DECOMPOSITION ANALYSIS OF SEGREGATION

Kazuo Yamaguchi

University of Chicago

Email: kyamagu@uchicago.edu

ABSTRACT

Although substantive studies on segregation, such as residential or school segregation by race and occupational segregation by gender, are many in sociology, the analytical methodology is almost exclusively focused on measurement issues. This paper introduces a set of two statistical models for the decomposition analysis of segregation. These models can be regarded as a tool to analyze whether one dimension of racial or gender inequality is related to another dimension of inequality, because they can assess, for example, the extent to which gender differences in human capital are related to gender segregation in occupation. One of the new models is a simple extension of the DiNardo-Fortin-Lemieux decomposition method of inequality, which implicitly assumes a supply-driven determination of positional status attainment, and another modifies it to incorporate demand-based macrosocial size constraints on positional status attainment, but both models rely on Rubin's conception of modeling counterfactual outcomes and inverse-probability-of-treatment weighting based on propensity score. An application focuses on gender segregation in occupation in Japan and will lead to a paradoxical result: equalizing human capital and labor supply characteristics between men and women increases, rather than decreases, gender segregation in occupation. Although the underlying behavioral mechanism for gender differences in occupational choice remains to be investigated, the analysis clarifies at least demographically why segregation increases under the counterfactual situation.

DECOMPOSITION ANALYSIS OF SEGREGATION

I. INTRODUCTION

Segregation analysis is a central topic in sociological research. Substantive research on segregation includes numerous studies of residential segregation and school segregation by race and ethnicity, and occupational segregation by gender. This paper introduces two alternative mutually complementary models for the decomposition of the extent of segregation into "explained" and "unexplained" components. Generally, this paper is concerned with the methodology of estimating the unique effect of the group variable X on segregation when the covariates of the group variable also affect segregation and, therefore, the observed effect of X on segregation is confounded by the effect of the association of X with covariates on segregation. The "explained" component of the segregation represents the extent of segregation that is explained by differences in the covariate distribution between the group with $X = 1$ and the group with $X = 0$, and the "unexplained" component represents the remaining extent of segregation. Substantive examples may include (1) a case where race and household income affect the choice of residential district, and we are concerned with the hypothetical extent of residential racial segregation after eliminating the indirect combined effect of (a) the association between race and household income and (b) the effects of household income on the choice of residential district; and (2) a case where gender is associated with the characteristics of educational attainment, such as years of education and college majors, and both gender and educational characteristics affect the choice of occupation, and we are concerned with the hypothetical extent of gender segregation in occupation after eliminating the indirect combined effect of (a) the association of gender and educational characteristics and (b) the effects of education on the choice of occupation. In both cases, the models described in this paper eliminate the indirect effect by removing factor (a) by attaining statistical independence between the group variable and covariates.

The two models introduced in this paper both rely on propensity-score weighting (Rubin 1985; Robins 1998), employed in Rubin's causal model (RCM). One of them is a simple extension of the DiNardo-Fortin-Lemieux (hereafter DFL) method (DiNardo et al. 1997) for decomposition analysis, extended for a polytomous outcome variable and applied to decompose the index of dissimilarity. The other model, which is newly introduced in this paper, is what I refer to as *the matching model*.[1] Like the RCM, the DFL method makes the SUTVA (stable unit treatment assignment) assumption, and this assumption leads to the fact that the distribution of outcomes, such as occupational distribution in the analysis of gender segregation in occupation and residential distribution in the analysis of residential racial segregation, depends only on the supply-side characteristics, namely, characteristics of people who attain those occupational or residential positions. As a result, when we consider a counterfactual situation, such as that which would be realized if women came to have the same human capital characteristics as men's occupational distribution changes according to changes in the supply-side labor characteristics. On the other hand, the matching model assumes that the outcome distribution is determined only by the demand-side characteristics, so that job vacancies or residential vacancies will be filled to the extent to which demand for the occupants of those positions exists, and, therefore, changes in the supply-side characteristics conceived under a counterfactual situation do not affect the outcome distribution of positions but alter only the matching between people and positions. The matching model does not satisfy the SUTVA assumption, because it introduces macrosocial

[1] The term "matching" here has nothing to do with the matching method of the RCM. As described below, matching here implies matching between people and positions.

constraints on the outcome distribution (Yamaguchi 2011). However, as shown in this paper, the matching model retains all other characteristics of the DFL method.

If in reality the attainment of social positions will neither be completely supply driven nor completely demand driven, but will depend on both the supply and the demand for those positions, the results from both the model based on the DFL method and the matching model identify a range of outcomes in which the counterfactual outcome will be realized.

Not unlike the DFL method, the two models for the decomposition of segregation are not methods for causal analysis, because I do not intend to claim that the covariates included exhaust all determinants of the outcome other than the group variable X, and, therefore, omitted-variable bias certainly exists. In addition, the issue of controlling endogenous intervening variables also exists for the DFL method in making a causal inference (Yamaguchi 2015). Instead, the purpose of the two models introduced in this paper is to separate the observed extent of segregation into a component explained by the association of the group variable X with a given set of covariates $\mathbf{V}$ that also affect the outcome and the remaining unexplained component—without claiming that the unexplained effect is a causal effect.

