample 1988. In our base sample below the correlation is 0.73.

[4]Currie and Yelowitz (2000) use a similar measured of educational attainment in the United States context.

of schooling by each age (i.e., who started school on time and have not repeated a grade). Grade repetition is common in Brazil, by age 15, when individuals should have completed all 8 years of primary school, average educational attainment is only 5.6 years (i.e., 0.70 grades per year). Less than 20 percent of 13 - 18 year olds are at the correct grade for their age, while 5 - 6 percent have failed to complete *any* grade by adolescence. Because our measure of educational attainment declines modestly with age, due to permanent school dropout during adolescence (see Figure 1), we use an age-adjusted version of it in the calculations which follow (see the notes to the Table 2 for description of the adjustment).

There is substantial variation in educational attainment across individuals by mid-adolescence. These differences are consequential for lifetime earnings: estimates of the returns to an additional year of schooling in Brazil are high, on the order of 10 to 15 percent (Lam and Schoeni, 1993, 1994) Since large differences in human capital acquisition develop during childhood, when parental influences are typically strong, it is of interest to study the role of parents own schooling in explaining them. If both parents' schooling does exert a strong influence on the educational attainment of their children, then assortative marriage by education may exacerbate inequality (cf., Kremer, 1997).

4.1 Construction of estimation sample

The Pesquisa Nacional por Amostra de Domicilios (PNAD) 1996 sampled 43,881 adolescents, defined as respondents between the ages of 13 and 18 on February 1st, 1996 (The first week of February is the start of the Brazilian school year). Of these 28,141 lived in a home where their birth mother was either the household head or the spouse of the household head. The balance of the 13 - 18 year olds either lived in homes without their birth mother, or, more often, in homes where their birth mother was neither head nor spouse of head. A total 27,547 of these respondents also lived with their mother's spouse, in most cases presumably their birth father (but also possibly a step-father).[5] Grades completed per year for this subsample are reported in Table 2 above.

[5]The household roster information available in the PNAD does not allow me to distinguish between fathers and step-fathers.

Table 2: Grades completed per year since age 7

Child age	13-18	13	14	15	16	17	18	13-18 (adjusted)
Grades per year	0.7020	0.7439	0.7312	0.7048	0.6915	0.6673	0.6380	0.7046
% 0 grades per year	5.6	5.9	5.3	5.7	5.5	5.7	5.8	-
% 1 grade per year	16.9	20.8	19.1	16.5	13.8	13.2	15.7	-
N	27,547	5,576	5,259	4,887	4,474	3,965	3,386	27,547

Source: PNAD 1996 and author's calculations. Averages weighted using the PNAD individual-level sampling weights. The construction of the sample is described in the main text. The right-hand column reports an age-adjusted measure of years of completed schooling per year. This measure is constructed by adding the residuals from a least squares fit of grades per year on a set of age dummies, to the average number of grades completed by 15 years olds in the sample. Individual-level sampling weights provided in the PNAD are used for these calculations.

Table 3: Age-adjusted grades completed by age 15 by years of parental schooling

A. Mother's years-of-schooling	All	0	1-3	4-7	8-10	11+
Grades per year	0.7622	0.4676	0.6340	0.8042	0.8887	0.9781
Total grades by age 15	6.10	3.74	5.07	6.43	7.11	7.83

B. Father's years-of-schooling	All	0	1-3	4-7	8-10	11+
Grades per year	0.7622	0.4768	0.6585	0.8106	0.8962	0.9771
Total grades by age 15	6.10	3.81	5.27	6.49	7.17	7.82

Source: PNAD 1996 and author's calculations. Sample consists the 11,220 families described in the main text. All averages use the household-level weights provided in the PNAD. Grades per year refers to the age adjusted measure described in the notes to Table 2 above.

We further focus on families where both parents are between the ages of 35 and 55. This age group includes most families with adolescent aged children. A total of 20,806 intact couples in this age group, with one member of the couple being the household head, are present in the PNAD. About 12,500 of these couples have adolescent aged children in their household. For 11,220 of these couples we have complete date on child's schooling, parents' schooling and grandparents' schooling. All results reported below are based on these 11,220 complete cases. In households with multiple adolescent siblings we use a household average of our "grades per year" measure as the outcome.

Table 3 reports grades completed per year between the ages of 7 and 15 by parental years of schooling. Consistent with prior research, parent education is strongly predictive of child's educational attainment (e.g., Lam and Duryea, 1999).

In Panel A of the table we see that while children of women who have themselves completed no schooling complete, on average, only 3.7 grades by age 15, children of women who have completed secondary school (11+ years of schooling) complete, on average, 7.8 years of schooling. Panel B reports the relationship with father's schooling and child's educational attainment. Father's education is equally predictive of child's schooling.

If both parents' educational background independently influences their child's educational attainment, then assortative matching by education in the marriage market will increase inequality in the next generation (e.g., Kremer, 1997; Torche, 2010). Depending on whether maternal and paternal education are complements or substitutes, such sorting may also raise or lower average education for the next generation relative to a hypothetical world where marriage is "random".

Table 4 reports the conditional frequency distribution of father's schooling given mother's schooling. Marriage is highly assortative in Brazil. If one's mother never attended school, there is almost a 60 percent chance that one's father also did not go to school. Likewise,

Table 4: Conditional distribution of father's school given mother's schooling

MOTHER/FATHER	0	1-3	4-7	8-10	11+	
0	59.0	19.8	18.0	2.2	1.1	100.0
1-3	25.8	39.3	28.6	4.3	2.0	100.0
4-7	10.2	17.5	52.2	12.1	8.0	100.0
8-10	3.4	8.3	30.2	32.0	26.1	100.0
11+	1.6	2.7	14.7	14.6	66.3	100.0
	17.9	17.7	32.4	12.2	19.8	100.0

Source: PNAD 1996 and author's calculations. Sample consists the 11,220 families described in the main text. Rows and columns respectively refer to ranges of years of completed schooling of mothers and fathers. Entries give the conditional frequency (multiplied by 100 to yield a percent) of various levels of father's education given mother's education. All calculations use the household-level weights provided in the PNAD.

Table 5: Conditional distribution of paternal given maternal grandparents schooling

Maternal Paternal (Grandmother/Grandfather)	NS/NS	NS/SS	SS/NS	SM/SM	16.6
NS/NS	66.1	11.8	8.5	13.7	100.0
NS/SS	37.4	22.8	7.9	31.9	100.0
SS/NS	39.3	12.8	16.9	31.1	100.0
SS/SS	16.6	12.0	8.2	63.1	100.0
	39.8	13.5	9.1	37.6	100.0

Source: PNAD 1996 and author's calculations. Sample consists the 11,220 families described in the main text. "NS" refers to no schooling or schooling unknown, "SS" refers to some schooling. Grandparents are ordered as grandmother first, grandfather second. Hence NS/SS means a grandmother with no schooling and a grandfather with some schooling. Entries give the conditional frequency (multiplied by 100 to yield a percent) of various levels of paternal grandparents' education given maternal grandparents' education. All calculations use the household-level weights provided in the PNAD.

there is only a 1 percent chance that one's father graduated from secondary school for this group (11+ years of school). In contrast, 66 percent of children of mothers who completed secondary school had fathers who also completed secondary school. Less than two percent of this group had fathers with no schooling. Assortative matching is strongest in the tails of the education distribution, with more mixing across intermediate levels of educational attainment.

The PNAD 1996 also includes information on grandparents schooling. Table 5 reports the conditional frequency distribution of paternal grandparents' schooling given maternal grandparents' schooling. Marriage is also highly assortative by parents' education in Brazil.

Finally Table 6 reports the conditional distribution of father's race given mother's race.

Table 6: Conditional distribution of father's race given mother's race

MOTHER/FATHER	White/Asian	Black/Mixed	
White/Asian	80.7	19.3	100.0
Black/Mixed	18.4	81.6	100.0
	44.1	56.0	100.0

Source: PNAD 1996 and author's calculations. Sample consists the 11,220 families described in the main text.

Marriage is also assortative by race in Brazil (albeit less so than in the United States).

4.2 Model specification

We categorize both mothers and fathers by 16 education levels (0 to 15+ years of schooling), two age levels (35 to 44 and 45 to 54) and two racial categories (White/Asian and Black/Mixed). This results in a total of $16 \times 2 \times 2 = 64$ "types" of women and men. The support points of W and X correspond to these types. Maternal and paternal grandparents' schooling serve as, respectively, R and S in our analysis. We categorize grandparents schooling using the four combinations presented in Table 5.

Assumption 2 requires that the unobserved determinants of marriage vary independently of the unobserved determinants of children's educational attainment *within* $W = w$, $R = r$ by $X = x$, $S = s$ cells. If individual sort on unobserved attributes that are themselves strongly correlated with child outcomes within $W = w$, $R = r$ by $X = x$, $S = s$ cells, and the distribution of these characteristics varies with $(W = w, R = r)$ and/or $(X = x, S = s)$, then Assumption 2 will fail. In our context, where potential marriage partners likely have substantially more information about one another than the econometrician, Assumption 2 is strong. On the other hand, as we will show below, parents' and grandparents' education are very strong predictors of child's educational attainment and hence our basic results may be reasonably robust to modest deviations from Assumption 2. We do not perform a formal sensitivity analysis in what follows, but developing such methods would be a useful area of future research.

Assumption 3 requires that $0 < \kappa \leq p_{wx}(r, s) \leq 1$ for all $(r, s) \in \mathbb{S}_{RS}(w, x)$ at all combinations of $W = w$ and $X = x$ of interest. In our sample all sixteen combinations of maternal and paternal grandparents' education are observed for each possible combination of parents' type, so that Assumption 3 is satisfied. If a finer partition of grandparents' education is used, however, overlap issues do arise. So in our case, satisfying Assumption 3 is partially an artifact of our definitions of R and S to include just four support points each.

Parametric analysis

Table 7 presents estimates of various regression functions mapping parental schooling into child schooling. Column one reports the coefficient on mother's schooling in a least squares fit of child's years of schooling per year on mother's years of completed schooling and demographics (see the notes to the Table for more details). Consistent with the summary statistics reported in Table 3, the coefficient on mother's schooling is large and precisely estimated. The estimated coefficient suggests that each year of mother's schooling is associated with an additional $8 \times 0.0335 \approx 0.27$ years of completed child schooling. Hence the schooling gap between the child of a women with no schooling versus a high school graduate is expected to be almost three years.

Column 2 adds father's years of completed schooling and demographics as additional covariates. In this specification both parental school variables enter significantly and are precisely estimated. Column 3 adds an interaction in parental schooling to the model (both parental schooling variables are deviated from their sample mean in forming the interaction). This variable enters with a negative sign and is precisely estimated. The column 3 estimate suggests that mother and fathers' schooling are substitutes. This, in turn, implies that assortative matching by education, in addition to generating inequality in educational attainment, may also be inefficient from the vantage point of educational attainment for the next generation. Columns 4 to 7 report the 1 to 3 specification with the additional inclusion of controls for grandparents' education. The inclusion of these controls modestly lowers the point estimates of the parental schooling coefficients, which remain precisely determined.

Nonparametric analysis

The joint support of (W, X, R, S) includes $64 \times 64 \times 4 \times 4 = 65,536$ points of support. The sample sizes available to us are not sufficiently large for a fully flexible specification of $q(w, x, r, s)$. Instead we proceed as follows. For each of the 16 subsamples defined in terms of unique combinations of realizations of R and S we compute a least squares fit of child's grades completed per year on a 4th order polynomial in parents' schooling (including all interactions) and dummies for parents' race and age.[6]

[6]Interactions in parents' race and age are also included. Separability of $q(w, x, r, s)$ in parents' education, age and race is maintained.

Table 7: Parametric estimates of $q(w, x, r, s)$

	(1)	(2)	(3)	(4)	(5)	(6)
MOTHERS'	0.0335	0.0194	0.0227	0.0306	0.0178	0.0214
YEARS-OF-SCHOOLING	(0.0006)	(0.0008)	(0.0009)	(0.0007)	(0.0009)	(0.0009)
FATHERS'		0.0181	0.0238		0.0168	0.0226
YEARS-OF-SCHOOLING		(0.0008)	(0.0008)		(0.0008)	(0.0009)
MOTHERS'×FATHERS'			−0.0032			−0.0030
YEARS-OF-SCHOOLING			(0.0001)			(0.0001)
Mothers' demographics	Yes	Yes	Yes	Yes	Yes	Yes
Fathers' demographics	No	Yes	Yes	No	Yes	Yes
Maternal grandparents schooling	No	No	No	Yes	Yes	Yes
Paternal grandparents schooling	No	No	No	No	Yes	Yes
R^2	0.3021	0.3463	0.3845	0.3111	0.3531	0.3877

Source: PNAD 1996 and author's calculations. Sample consists the 11,220 families described in the main text. Demographics include a dummy for race (White/Asian vs. Black/Mixed) and a quadratic function in age. The grandparents' schooling controls include dummies for each of the four possible schooling combinations used in Table 5. All controls are entered linearly with no interactions. Estimation is by least squares using the household-level sampling weights provided in the PNAD. Standard errors, in parentheses, are heteroscedastic robust.

The 4th order specification was chosen to balance parsimony with flexibility. Recall that the efficiency properties of any matching are strongly determined by the complementarity properties of the average match function (AMF). The 4th order specification implies that the "cross-derivative" of the AMF is a second order polynomial in parents' schooling (including an interaction). This specification does not, a priori, impose super- or sub-modularity on the match function. It is flexible enough to allow for parents' education to be complementary inputs at certain combinations of parental education and substitutable inputs at other combinations. In total our model for $q(w,x,r,s)$ includes $16 \times 21 = 336$ parameters. With this estimate of $q(w,x,r,s)$ in hand, the AMF is computed for various combinations of $W = w$ and $X = x$ using (17). Nonparametric cell mean estimators are used to compute ρ_w, λ_x, $p_w(r)$ and $p_x(s)$.