Methodological interest in segregation analysis has been centered on which measure has the most desirable characteristics (e.g., Lieberson 1976; Winship 1977; Blackburn and Marsh 1991; Reardon and Firebaugh 2002; Reardon and O'Sullivan 2004), given the fact that the index of dissimilarity, which is still a representative measure of segregation, may be better modified to reflect certain additional characteristics such as distance among the units of segregation for measuring residential segregation. This is an important methodological issue but is largely independent of the topic of this paper, because the decomposition methods introduced in this paper can be combined with different segregation indices of the researcher's choice as long as they are modifications of the index of dissimilarity based on the measurement of differences between groups in the categorical distribution of segregation units. Charles and Grusky (1995, 2004) introduced loglinear and log-multiplicative models for segregation focusing on gender segregation in occupation. Their method measures segregation based on characteristics of the odds ratio and requires a distinct methodological approach to the decomposition of segregation. This paper, however, focuses on the decomposition of segregation based on the index of dissimilarity. Below, I introduce the decomposition of the index of dissimilarity by an extension of the DFL method and by the matching model and apply them to an analysis of gender segregation of occupation in Japan.

II. METHODS

II_1 **A brief review of the DiNardo-Fortin-Lemieux method**

Unlike the Blinder-Oaxaca decomposition method (Blinder 1973; Oaxaca 1973), which relies on a pair of regression equations for the outcome, the DFL method does not assume any parametric model for the outcome. As Barskey et al. (2002) point out, a major limitation of the Blinder-Oaxaca method is that it requires a linear relationship between the dependent variable and its covariates. On the other hand, the DFL method, which does not assume such a relationship, is considered (Fortin et al. 2011) to be a method that generates a more robust estimate for decomposition than the methods that rely on regression models.

The pair of equations assumed for men and women can be expressed in the DFL method as $y^M = \phi(\mathbf{V}^M, \boldsymbol{\theta}^M) + \varepsilon$ for men and $y^W = \phi(\mathbf{V}^W, \boldsymbol{\theta}^W) + \varepsilon$ for women, where ϕ is an unspecified function, $\mathbf{V}^M$ and $\mathbf{V}^W$ are covariates of the outcome, and $\boldsymbol{\theta}^M$ and $\boldsymbol{\theta}^W$ are parameters that indicate the covariate effects of the outcome, for men and women, respectively. Under the assumed independence of the error term from the covariates, we may obtain the following decomposition:

$$\overline{y}^W - \overline{y}^M = \left[\overline{\phi}(\mathbf{V}^W, \boldsymbol{\theta}^W) - \overline{\phi}(\mathbf{V}^M, \boldsymbol{\theta}^W))\right] + \left[\overline{\phi}(\mathbf{V}^M, \boldsymbol{\theta}^W) - \overline{\phi}(\mathbf{V}^M, \boldsymbol{\theta}^M))\right]. \quad (1)$$

The first component—the "explained" component—of equation (1) reflects inequality in the mean of Y that would be eliminated if women had men's covariate distribution, and the second component—the "unexplained" component—reflects inequality that would remain because women having the same covariate distribution as men's are treated differently in the society. Since the estimates of $\overline{\phi}(\mathbf{V}^M, \boldsymbol{\theta}^M)$ and $\overline{\phi}(\mathbf{V}^W, \boldsymbol{\theta}^W)$ are simply sample means, we only need an estimate of $\overline{\phi}(\mathbf{V}^M, \boldsymbol{\theta}^W)$, which can be obtained from the following equation for the gender dummy variable X that takes a value of 0 for men and 1 for women:

$$\begin{aligned}\overline{\phi}(\mathbf{V}^M, \boldsymbol{\theta}^W) &\equiv \int_{\mathbf{v}} \phi(\mathbf{v}, \boldsymbol{\theta}^W) p(\mathbf{v} \mid X=0) d\mathbf{v} = \int_{\mathbf{v}} E(Y^W \mid \mathbf{v}) p(\mathbf{v} \mid X=0) d\mathbf{v} \\ &= \int_{\mathbf{v}} \omega(\mathbf{v}) E(Y^W \mid \mathbf{v}) p(\mathbf{v} \mid X=1) d\mathbf{v} = E_{\omega}(Y^W), \qquad (2)\end{aligned}$$

where $p(\mathbf{v} \mid X)$ indicates the conditional probability density of $\mathbf{V}$, and E_{ω} indicates the weighted mean with weights:

$$\omega(\mathbf{v}) \equiv \frac{p(\mathbf{v} \mid X=0)}{p(\mathbf{v} \mid X=1)} = \frac{p(X=0 \mid \mathbf{v}) p(\mathbf{v}) / p(X=0)}{p(X=1 \mid \mathbf{v}) p(\mathbf{v}) / p(X=1)} = \frac{p(X=1) p(X=0 \mid \mathbf{v})}{p(X=0) p(X=1 \mid \mathbf{v})}. \quad (3)$$

Unlike the Blinder-Oaxaca method, the DFL method can be applied to the decomposition of difference in proportion, as in the decomposition analysis of gender difference in the proportion of managers, because it does not assume any outcome regression model. The method introduced in this paper takes this advantage of the DFL method further by extending its application to the case of a polytomous dependent variable.