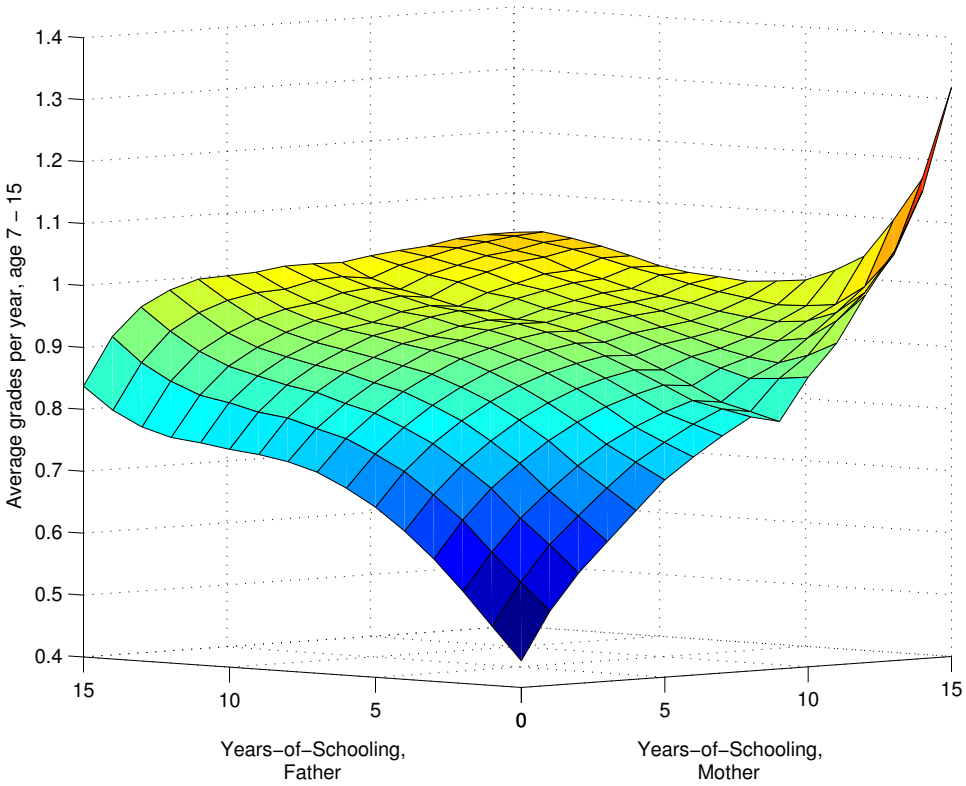
Figure 2 plots $\hat{\beta}(w,x)$ in mother and fathers' years of completed schooling space, holding both parents race fixed at White/Asian and their age fixed at 35 - 44. Consistent with the parametric analysis, child's education is increasing in both mother and father's years of completed schooling, but important nonlinearities are also apparent in the figure. First, if father's schooling is low, then the AMF increases relatively steeply with mother's schooling. Likewise if mother's schooling is low, the rise in the AMF with father's schooling is relatively steep. Increases in parental education are especially important, when one parent has very low levels of schooling. Second, for higher levels of mother's schooling, the effect of additional years of father's schooling is very modest to flat and possibly even negative at the highest levels of mother's schooling (although this portion of the AMF, the upper-right-hand region of the figure, is imprecisely estimated). Third, even at high levels of father's schooling, there are strong returns to increases in mother's schooling from very low, to modest levels.

Figure 3 shows the relationship between mother's years of completed schooling and child's educational attainment observed in our sample ("status quo"/ light gray line) and under two different counterfactual marriage distributions. The first, labelled "conditionally random" in the figure, replaces $f_{WXRS}(w,x,r,s)$ with $f_{WX}(w,x) f_{R|W}(r|w) f_{S|X}(s|x)$. This assignment maintains the observed degree of assortativeness of across parental characteristics, but eliminates any matching on R and S (grandparents' schooling). The dark line in the upper-left-hand graph corresponds to $\sum_{k=1}^{K} \beta(w,x_k) f_{X|W}(x_k|w)$.[7]

The difference between the dark (conditionally random) and light (status quo) assignments is plotted in the lower-left-hand figure. This difference can be viewed as the effect of "matching bias" on the observed distribution of child's educational attainment under our maintained assumptions. For example, part of the observed lower levels of education attainment amongst

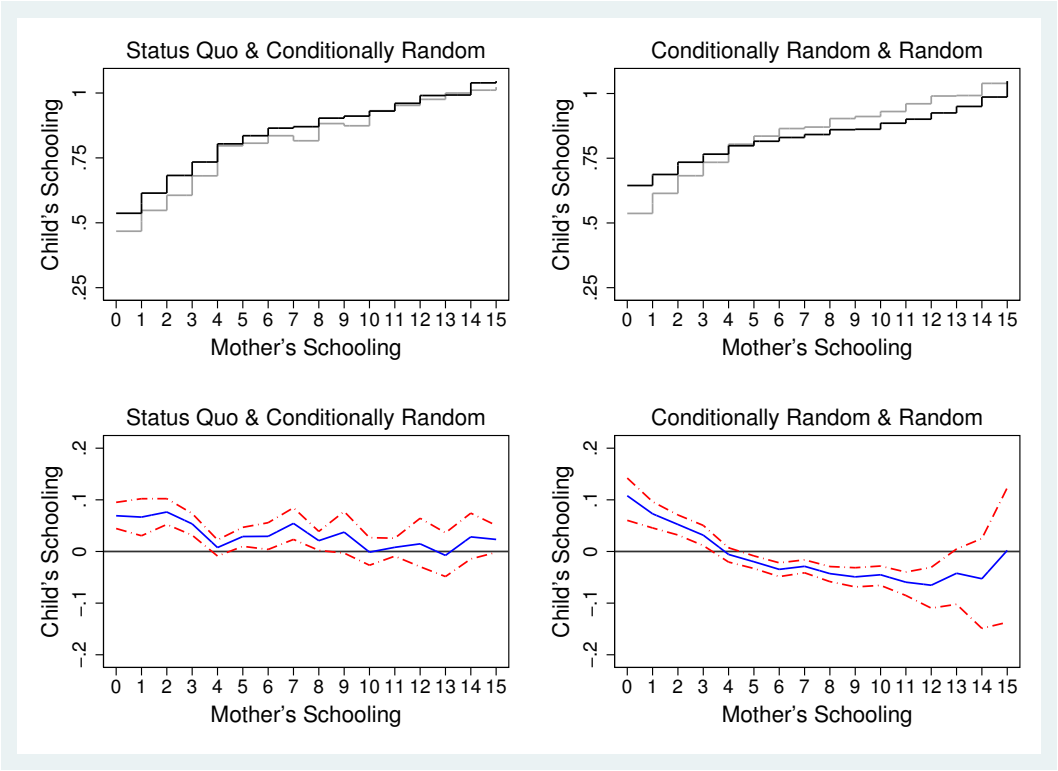
[7] In fact it is an average of this object across different combinations of couple race and age.

Figure 2: Average match function for White/Asian couples aged 35 to 44



Source: PNAD 1996 and author's calculations. Figure graphs $\hat{\beta}(w, x)$ in mother's and father's years-of-schooling space holding race and age fixed at, respectively, White/Asian and 35 to 44. Estimation procedure is as described in the main text.

Figure 3: Mother and child's schooling: status quo vs. counterfactuals



Source: PNAD 1996 and author's calculations.

children of poorly educated mothers is due to the fact that, not only do such women tend to marry men with low levels of education, but these men tend to also have unobserved characteristics which are themselves associated with lower levels of child schooling. Put differently, not only is a women with no education more likely to marry a man with no education, amongst men with no education she is more likely to marry one whose parents were also of low socioeconomic status. The upper-left-hand figure plots the counterfactual mapping from mother to child's schooling if this last source of "matching bias" were eliminated. Eliminating matching bias modestly raises education for children of mothers with low levels of schooling and leaves attainment for those of more educated mothers virtually unchanged.

In the second counterfactual marriage assignment, $f_{WXRS}(w,x,r,s)$ is replaced with

$$f_W(w) f_X(x) f_{R|W}(r|w) f_{S|X}(s|x).$$

The mapping from mother to child's schooling under this assignment is plotted in the upper-right-hand figure (dark line). Specifically the dark line corresponds to the ANSO, $\gamma(w) = \sum_{k=1}^{K} \beta(w, x_k) f_X(x_k)$. The mapping which only eliminates matching on R and S is also

plotted for comparison purposes in light gray. This figure gives an estimate of what a child's educational attainment would be for different levels of mother's education is marriage was completely at random. [Need to comment on relationship to "cross partial" of the AMF] The bottom-right-hand figure indicates that eliminating assortative matching on the marriage market would raise education attainment amongst children of poorly educated mothers and lower it among those of highly educated mothers. However, the beneficial effects on the former, are larger than the adverse effects on the latter. This is consist with Figure 2 which suggests that parents schooling are substitutes over much of the joint support of W and X (see also the negative interaction coefficient in Table 7).

Table 8 reports differences in grades completed per year across individuals with different levels of parental education. These differences are reported as observed (status quo) as well as under the counterfactual random matching. In Brazil the children of mother's who have completed the first cycle of primary school (4 years) accumulate on average 0.3289 more years of school per year between the ages of 7 and 15, than children of mother's with no education. A gap that accumulates into a difference of over 2.5 years of completed schooling by age 15. Under the counterfactual random matching this gap falls by over 50 percent. Eliminating assortative matching leads to more modest declines in the schooling gap across other levels of mother's education.

Table 9 reports estimates of the change in average child's education associated with moving from the status quo to random matching. Such a change would, under the maintained assumptions, raises average schooling by almost half a year (a little less than a 10 percent gain).

Table 8: Gaps in grades completed per year by mother's education and race

	4 vs. 0 years	8 vs. 4 years	11-14 vs. 8 years	15+ vs. 11-14 years	White	Black	Dif.
STATUS QUO	0.3289	0.0858	0.0700	0.0704	0.8370	0.6402	0.1967
	(0.0099)	(0.0096)	(0.0097)	(0.0103)	(0.0034)	(0.0051)	(0.0062)
RANDOM MATCHING	0.1536	0.0621	0.0405	0.1472	0.8421	0.7411	0.1010
	(0.0184)	(0.0085)	(0.0117)	(0.0663	(0.0087)	(0.0090)	(0.0095)

Source: PNAD 1996 and author's calculations.

Table 9: Aggregate education: status quo vs. random matching

	Grades per year	Grades by age 15
STATUS QUO	0.7622	6.0973
	(0.0033)	(0.0263)
RANDOM MATCHING	0.8162	6.5296
	(0.0085)	(0.0680)

Source: PNAD 1996 and author's calculations.

5 Conclusion

[To be completed]

References

[1] Almeida dos Reis, Jose Guilherme and Ricardo Paes de Barros. (1991). "Wage inequality and the distribution of education: A study of the evolution of regional differences in inequality in metropolitan Brazil," *Journal of Development Economics* 36 (1): 117 - 143

[2] Aradillas-Lopez, Andres, Bo E. Honoré and James L. Powell. (2007). "Pairwise difference estimation with nonparametric control variables," *International Economic Review* 48 (4): 1119 - 1158.

[3] Behrman, Jere R., Alejandro Gaviria and Miguel Szekely. (2001). "Intergenerational mobility in Latin America," *Economia* 2 (1): 1 - 44.

[4] Bickel, Peter J., Chris A.J. Klaassen, Ya'acov Ritov and Jon A. Wellner. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* New York: Springer-Verlag.

[5] Blundell, Richard and James L. Powell. (2003) "Endogeneity in nonparametric and semiparametric regression models," *Advances in Economics and Econometrics: Theory and Applications* II: 312 - 357 (M. Dewatripont, L.P. Hansen & S.J. Turnovsky, Eds.). Cambridge: Cambridge University Press.

[6] Chen, Xiaohong, Han Hong and Alessandro Tarozzi. (2008). "Semiparametric efficiency in GMM models with auxiliary data," *Annals of Statistics* 36 (2): 808 - 843.

[7] Choo, Eugene and Aloysius Siow. (2006a). "Who marries whom and why?" *Journal of Political Economy* 114 (1): 175 - 201.

[8] Choo, Eugene and Aloysius Siow. (2006b). "Estimating a marriage matching model with spillover effects," *Demography* 43 (3): 464 - 490.

[9] Chamberlain, Gary. (1984). "Panel data," *Handbook of Econometrics* 2: 1247 - 1318 (Z. Griliches & M. D. Intriligator, Eds.). Amsterdam: North-Holland.

[10] Currie, Janet and Aaron Yelowtiz. (2000). "Are public housing projects good for kids?" *Journal of Public Finance* 75 (1): 99 - 124.

[11] Blom, Andreas, Lauritz Holm-Nielsen and Dorte Verner. (2001). "Education, earnings, and Inequality in Brazil, 1982 - 1998: implications for education policy," *Peabody Journal of Education* 76 (3-4): 180-221.

[12] Boyd, Donald, Hamilton Lankford, Susanna Loeb and James Wyckoff. (2013). "Analyzing the determinants of the matching of public school teachers to jobs: disentangling the preferences of teachers and employers," *Journal of Labor Economics* 31 (1): 83 - 117.

[13] Fortin, Nicole, Thomas Lemieux and Sergio Firpo. (2011). "Decomposition methods in economics," *Handbook of Labor Economics* 4A: 1 - 102 (D. Card & O. Ashenfelter, Eds.). Amsterdam: North-Holland.

[14] Griliches, Zvi and Jacques Mairesse. (1998). "Production functions: the search for identification," *Econometrics and Economic Theory in the 20th Century: The Ragner Frisch Memorial Symposium*: 169 - 203 (S. Strom, Ed.). Cambridge: Cambridge University Press.

[15] Graham, Bryan S. (2008). "Identifying social interactions through conditional variance restrictions," *Econometrica* 76 (3): 643 - 660.

[16] Graham, Bryan S. (2011a). "Econometric methods for the analysis of assignment problems in the presence of complementarity and social spillovers," *Handbook of Social Economics* 1B: 965 - 1052 (J. Benhabib, A. Bisin, & M. Jackson, Eds.). Amsterdam: North-Holland.

[17] Graham, Bryan S. (2011b). "Efficiency bounds for missing data models with semiparametric restrictions," *Econometrica* 79 (2): 437 - 452.

[18] Graham, Bryan S., Guido W. Imbens and Geert Ridder. (2007). "Redistributive effects for discretely-valued inputs," *IEPR Working Paper No. 07.7.*

[19] Graham, Bryan S., Guido W. Imbens and Geert Ridder. (forthcoming). "Complementarity and aggregate implications of assortative matching: a nonparametric analysis," *Quantitative Economics*.

[20] Hahn, Jinyong. (1998). "On the role of the propensity score in efficient estimation of average treatment effects," *Econometrica* 66 (2): 315 - 311.

[21] Hahn, Jinyong, Keisuke Hirano and Dean Karlan. (2011). "Adaptive experimental design using the propensity score," *Journal of Business and Economic Statistics* 29 (1): 96 - 108.