II_2. **The DFL model—where the marginal distribution of outcomes is endogenously determined under counterfactual situations**

I introduce below a decomposition method, which I call the DFL model, for the index of dissimilarity (hereafter the ID), using an example of gender segregation in occupation, by extending the DFL method. We are interested in the decomposition of $P_j^M - P_j^W$, for a set of occupations with $j = 1, \dots, J$, where P_j^M is the proportion, among men who have jobs, of those in the j-th occupation, and P_j^W is the proportion, among women who have jobs, of those in the j-th occupation, into the component attributable to gender differences in covariates **V**, such as educational attainment, age, marital status, and hours of work, and the unexplained component. If this decomposition is attained, then we can decompose the ID, defined as $(1/2)\sum_j |P_j^M - P_j^W|$, into the unexplained component, defined as $(1/2)\sum_j |(P_j^M - P_j^W)_{\text{unexplained}}|$ —where $(P_j^M - P_j^W)_{\text{unexplained}}$ is the unexplained gender difference in the proportion estimated by the method described below—and the explained component, which is the difference between the observed ID value and the unexplained ID value.

Generally, we assume that the marginal distribution of the group variable X, such as gender or race, is fixed under any counterfactual situation. On the other hand, the marginal distribution of covariates **V** summed over the states of X depends on a particular counterfactual situation that we assume but is assumed nonetheless to be exogenously determined independently of the association between X and **V**. First, in this section we assume further a model that implicitly makes the SUTVA assumption (Morgan and Winship 2007) for potential outcomes. With this assumption, we can define the saturated linear probability model for the outcome with categorical covariates for the probability of having outcome *j* as a function of the covariates of individuals and outcome-specific parameters as follows:

$$\Pr(Y_i = j) \equiv P_{ij} = \alpha_j(\mathbf{v}_i;\boldsymbol{\theta}_{0j}) + \beta_j(\mathbf{v}_i;\boldsymbol{\theta}_{1j})X_i, \quad \text{for } j = 1,\ldots,J-1, \tag{4}$$

where Y_i is a categorical outcome variable for person *i*, P_{ij} is the probability for person *i* with covariates $\mathbf{v}_i$, where $\mathbf{v}_i$ is the value of $\mathbf{V}$ observed for person *i*, to attain the *j*-th occupation, $\alpha_j(\mathbf{v}_i;\boldsymbol{\theta}_{0j})$ represents the "main effects" of covariates **V** with the outcome-specific parameters $\boldsymbol{\theta}_{0j}$, X_i is a gender dummy variable that takes a value of 0 for women and a value of 1 for men for person i, and $\beta_j(\mathbf{v}_i;\boldsymbol{\theta}_{1j})X_i$ represents heterogeneous effects of gender, with heterogeneity varying with covariates $\mathbf{V}$, on the outcome with the outcome-specific parameters $\boldsymbol{\theta}_{1j}$. Note that although it may be thought that a linear probability model will generate inconsistent parameter estimates, we assume here a saturated model with categorical covariates where the estimated probabilities are equal to sample proportions for each combined state of covariates, and for any saturated model no inconsistency issue exists. Note also that the marginal distribution of the *j*-th occupation is determined by the effects of *X* and **V** on P_{ij} characterized by parameters $\alpha_j(\mathbf{v}_i;\boldsymbol{\theta}_{0j})$ and $\beta_j(\mathbf{v}_i;\boldsymbol{\theta}_{1j})$. In other words, the size of each occupation here is implicitly assumed to be endogenously determined by supply-side characteristics of the labor market, that is, the joint distribution of $(X,\mathbf{V})$ for a given set of parameters $\{\boldsymbol{\theta}_{0j},\boldsymbol{\theta}_{1j}\}$.

From equation (1), we obtain $\Pr(Y = j \mid X = 0) \equiv P_j^W = E\left(\alpha_j(\mathbf{v}) \mid X = 0\right)$ and $\Pr(Y = j \mid X = 1) \equiv P_j^M = E\left(\alpha_j(\mathbf{v}) \mid X = 1\right) + E\left(\beta_j(\mathbf{v}) \mid X = 1\right)$, where P_j^W and P_j^M are the means of P_{ij} for men and women, respectively, with each outcome *j*, and those means for men and women are equal to their sample proportions for saturated models. It follows that

$$P_j^M - P_j^W = \left\{E\left(\alpha_j(\mathbf{v}) \mid X=1\right) - E\left(\alpha_j(\mathbf{v}) \mid X=0\right)\right\} + E\left(\beta_j(\mathbf{v}) \mid X=1\right). \qquad (5)$$

Equation (5) indicates that the observed gender difference in the proportion of having the *j*-th job is the sum of the gender difference in the main average covariate effects and the average gender effect among men.