[22] Heckman, James J. and Richard Robb (1985). "Alternative methods for evaluating the impact of interventions," *Longitudinal analysis of labor market data*: 156 - 246 (J.J. Heckman & B. Singer). Cambridge: Cambridge University Press.

[23] Heckman, James J. and Edward Vytlacil. (2007a). "Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation," *Handbook of Econometrics* 6B: 4779 – 4874 (J. J. Heckman & E. E. Leamer, Eds.). Amsterdam: North-Holland.

[24] Heckman, James J. and Edward Vytlacil. (2007b). "Econometric evaluation of social programs, part II: using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments *Handbook of Econometrics* 6B: 4875 - 5143 (J. J. Heckman & E. E. Leamer, Eds.). Amsterdam: North-Holland.

[25] Hirano, Keisuke and Jack R. Porter. (2009). "Asymptotics for statistical treatment rules," *Econometrica* 77 (5): 1683 - 1701.

[26] Ho, Kate. (2009). "Insurer-provider networks in the medical care market" *American Economic Review* 99 (1): 393 - 430.

[27] Holland, Paul W. (1986). "Statistics and causal inference," *Journal of the American Statistical Association* 81 (396): 945 - 960.

[28] Honoré, Bo E. and James L. Powell. (1994). "Pairwise difference estimators of censored and truncated regression models," *Journal of Econometrics* 64 (1-2): 241–278.

[29] Honoré, Bo E. and James L. Powell. (2005). "Pairwise difference estimators for nonlinear models," *Identification and Inference for Econometric Models: Essays in Honor*

of Thomas Rothenberg: 520 - 553 (D.W.K. Andrews & J.H. Stock, Eds.). Cambridge: Cambridge University Press.

[30] Imbens, Guido W. (2004). "Nonparametric estimation of average treatment effects under exogeneity," *Review of Economics and Statistics* 86 (1): 4 - 29.

[31] Imbens, Guido W. and Whitney K. Newey. (2009). "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica* 77 (5): 1481 - 1512.

[32] Kremer, Michael. (1997). "How much does sorting increase inequality," *Quarterly Journal of Economics* 112 (1): 115 - 139.

[33] Kremer, Michael and Eric Maskin. (1996). "Wage inequality and segregation by skill," *NBER Working Paper 5718.*

[34] Lam, David. (1999). "Generating extreme inequality: schooling, earnings and intergenerational transmission of human capital in South Africa and Brazil," *Population Studies Center Report No. 99-439.*

[35] Lam, David and Suzanne Duryea. (1999). "Effects of schooling on fertility, labor supply, and investments in children, with evidence from Brazil," *Journal of Human Resources* 34 (1): 160 - 192.

[36] Lam, David and Robert F. Schoeni. (1993). "Effects of family background on earnings and returns to schooling: evidence from Brazil," *Journal of Political Economy* 101 (4): 710 - 740.

[37] Lam, David and Robert F. Schoeni. (1994). "Family ties and labor markets in the United States and Brazil," *Journal of Human Resources* 29 (4): 1235 - 1258.

[38] Manski, Charles F. (1975). "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics* 3 (3): 205 - 228.

[39] Manski, Charles F. (1985). "Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator," *Journal of Econometrics* 27 (3): 313 - 333.

[40] Newey, Whitney K. (1990). "Semiparametric efficiency bounds," *Journal of Applied Econometrics* 5 (2): 99 - 135.

[41] Newey, Whitney K. (1994a). "Kernel estimation of partial means and a general variance estimator," *Econometric Theory* 10 (2): 233 - 253.

[42] Newey, Whitey K. (1994b). "The asymptotic variance of semiparametric estimators," *Econometrica* 62 (6): 1349 - 1382.

[43] Olley, G. Steven and Ariel Pakes. (1996). "The dynamics of productivity in the telecommunications equipment industry," *Econometrica* 64 (6): 1263 - 1297.

[44] Schwartz, Christine. (2010). "Earnings inequality and the changing association between spouse's earnings," *American Journal of Sociology* 115 (5): 1524 - 1557.

[45] Topkis, Donald M. (1998). *Supermodularity and Complementarity.* Princeton, NJ: Princeton University Press.

[46] Torche, Florencia. (2010). "Educational assortative matching and economic inequality: a comparative analysis of three Latin American countries," *Demography* 47 (2): 481 - 502.

[47] Wooldridge, Jeffrey. (2005). "Unobserved heterogeneity and the estimation of average partial effects," *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*: 27 - 55 (D.W.K. Andrews & J.H. Stock, Eds.). Cambridge: Cambridge University Press.

A Proof of Theorem 1

In calculating the semiparametric efficiency bound for the model defined by (1) and Assumptions 1 to 3 above I follow the general approach of Bickel, Klaassen, Ritov and Wellner (1993) and, especially, Newey (1990, Section 3). First, I characterize the nuisance tangent space. Second, I demonstrate pathwise differentiability of the average match function $\beta_{jk} = \beta(w_j, x_k)$. The efficient influence function is the projection of the pathwise derivative onto the nuisance tangent space. In the present case the pathwise derivative is an element of the tangent space and therefore coincides with the required projection (i.e., β_{jk} is a parameter of an unrestricted distribution and hence the pathwise derivative is unique; cf. Newey (1994b)). The main result then follows from Theorem 3.1 of Newey (1990, p. 106).

Step 1: Characterization of tangent space

The joint density function of $Z = (W, X, Y, R', S')'$, recalling that

$$p_{jk}(r, s) = \Pr(W = w_j, X = x_k | R = r, S = s),$$

$\rho_j = \Pr(W = w_j)$ and $\lambda_k = \Pr(X = x_k)$, is conveniently factorized as follows:

$$\begin{aligned} f(y,w,x,r,s) &= \prod_{j=1}^{J}\prod_{k=1}^{K} f(y|\, w_j,x_k,r,s)^{d_j e_k} f(r,s|\, w_j,x_k)^{d_j e_k} \Pr(W=w_j,X=x_k)^{d_j e_k} \\ &= \prod_{j=1}^{J}\prod_{k=1}^{K} f(y|\, w_j,x_k,r,s)^{d_j e_k} \left[\frac{f(w_j,x_k,r,s)}{f(w_j,r)\,f(x_k,s)} f(r|\, w_j)\, f(s|\, x_k)\,\rho_j\lambda_k\right]^{d_j e_k} \\ &= \prod_{j=1}^{J}\prod_{k=1}^{K} f(y|\, w_j,x_k,r,s)^{d_j e_k} \left[\frac{p_{jk}(r,s)}{p_j(r)\,p_k(s)}\frac{f(r,s)}{f(r)\,f(s)} f(r|\, w_j)\, f(s|\, x_k)\,\rho_j\lambda_k\right]^{d_j e_k}, \end{aligned}$$

where I suppress the functional dependence of d_j on w and e_k on x.[8] Recall also that $p_j(r) = \Pr(W = w_j|\, R = r)$ and $p_k(s) = \Pr(X = x_k|\, S = s)$.
Consider a regular parametric submodel with $f(y,w,x,r,s;\eta) = f(y,w,x,r,s)$ at $\eta = \eta_0$. The submodel joint density is given by

$$\begin{aligned} f(y,w,x,r,s;\eta) &= \prod_{j=1}^{J}\prod_{k=1}^{K} f(y|\, w_j,x_k,r,s;\eta)^{d_j e_k}. \\ &\quad \times \left[\frac{p_{jk}(r,s;\eta)}{p_j(r;\eta)\,p_k(s;\eta)}\frac{f(r,s;\eta)}{f(r;\eta)\,f(s;\eta)} f(r|\, w_j;\eta)\, f(s|\, x_k;\eta)\,\rho_j(\eta)\,\lambda_k(\eta)\right]^{d_j e_k} \end{aligned}$$