Suppose that we realize a counterfactual situation in sample data where the covariate distribution becomes statistically independent of *X* by multiplying weights $\omega(\mathbf{v})$ to sample observations. Then, since $E_\omega\left(\alpha_j(\mathbf{v}) \mid X=1\right) = E_\omega\left(\alpha_j(\mathbf{v}) \mid X=0\right)$ holds, where E_ω indicates the weighted mean, under $\mathbf{V} \perp X$, we obtain

$$E_\omega(Y=j \mid X=1) - E_\omega(Y=j \mid X=0) = E_\omega\left(\beta_j(\mathbf{v}) \mid X=1\right) = E_\omega\left(\beta_j(\mathbf{v})\right), \qquad (6)$$

and, therefore, $E_\omega(Y=j \mid X=1) - E_\omega(Y_j=j \mid X=0)$ reflects the weighted average of the gender effect. Note that the conditioning by $X = 1$ for β can be omitted, because the weighted average of β depends only on the weighted distribution of **V**, which is made equal between $X = 1$ and $X = 0$.

Regarding the distribution of **V**, which is to be made equal between the group with $X = 1$ and the group with $X = 0$ for the weighted data, we can consider three meaningful alternative situations:

(1) a counterfactual situation where both men and women come to have the average covariate distribution $P(\mathbf{v})$,

(2) a counterfactual situation where women come to have men's covariate distribution $P(\mathbf{v} \mid X=1)$, and

(3) a counterfactual situation where men come to have women's covariate distribution $P(\mathbf{v} \mid X=0)$.

As shown below, these three counterfactual situations are similar to those considered for the estimation of the average treatment effect, the average treatment effect for the treated, and the average treatment effect for the untreated (Morgan and Winship 2007)—although we do not make a causal interpretation of those "treatment effects," for reasons explained above. In addition, the distinction between $X = 1$ and $X = 0$ is not a distinction between the treatment group and the control group here, and we will therefore refer to the effect of X under the three alternative situations for the covariate distribution as the average group effect (AGE), the average group effect for group 1 (AGE1) and the average group effect for group 0 (AGE0).

As initially shown by Rubin (1985; see also Robins [1998], and described by Morgan and Winship [2007]), the AGE can be estimated by using the following sets of inverse-probability-of-treatment (IPT) weights:

$$\omega_1(\mathbf{v}) = \frac{P(\mathbf{v})}{P(\mathbf{v} \mid X = 1)} = \frac{P(X = 1)}{P(X = 1 \mid \mathbf{v})} \qquad (7)$$

for men, for whom $X = 1$ holds, and

$$\omega_0(\mathbf{v}) = \frac{P(\mathbf{v})}{P(\mathbf{v} \mid X = 0)} = \frac{P(X = 0)}{P(X = 0 \mid \mathbf{v})} \qquad (8)$$

for women, for whom $X = 0$ holds.

By those weightings, the covariate distributions for both men and women become nearly equal when the propensity score, $P(X = 1 \mid \mathbf{v})$, is adequately estimated. Note that when $P(X = 1 \mid \mathbf{v})$ can be estimated nonparametrically because we have observations for both $X = 1$ and $X = 0$ for each combined state of the covariates, the conversion of $P(\mathbf{v} \mid X = 1)$ and $P(\mathbf{v} \mid X = 0)$ to $P(\mathbf{v})$ by using their nonparametric estimates for the IPT weighting will be perfect. For a parametric estimate of $P(\mathbf{v} \mid x)$, we expect that the weighting, when the propensity score is adequately estimated, will attain statistical independence between X and $\mathbf{V}$. However, we need to diagnose this expected property of the IPT weights, since this property may not be attained either when we have a model of $P(\mathbf{v} \mid x)$ that does not fit the data or when observations that lie outside the "common area of support" for the propensity score generate nonnegligible bias in attaining statistical independence between *X* and **V**. This issue will be addressed in the application.

It is also noteworthy that in order to retain the marginal distribution of *X*, weights $\omega_1(\mathbf{v})$ and $\omega_0(\mathbf{v})$ need to be replaced by the adjusted weights $\omega_1{*}(\mathbf{v})$ and $\omega_0{*}(\mathbf{v})$ for the ratio estimation so that the sum of the weights becomes equal to the sample size in each group of *X*.

$$\omega_1{*}(\mathbf{v}_i) = \frac{N_1 \omega_1(\mathbf{v}_i)}{\sum_{k=1|X=1}^{N_1} \omega_1(\mathbf{v}_k)} \text{ and } \omega_0{*}(\mathbf{v}_i) = \frac{N_0 \omega_0(\mathbf{v}_i)}{\sum_{k=1|X=0}^{N_0} \omega_0(\mathbf{v}_k)}, \tag{9}$$

where N_1 and N_0 indicate the number of sample observations with $X = 1$ and $X = 0$, respectively. Then the unexplained difference in the proportion in unit *j* between men and women is given as

$$E_{\omega_1{*}}(Y = j \mid X = 1) - E_{\omega_0{*}}(Y = j \mid X = 0), \tag{10}$$

where $E_{\omega_1{*}}(Y = j \mid x)$ indicate the weighted average probability of being in unit *j* with adjusted weights of equation (9). This is the estimate for the AGE.

When we consider an alternative counterfactual situation where women have the covariate distribution of men, the weights for women and their adjustment for the ratio estimation are given, as in equation (3), as

$$\omega_{01}(\mathbf{v}) \equiv \frac{P(\mathbf{v} \mid X = 1)}{P(\mathbf{v} \mid X = 0)} = \frac{P(X = 0)P(X = 1 \mid \mathbf{v})}{P(X = 1)P(X = 0 \mid \mathbf{v})}, \text{ and} \tag{11}$$

$$\omega_{01}{*}(\mathbf{v}_i) = \frac{N_0 \omega_{01}(\mathbf{v}_i)}{\sum_{k=1|X=0}^{N_0} \omega_{01}(\mathbf{v}_k)}. \tag{12}$$

Then the unexplained difference in the proportion in unit *j* between men and women is given as

$$E(Y=j \mid X=1)-E_{\omega_{01}*}(Y=j \mid X=0). \qquad (13)$$

This is the estimate for the AGE1. Similarly, the estimate for the AGE0 is obtained for the difference in the proportion in the counterfactual situation where men have the covariate distribution of women by switching men and women in formulas (11) through (13).

II_3. **The matching model—where the marginal distribution of outcomes is exogenously determined under counterfactual situations**

In the model introduced in the previous section, it is implicitly assumed that the size of the outcome categories such as the proportion of people in each occupation, is endogenously determined and can change depending on supply-side characteristics, characterized by the joint distribution of the group variable X and covariates **V**. Let us demonstrate this first by using a simplified case of dichotomous occupational distinction, such as managers and other employees, and dichotomized covariate states, which we denote by V. We assume that the probability distribution of the cross-classification of gender and covariates is as given in Table 1.

(Table 1 about Here)

Then the saturated linear probability of being a manager at the individual level, its average, and the difference in the proportion of managers between men and women are given, respectively, as

$$P_i=\alpha+\beta_1 X_i+\beta_2 V_i+\beta_3 X_i V_i \ , \qquad (14)$$

$$\begin{aligned}\overline{P}&=\alpha+(\beta_1+\beta_2+\beta_3)Q+\beta_1 R+\beta_2 S\\ &=\alpha+\beta_1(Q+R)+\beta_2(Q+S)+\beta_3 Q, \text{ and} \qquad (15)\end{aligned}$$

$$\begin{aligned}\overline{P}^M-\overline{P}^W&=\frac{\alpha(Q+R)+(\beta_1+\beta_2+\beta_3)Q+\beta_1 R}{Q+R}-\frac{\alpha(S+T)+\beta_2 S}{S+T}\\ &=\beta_1+\beta_3\left(\frac{Q}{Q+R}\right)+\beta_2\left(\frac{Q}{Q+R}-\frac{S}{S+T}\right). \qquad (16)\end{aligned}$$

We retain the proportions of men and women, Q+R and S+T, for any of the counterfactual situations considered here. The proportion of $V = 1$, Q+S, is retained only for the average group effect (AGE) but not for the average group effect for group 0 (AGE0) or that for group 1 (AGE1). Then equation (15) shows that neither the AGE0 nor the AGE1 preserves the average proportion of managers under each counterfactual situation, because Q+S changes for each counterfactual situation. On the other hand, the AGE preserves the average proportion of managers if the effects of X and V are additive, but this does not hold when an interaction effect of X and V exists, because then $\overline{P}$ changes with Q, which does not remain constant under any of the three counterfactual situations. This holds in the general case where the covariate V has many more categories, because equation (15) holds for the dichotomous distinction between each category and the rest of the categories combined. Equation (16) simply demonstrates the fact that the confounding covariate effect will be eliminated when X becomes statistically independent of V and, therefore, $Q/(Q+R) = S/(S+T)$ holds.

The model of equation (4) assumes that the sizes of the outcome categories vary freely with the supply-side characteristics of labor under the counterfactual situation. This assumption may be unrealistic, however. In the case of residential segregation, change in the association between race and household income may not change the number of people living in each residential district. Similarly, for the case of occupational segregation by gender, change in the association between gender and education may not change the number of people in each occupation but change only the matching between occupation and people. Below, we consider an alternative model, the matching model, where the sizes of outcome categories are invariant under a counterfactual situation where the group variable and covariates become independent. In the analysis of occupational segregation, this leads to a model where demand by employers, rather than supply-side characteristics, determines the number of occupants of each occupation.