The submodel log likelihood is

$$\begin{aligned} \ln f(y,w,x,r,s;\eta) &= \sum_{j=1}^{J}\sum_{k=1}^{K} d_j e_k \ln f(y|\, w_j,x_k,r,s;\eta) \\ &\quad + \sum_{j=1}^{J}\sum_{k=1}^{K} d_j e_k \ln\left\{\frac{p_{jk}(r,s;\eta)}{p_j(r;\eta)\,p_k(s;\eta)}\frac{f(r,s;\eta)}{f(r;\eta)\,f(s;\eta)}\right\} \\ &\quad + \sum_{j=1}^{J}\sum_{k=1}^{K} d_j e_k \ln f(r|\, w_j;\eta) + \sum_{j=1}^{J}\sum_{k=1}^{K} d_j e_k \ln f(s|\, x_k;\eta) \\ &\quad + \sum_{j=1}^{J}\sum_{k=1}^{K} d_j e_k \ln\rho_j(\eta) + \sum_{j=1}^{J}\sum_{k=1}^{K} d_j e_k \ln\lambda_k(\eta) \end{aligned}$$

[8] That is $D_j = D_j(W) = 1$ if $W = w_j$ and zero otherwise and that $E_k = E_k(X) = 1$ if $X = x_k$ and zero otherwise

Using the fact that $\sum_{j=1}^{J}\sum_{k=1}^{K} d_j e_k \ln f(r|\, w_j;\eta) = \sum_{j=1}^{J} d_j \ln f(r|\, w_j;\eta)$, and a similar simplification holding for the last three terms in the above likelihood, we get

$$\begin{aligned}
\ln f(y,w,x,r,s;\eta) &= \sum_{j=1}^{J}\sum_{k=1}^{K} d_j e_k \ln f(y|\, w_j,x_k,r,s;\eta) \\
&\quad + \sum_{j=1}^{J}\sum_{k=1}^{K} d_j e_k \ln\left\{\frac{p_{jk}(r,s;\eta)}{p_j(r;\eta)\,p_k(s;\eta)}\frac{f(r,s;\eta)}{f(r;\eta)\,f(s;\eta)}\right\} \\
&\quad + \sum_{j=1}^{J} d_j \ln f(r|\, w_j;\eta) + \sum_{k=1}^{K} e_k \ln f(s|\, x_k;\eta) \\
&\quad + \sum_{j=1}^{J} d_j \ln \rho_j(\eta) + \sum_{k=1}^{K} e_k \ln \lambda_k(\eta),
\end{aligned}$$

with an associated score vector of

$$\begin{aligned}
s_\eta(y,w,x,r,s;\eta) &= \sum_{j=1}^{J}\sum_{k=1}^{K} d_j e_k s_\eta(y|\, w_j,x_k,r,s;\eta) \\
&\quad + \sum_{j=1}^{J}\sum_{k=1}^{K} d_j e_k k_\eta(w_j,x_k,r,s;\eta) \\
&\quad + \sum_{j=1}^{J} d_j t_\eta(r|\, w_j;\eta) + \sum_{k=1}^{K} e_k t_\eta(s|\, x_k;\eta) \\
&\quad + \sum_{j=1}^{J} d_j \rho_{j\eta}(\eta) + \sum_{k=1}^{K} e_k \lambda_{k\eta}(\eta),
\end{aligned} \tag{18}$$

where

$$\begin{aligned}
s_\eta(y|\, w_j,x_k,r,s;\eta) &= \nabla_\eta \ln f(y|\, w_j,x_k,r,s;\eta) \\
k_\eta(w_j,x_k,r,s;\eta) &= \nabla_\eta \ln p_{jk}(r,s;\eta) - \nabla_\eta \ln p_j(r;\eta) - \nabla_\eta \ln p_k(s;\eta) \\
&\quad + \nabla_\eta \ln f(r,s;\eta) - \nabla_\eta \ln f(r;\eta) - \nabla_\eta \ln f(s;\eta) \\
t_\eta(r|\, w_j;\eta) &= \nabla_\eta \ln f(r|\, w_j;\eta) \\
t_\eta(s|\, x_k;\eta) &= \nabla_\eta \ln f(s|\, x_k;\eta) \\
\rho_{j\eta}(\eta) &= \frac{\partial \ln \rho_j(\eta)}{\partial \eta} \\
\lambda_{k\eta}(\eta) &= \frac{\partial \ln \lambda_k(\eta)}{\partial \eta}.
\end{aligned}$$

By the usual conditional mean zero property of the score function,

$$\begin{aligned}
\mathbb{E}\left[s_\eta\left(Y\middle|W,X,R,S\right)\middle|W,X,R,S\right] &= 0 \\
\mathbb{E}\left[k_\eta\left(W,X,R,S\right)\right] &= 0 \\
\mathbb{E}\left[t_\eta\left(R\middle|W\right)\middle|W\right] &= 0 \\
\mathbb{E}\left[t_\eta\left(S\middle|X\right)\middle|X\right] &= 0, \qquad (19)
\end{aligned}$$

where the suppression of η in a function means that it is evaluated at is population value (e.g., $t_\eta\left(S\middle|X\right) = t_\eta\left(S\middle|X;\eta_0\right)$).

From (18) and (19) the tangent set is therefore given by

$$\begin{aligned}
\mathcal{T} = \Bigg\{ & \sum_{j=1}^{J}\sum_{k=1}^{K} d_j e_k s\left(y\middle|w_j,x_k,r,s\right) + \sum_{j=1}^{J}\sum_{k=1}^{K} d_j e_k k\left(w_j,x_k,r,s\right) \\
& + \sum_{j=1}^{J} d_j t\left(r\middle|w_j\right) + \sum_{k=1}^{K} e_k t\left(s\middle|x_k\right) + \sum_{j=1}^{J} d_j a_j + \sum_{k=1}^{K} e_k b_k \Bigg\}, \qquad (20)
\end{aligned}$$

where a_j and b_k are finite constants for $j = 1,\ldots,J$ and $k = 1,\ldots,K$ and $s\left(y\middle|w,x,r,s\right)$, $k\left(w,x,r,s\right)$, $t\left(r\middle|w\right)$ and $t\left(s\middle|x\right)$ satisfy

$$\begin{aligned}
\mathbb{E}\left[s\left(Y\middle|W,X,R,S\right)\middle|W,X,R,S\right] &= 0 \\
\mathbb{E}\left[k\left(W,X,R,S\right)\right] &= 0 \\
\mathbb{E}\left[t\left(R\middle|W\right)\middle|W\right] &= 0 \\
\mathbb{E}\left[t\left(S\middle|X\right)\middle|X\right] &= 0.
\end{aligned}$$