For the matching model, we need to modify the model of equation (4) as follows:

$$\Pr(Y_i = j) \equiv P_{ij} = \phi_j(\mathbf{V}, \mathbf{x})\left\{\alpha_j(\mathbf{v}_i;\boldsymbol{\theta}_{0j}) + \beta_j(\mathbf{v}_i;\boldsymbol{\theta}_{1j})X_i\right\}, \quad \text{for } j = 1, ..., J-1, \qquad (17)$$

where ϕ_j is an outcome-category-specific parameter and is a function of the joint distribution of $\mathbf{V} \equiv \{\mathbf{v}_i\}$ and $\mathbf{x} = \{x_i\}$. Parameters ϕ_j are constant and take a value of 1 for observed data and take values under a given counterfactual situation to make the average probability of attaining each outcome status remain the same as the proportion of people that is observed in the unit. Note that since parameters $\{\phi_j\}$ depend on the macrosocial joint distribution of X and $\mathbf{V}$, the matching model does not satisfy the SUTVA assumption.

The model of equation (17) might be thought to be underidentified, because we assume that $\alpha_j(\mathbf{v}_i;\boldsymbol{\theta}_{0j}) + \beta_j(\mathbf{v}_i;\boldsymbol{\theta}_{1j})X_i$ is the saturated model for observed data with zero degrees of freedom. In fact, equation (17), as well as equation (4), is a model for predicting outcomes for a counterfactual situation. Hence, if conditions to be met for a particular counterfactual situation are given, it can be identified. For the model of equation (17), we assume that a counterfactual situation satisfies the following five conditions four of which, (a) through (d), are shared by the DFL model, and one of which, condition (e), is unique to the matching model:

(a) Parameters $\alpha_j(\mathbf{v}_i;\boldsymbol{\theta}_{0j})$ and $\beta_j(\mathbf{v}_i;\boldsymbol{\theta}_{1j})$ are fixed for each given $\mathbf{v}_i$.

(b) The marginal distribution of the group variable X is fixed and is equal to the observed distribution.

(c) The marginal distribution of the covariates is exogenously imposed on one of the following depending on whether we estimate the AGE, AGE0, or AGE1: $P(\mathbf{v}), P(\mathbf{v} \mid x=1), \text{or } P(\mathbf{v} \mid x=0)$.

(d) The group variable and covariates become statistically independent.

(e) The marginal distribution of the outcome categories $\Pr(Y=j)$ is fixed and is equal to the observed distribution, and this is attained by the adjustment of parameter ϕ_j for each counterfactual joint distribution of *X* and V.

Given the model of equation (17), we obtain the following equalities:

$$E(Y=j \mid X=0) \equiv P_j^W = \phi_j E\left(\alpha_j(\mathbf{v}) \mid X=0\right), \text{and} \tag{18}$$

$$E(Y=j \mid X=1) \equiv P_j^M = \phi_j E\left(\alpha_j(\mathbf{v}) \mid X=1\right) + \phi_j E\left(\beta_j(\mathbf{v}) \mid X=1\right). \tag{19}$$

Hence, under the counterfactual situation where *X* becomes independent of covariates **V** by the use of weights $\omega_M(\mathbf{v})$, which differ from the weights for the DFL method (hereafter, "the DFL weights") as explained below, $E_{\omega_M}\left(\alpha_j(\mathbf{v}) \mid X=1\right) = E_{\omega_M}\left(\alpha_j(\mathbf{v}) \mid X=0\right)$ holds.

Hence, we obtain;

$$E_{\omega_M}(Y=j \mid X=1) - E_{\omega_M}(Y=j \mid X=0) = \phi_j E_{\omega_M}\left(\beta_j(\mathbf{v})\right). \tag{20}$$

The remaining issue is how to obtain estimates of the set of $\{\phi_j\}$ that satisfies conditions (a) through (e) above. Note that ϕ_j is not simply equal to $E(Y=j)/E_{\omega_{DFL}}(Y=j)$, where ω_{DFL} denotes the DFL weights, because the adjustment by $E(Y=j)/E_{\omega_{DFL}}(Y=j)$ does not preserve the statistical independence between *X* and **V**.

Let us denote by $f(x, y, \mathbf{v})$ the observed joint frequency distribution, and denote by $F_1(x, y, \mathbf{v})$ the initial estimate of the counterfactual joint frequency distribution by the DFL method. $F_1(x, y, \mathbf{v})$ is an estimate that, when the propensity score is estimated nonparametrically, satisfies conditions (a) through (d) but may not satisfy condition (e). When the propensity score is estimated parametrically using logit or probit regression, we need to diagnose and make sure that condition (d), regarding statistical independence between X and $\mathbf{V}$, is satisfied in the sample with the DFL weights. When $F_1(x, y, \mathbf{v})$ satisfies conditions (a) through (d), we then need to estimate the expected joint distribution $F(x, y, \mathbf{v})$ that satisfies both $F(x, +, \mathbf{v}) = F_1(x, +, \mathbf{v})$, in consequence of which $F(x, y, \mathbf{v})$ will satisfy conditions (b) through (d), and $F(+, y, +) = f(+, y, +)$, in consequence of which $F(x, y, \mathbf{v})$ will satisfy condition (e). Here the + sign indicates the sum across categories. Condition (a) is also satisfied by the use of multiplicative iterative adjustments of frequencies described below. In the case of the AGE that preserves the marginal distribution of covariates, the above-described model is in fact a special case of the marginal model (Bergsma and Rudas 2002), for which $F(x, +, \mathbf{v}) = f(x, +, +) f(+, +, \mathbf{v})$ and $F(+, y, +) = f(+, y, +)$ hold.