Step 2: Demonstration of Pathwise Differentiability

Under the parametric submodel $\beta\left(\eta\right)$ is identified by

$$\beta\left(w,x;\eta\right) = \int\int\left[\int y f\left(y\middle|w,x,r,s;\eta\right)\mathrm{d}y\right] f\left(r\middle|w;\eta\right) f\left(s\middle|x;\eta\right)\mathrm{d}r\mathrm{d}s.$$

Differentiating under the integral and evaluating at $\eta = \eta_0$ gives

$$\begin{aligned}
\frac{\partial \beta(w,x;\eta_0)}{\partial \eta'} &= \int\int \mathbb{E}\left[Y s_\eta(Y|W,X,R,S;\eta_0)|\, w,x,r,s\right] f(r|\, w;\eta_0)\, f(s|\, x;\eta_0)\,\mathrm{d}y\mathrm{d}r\mathrm{d}s \\
&\quad + \int\int q(w,x,r,s;\eta_0)\,\frac{\partial \log f(r|\, w;\eta_0)}{\partial \eta'} f(r|\, w;\eta_0)\, f(s|\, x;\eta_0)\,\mathrm{d}r\mathrm{d}s \\
&\quad + \int\int q(w,x,r,s;\eta_0)\, f(r|\, w;\eta_0)\,\frac{\partial \log f(s|\, x;\eta_0)}{\partial \eta'} f(s|\, x;\eta_0)\,\mathrm{d}r\mathrm{d}s \\
&= \int\int \mathbb{E}\left[Y s_\eta(Y|W,X,R,S;\eta_0)|\, w,x,r,s\right] f(r|\, w;\eta_0)\, f(s|\, x;\eta_0)\,\mathrm{d}y\mathrm{d}r\mathrm{d}s \\
&\quad + \int e_S(w,x,r;\eta_0)\,\frac{\partial \log f(r|\, w;\eta_0)}{\partial \eta'} f(r|\, w;\eta_0)\,\mathrm{d}r \\
&\quad + \int e_R(w,x,s;\eta_0)\,\frac{\partial \log f(s|\, x;\eta_0)}{\partial \eta'} f(s|\, x;\eta_0)\,\mathrm{d}s \\
&= \int\int \mathbb{E}\left[Y s_\eta(Y|W,X,R,S;\eta_0)|\, w,x,r,s\right] f(r|\, w;\eta_0)\, f(s|\, x;\eta_0)\,\mathrm{d}y\mathrm{d}r\mathrm{d}s \\
&\quad + \mathbb{E}\left[e_S(w,x,R;\eta_0)\,\frac{\partial \log f(R|\, w;\eta_0)}{\partial \eta'}\,\middle|\, W = w\right] \\
&\quad + \mathbb{E}\left[e_R(w,x,S;\eta_0)\,\frac{\partial \log f(S|\, x;\eta_0)}{\partial \eta'}\,\middle|\, X = x\right], \qquad (21)
\end{aligned}$$

where

$$\begin{aligned}
e_S(w,x,r;\eta_0) &= \int q(w,x,r,s;\eta_0)\, f(s|\, x;\eta_0)\,\mathrm{d}s \\
e_R(w,x,s;\eta_0) &= \int q(w,x,r,s;\eta_0)\, f(r|\, w;\eta_0)\,\mathrm{d}r.
\end{aligned}$$

To demonstrate pathwise differentiability of $\beta_{jk} = \beta(w_j, x_k)$ we require $F_{jk}(Y,W,X,R,S)$ such that

$$\frac{\partial \beta(w_j,x_k;\eta_0)}{\partial \eta'} = \mathbb{E}\left[F_{jk}(Y,W,X,R,S)\, s_\eta(Y,W,X,R,S;\eta_0)'\right]. \qquad (22)$$

With some work it is possible to show that condition (22) holds for

$$\begin{aligned}
F_{jk}(Y,W,X,R,S) &= \frac{f(R|\, W)\, f(S|\, X)}{f(R,S)}\,\frac{D_j E_k}{p_{jk}(R,S)}\left(Y - q(w_j,x_k,R,S)\right) \\
&\quad + \frac{D_j}{\rho_j}\left(e_S(w_j,x_k,R) - \beta_{jk}\right) + \frac{E_k}{\lambda_k}\left(e_R(w_j,x_k,S) - \beta_{jk}\right). \qquad (23)
\end{aligned}$$

I evaluate the covariance of each of the three terms in (23) with $s_\eta(Y,W,X,R,S;\eta)$ in turn.

I begin with

$$\begin{aligned}
&\mathbb{E}\left[\frac{f\left(R\middle|W\right)f\left(S\middle|X\right)}{f\left(R,S\right)}\frac{D_jE_k}{p_{jk}\left(R,S\right)}\right. \\
&\qquad\left.\times\left(Y-q\left(w_1,x_1,R,S\right)\right)\right. \\
&\quad\left.\times D_jE_ks_\eta\left(Y,W,X,R,S;\eta_0\right)\right] &=& \mathbb{E}\left[\frac{f\left(R\middle|W\right)f\left(S\middle|X\right)}{f\left(R,S\right)}\frac{D_jE_k}{p_{jk}\left(R,S\right)}\right. \\
&&&\left.\times\left(Y-q\left(w_j,x_k,R,S\right)\right)D_jE_ks_\eta\left(Y\middle|W,X,R,S;\eta_0\right)\right] \\
&&&+\mathbb{E}\left[\frac{f\left(R\middle|W\right)f\left(S\middle|X\right)}{f\left(R,S\right)}\frac{D_jE_k}{p_{jk}\left(R,S\right)}\right. \\
&&&\left.\times\left(Y-q\left(w_j,x_k,R,S\right)\right)D_jE_ks_\eta\left(W,X,R,S;\eta_0\right)\right] \\
&&=& \mathbb{E}\left[\frac{f\left(R\middle|W\right)f\left(S\middle|X\right)}{f\left(R,S\right)}\frac{D_jE_k}{p_{jk}\left(R,S\right)}Ys_\eta\left(Y\middle|W,X,R,S;\eta_0\right)\right] \\
&&=& \mathbb{E}\left[\frac{f\left(R\middle|W\right)f\left(S\middle|X\right)}{f\left(R,S\right)}\frac{D_jE_k}{p_{jk}\left(R,S\right)}\right. \\
&&&\left.\times\mathbb{E}\left[Ys_\eta\left(Y\middle|W,X,R,S;\eta_0\right)\middle|W,X,R,S\right]\right] \\
&&=& \int\int\sum_{l=1}^{J}\sum_{m=1}^{K}\frac{f\left(r\middle|w_l\right)f\left(s\middle|x_m\right)}{f\left(r,s\right)}\frac{D_j\left(w_l\right)E_k\left(x_m\right)}{p_{jk}\left(r,s\right)} \\
&&&\times\mathbb{E}\left[Ys_\eta\left(Y\middle|W,X,R,S;\eta_0\right)\middle|w_l,x_m,,r,s\right]p_{lm}\left(r,s\right)\right]f\left(r,s\right)\mathrm{d}r\mathrm{d}s \\
&&=& \int\int\mathbb{E}\left[Ys_\eta\left(Y\middle|W,X,R,S;\eta_0\right)\middle|W=w_j,X=x_k,r,s\right] \\
&&&\times f\left(r\middle|w_j\right)f\left(s\middle|x_k\right)\mathrm{d}r\mathrm{d}s,
\end{aligned}$$

which coincides with the first component of (21). The second equality above follows by iterated expectations and the conditional mean zero property of the score function. The third and fourth equalities follow from applications of iterated expectations.