It might be objected that while the marginal model is a loglinear model, we have a modification of the additive model here. Note, however, that the additive component is saturated for the observed data, and we fix it for any counterfactual situation and, therefore, do not need to estimate those parameters. On the other hand, the model of equation (17) is multiplicative, or loglinear, regarding the set of ϕ parameters that we need to estimate and, therefore, we can employ the method of model fitting and parameter estimation for the loglinear model.

We can employ the iterative proportional adjustment, across rounds of iterative estimation starting with $t = 1$, to obtain the final estimate of $F^*(x, y, \mathbf{v})$ that should be attained when the adjustments lead to a convergence.

(1) Calculate $F_{2t-1}(+,y,+)=\sum_{x}\sum_{\mathbf{v}}F_{2t-1}(x,y,\mathbf{v})$.

(2) Calculate $F_{2t}(x,y,\mathbf{v})=F_{2t-1}(x,y,\mathbf{v})\dfrac{f(+,y,+)}{F_{2t-1}(+,y,+)}$.

(3) Calculate $F_{2t}(x,+,\mathbf{v})=\sum_{y}F_{2t}(x,y,\mathbf{v})$.

(4) For the case where $F_1(x,+,\mathbf{v})>0$, calculate $F_{2t+1}(x,y,\mathbf{v})=F_{2t}(x,y,\mathbf{v})\dfrac{F_1(x,+,\mathbf{v})}{F_{2t}(x,+,\mathbf{v})}$.

Then, the ratio of the final estimate to the initial estimate can be expressed as

$$F^*(\mathbf{v},j,x)/F_1(\mathbf{v},j,x)=\hat{\phi}_j\varphi(\mathbf{v},x), \qquad (21)$$

where $\hat{\phi}_j$ is the product of iterative adjustments made in step (2) of the iterative procedure and is the estimate of parameter ϕ_j in equation (17), and $\varphi(\mathbf{v},x)$ is the product of iterative adjustments made in step (4) of the iterative procedure and the modifier of the DFL weights. Note that weights $\omega_{DFL}(\mathbf{v},x)\varphi(\mathbf{v},x)$ preserve the statistical independence between X and $\mathbf{V}$ in $F^*(x,+,\mathbf{v})$ because $F^*(x,+,\mathbf{v})=F_1(x,+,\mathbf{v})$, and they are the weights for the matching model, that is, $\omega_M(\mathbf{v},x)=\omega_{DFL}(\mathbf{v},x)\varphi(\mathbf{v},x)$.

III. APPLICATION

III_1. **Data and the main descriptive characteristics of occupational segregation**

As an illustrative analysis, I focus on gender segregation in occupation among employees of ages 23–59 in Japan based on the 2005 survey of Social Stratification and Mobility (SSM). The SSM survey collects data from a national representative sample of men and women of ages 20–69. Those under age 23 and those over 60 are excluded in this analysis in order to reduce selection bias caused by gender-specific college attendance rates and gender-specific retirement processes. This survey is chosen for this analysis because of its detailed collection of information on occupation and education, including high school types and college majors for the latter.

Gender segregation in occupation in Japan exists for several reasons. Among nonmanual workers, women are overrepresented in clerical occupations and the “middle-and-lower-status” human service professions, which we will refer to as “type-II professionals,” distinguished from other professionals, which we will refer to as “type-I professionals.” By human services, I mean occupations related to medicine, education, human care, and welfare. By the term “middle-and-lower-status,” I imply the exclusion of the high-status professionals in medicine and education, namely, physicians, surgeons, and dentists in medicine, and post-secondary-school teachers in education. Among manual workers, women are overrepresented in service work. Table 2 presents the composition of occupation by gender and the ID, and, for comparison, it also presents numbers for the 1995 SSM survey for Japan and numbers for the United States based on the 2010 population census. The results in Japan are weighted by sampling weights provided by the researchers who conducted the survey. Information on the correspondence of the eight occupational categories to the 2010 USA census classification codes for occupations is available from the author on request.

Table 2 shows that a great deal of gender segregation in occupation exists in both Japan and the United States, and the extent is greater in Japan than in the United States. However, a more important fact is that while there are commonalities and differences between the two countries in the pattern of gender segregation in occupation, those differences always indicate greater disadvantages of women in Japan than in the United States, as explained below.

(Table 2 about here)

Commonalities between Japan and the United States are that women are significantly overrepresented in type-II professional work and clerical work. The differences are that in Japan women are strongly underrepresented in type-I professional and managerial positions and strongly overrepresented in service work. While gender differences exist for those occupations in the United States as well, the extent of gender difference is much smaller. Those differences between Japan and the United States indicate relative disadvantages of women in Japan in comparison to women in the United States. On the other hand, men are much more overrepresented in nonservice manual occupations in the United States than in Japan. This difference between Japan and the United States indicates a relative disadvantage of men in the United States in comparison to men in Japan, given that average earnings are lower for manual workers than for nonmanual workers. Hence, all differences between Japan and the United States in gender segregation in occupation indicate stronger disadvantages of women in Japan than in the United States.