To evaluate the covariance of the second two terms in (23) with $s_\eta\left(Y,W,X,R,S;\eta_0\right)$ the following alternative density factorizations will prove useful:

$$\begin{aligned}
f\left(w,x,r,s;\eta\right) &= f\left(r\middle|w;\eta\right)f\left(x,s\middle|w,r;\eta\right)f\left(w;\eta\right) \\
f\left(w,x,r,s;\eta\right) &= f\left(s\middle|x;\eta\right)f\left(w,r\middle|x,s;\eta\right)f\left(x;\eta\right).
\end{aligned}$$

These give, in an abuse of notation, the score decompositions

$$\begin{aligned}
s_\eta\left(Y,W,X,R,S;\eta\right) &= s_\eta\left(Y\middle|W,X,R,S;\eta\right)+t_\eta\left(R\middle|W;\eta\right)+s_\eta\left(X,S\middle|W,R;\eta\right)+s_\eta\left(W;\eta\right) \\
s_\eta\left(Y,W,X,R,S;\eta\right) &= s_\eta\left(Y\middle|W,X,R,S;\eta\right)+t_\eta\left(S\middle|X;\eta\right)+s_\eta\left(W,R\middle|X,S;\eta\right)+s_\eta\left(X;\eta\right).
\end{aligned}$$

By the conditional mean zero property of the score function

$$\mathbb{E}\left[\frac{D_j}{\rho_j}\left(e_S\left(w_j, x_k, R\right) - \beta_{jk}\right) s_\eta\left(Y \middle| W, X, R, S\right)\right] = 0.$$

Using iterated expectations further yields

$$\begin{aligned}
\mathbb{E}\left[\frac{D_j}{\rho_j}\left(e_S\left(w_j, x_k, R\right) - \beta_{jk}\right) s_\eta\left(W\right)\right] &= \mathbb{E}\left[s_\eta\left(W\right)\mathbb{E}\left[\frac{D_j}{\rho_j}\left(e_S\left(w_j, x_k, R\right) - \beta_{jk}\right)\middle| W\right]\right] \\
&= s_\eta\left(w_j\right)\mathbb{E}\left[\left(e_S\left(w_j, x_k, R\right) - \beta_{jk}\right)\middle| W = w_j\right] \\
&= s_\eta\left(w_j\right)\left(\beta_{jk} - \beta_{jk}\right) \\
&= 0.
\end{aligned}$$

Similarly, using iterated expectations and the conditional mean zero property of the score function, yields

$$\begin{aligned}
\mathbb{E}\left[\frac{D_j}{\rho_j}\left(e_S\left(w_j, x_k, R\right) - \beta_{jk}\right) s_\eta\left(X, S \middle| W, R\right)\right] &= \mathbb{E}\left[\frac{D_j}{\rho_j}\left(e_S\left(w_j, x_k, R\right) - \beta_{jk}\right)\mathbb{E}\left[s_\eta\left(X, S\middle| W, R\right)\middle| W, R\right]\right] \\
&= \mathbb{E}\left[\frac{D_j}{\rho_j}\left(e_S\left(w_j, x_k, R\right) - \beta_{jk}\right)\cdot 0\right] \\
&= 0.
\end{aligned}$$

Finally

$$\begin{aligned}
\mathbb{E}\left[\frac{D_j}{\rho_j}\left(e_S\left(w_j, x_k, R\right) - \beta_{jk}\right) t_\eta\left(R\middle| W\right)\right] &= \mathbb{E}\left[\frac{D_j}{\rho_j}\left(e_S\left(w_j, x_k, R\right) - \beta_{jk}\right)\frac{\partial \log f\left(R\middle| W; \eta_0\right)}{\partial \eta'}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{D_j}{\rho_j}\left(e_S\left(w_j, x_k, R\right) - \beta_{jk}\right)\frac{\partial \log f\left(R\middle| W; \eta_0\right)}{\partial \eta'}\middle| W\right]\right] \\
&= \mathbb{E}\left[\left(e_S\left(w_j, x_k, R\right) - \beta_{jk}\right)\frac{\partial \log f\left(R\middle| W; \eta_0\right)}{\partial \eta'}\middle| W = w_j\right] \\
&= \mathbb{E}\left[e_S\left(w_j, x_k, R\right)\frac{\partial \log f\left(R\middle| W; \eta_0\right)}{\partial \eta'}\middle| W = w_j\right],
\end{aligned}$$

again using the conditional mean zero property of the score function in moving the second-to-last to last equality. Putting these results together gives

$$\mathbb{E}\left[\frac{D_j}{\rho_j}\left(e_S\left(w_j, x_k, R\right) - \beta_{jk}\right) s_\eta\left(Y, W, X, R, S; \eta_0\right)\right] = \mathbb{E}\left[e_S\left(w_j, x_k, R\right)\frac{\partial \log f\left(R\middle| w_j; \eta_0\right)}{\partial \eta'}\middle| W = w_j\right].$$

Analogous calculations yield the expression

$$\mathbb{E}\left[\frac{E_k}{\lambda_k}\left(e_R\left(w_j,x_k,S\right)-\beta_{jk}\right)s_\eta\left(Y,W,X,R,S;\eta_0\right)\right]=\mathbb{E}\left[e_R\left(w_j,x_k,S;\eta_0\right)\frac{\partial\log f\left(S|\,x_k;\eta_0\right)}{\partial\eta'}\middle|\,X=x_k\right].$$

These expressions coincide with the second and third components of (21). Condition (22) then holds for $F_{jk}\left(Y,W,X,R,S\right)$ as defined in (23).

Step 3: Calculation of projection

The semiparametric variance bound for β_{jk} is the expected square of the projection of $F_{jk}\left(Y,W,X,R,S\right)$ onto $\mathcal{T}$. Since $F_{jk}\left(Y,WX,R,S\right)\in\mathcal{T}$ it coincides with the required projection and is therefore the efficient influence function as claimed. Here

$$\frac{f\left(R|\,w_j\right)f\left(S|\,x_k\right)}{f\left(R,S\right)}\frac{D_jE_k}{p_{jk}\left(R,S\right)}\left(Y-q\left(w_j,x_k,R,S\right)\right)$$

plays the role of $\sum_{j=1}^{J}\sum_{k=1}^{K}d_je_ks\left(y|\,w_j,x_k,r,s\right)$ and $\frac{D_j}{\rho_j}\left(e_S\left(w_j,x_k,R\right)-\beta_{jk}\right)$ and $\frac{E_k}{\lambda_k}\left(e_R\left(w_j,x_k,S\right)-\beta_{jk}\right)$ the roles of, respectively, $d_jt\left(r|\,w_j\right)$ and $e_kt\left(s|\,x_k\right)$. Zeros plays the role of the remaining terms. Note that the first term of the efficient influence function is conditionally mean zero given (W,X,R,S) as required. The second and third terms are conditional mean zero given, respectively, R and S as required.

The form of the efficient influence function given in the statement of Theorem 1 coincides $F_{jk}\left(Y,W,X,R,S\right)$ and the variance bound with $\mathbb{E}\left[F_{jk}\left(Y,W,X,R,S\right)^2\right]$.

B Proof of Theorem 2

[To be completed]