A comparison of the 1995 and 2005 results in Japan also shows that the extent of gender segregation in occupation increased overtime in Japan. This change occurred because both the extent of women's overrepresentation in type-II professional occupations and the extent of women's underrepresentation in nonservice manual work became greater in 2005 than in 1995 despite an increase in gender equality of educational attainment over the decade.

Table 3 also shows that Japanese women are disadvantaged in other respects, that is, in postsecondary education and in employment status. Women are severely underrepresented in receiving a four-year college education in Japan and are overrepresented in irregular employment. The overrepresentation of women in irregular employment comes mainly from a combination of the facts that the majority of women work either part-time or full-time without overtime work and that the majority of men work full-time with overtime work, as shown in Table 3, and there is a very strong association between hours of work and employment status in Japan, such that among those who work less than 35 hours per week, in which women are heavily overrepresented, 91.1% are irregularly employed, among those who work 35-40 hours per week, in which women are somewhat overrepresented, 24.5% are irregularly employed, and among those who work 41 hours or more per week, in which women are strongly underrepresented, 8.7% are irregularly employed. Gender inequality in educational attainment also varies strongly with age. We will examine whether gender inequality in education, age, and labor supply explains part of gender segregation in occupation.

(Tables 3 about here)

III_2. **Decomposition analysis of occupational segregation by gender**

Table 4 presents the main results of the decomposition analysis of occupational segregation based on the data from the 2005 SSM survey for the counterfactual situation where women's covariate distributions become equal to men's (AGE1). They are based on the use of data weighted by sampling weights. Model 1 includes age and education, including the category-by-category interaction effects of education and age, which are found to be significant, as covariates. Model 2 adds to model 1 marital status (married versus single) and number of children (0, 1, 2, and 3 or more) as covariates,[2] including the category-by-category interaction

[2] In the preliminary analysis, I tested the use of a three-category distinction (never married, married, divorced/widowed) for marital status. However, since very few samples with children existed among the never married in Japan, a full cross-classification of this marital-status variable with the number of children could not be made. Hence, I tested a nine-category distinction of combined marital-and-child states by treating "never married" as a single category and cross-classifying "married" and "divorced/widowed" with the number of children. The use of this variable did not improve the fit of the model for predicting the propensity score compared with the use of the combination of number of children with the dichotomous marital status employed in the final analysis.

effects of age and marital status, and the category-by-category interaction effects of marital status and number of children, which are also found to be significant. Model 3 adds to model 2 the six-category distinction of hours of work per week (less than 35 hours, 35-40 hours, 41-49 hours, 50-59 hours, 60 hours or more, and “missing”) and the two interaction effects that are found to be significant in predicting propensity score and are necessary to attain statistical independence between gender and the hours of work in the IP-weighted sample. Those interaction effects are (1) the interaction effect of being married and working less than 35 hours per work and (2) the interaction effect of being a college graduate and working 60 hours or more per week, both of which imply that women are more likely to have those combinations beyond the additive effects of the two variables.

A surprising finding in the results of Table 4 is that, contrary to our expectation, the equalization of those achieved statuses and attributes between men and women increases, rather than decreases, the extent of occupational segregation. The results of Table 4 indicate that the greater the range of achieved statuses in which women obtain status distributions equal to men’s, the greater the extent of segregation, and the tendency is even stronger in the supply-driven DFL model than in the matching model.

(Table 4 about here)

Table 5 presents the predicted occupational compositions for models 2 and 3. Since the men’s occupational composition remains the same as the sample composition for the DFL estimates, the results for the DFL estimates are presented only for women.

(Table 5 about here)

The results in Table 5 indicate that the increase in gender segregation in occupation when women come to have the same distribution of age, education, marital status, and number of children as men in the results of model 2 and the further increase in segregation when women come to have the same distribution of work hours as men's in the results of model 3 come from the fact that the gender gaps in the proportion of type-II professionals and in the proportion of nonservice manual workers both increase when those individual characteristics are equalized. This outcome occurs because when women's educational attainment and work hours, the means of which are lower than men's means, as shown in table 3, are equalized with men's, (a) many more women attain type-II professional occupations, in which women are already overrepresented, than type-I professional occupations, in which women are underrepresented, and (b) many fewer women obtain nonservice manual work, in which women are underrepresented, than manual service occupations, in which women are overrepresented. These two expected changes under the counterfactual situation are very similar to actual changes that took place over a decade from 1995 to 2005 as shown in Table 2.

The increase in the extent of segregation predicted by the matching model is less, because, in the matching model, an increase in type-II professional occupations and a decrease in nonservice manual work are both smaller because constraints on the size of those occupations make changes in men's occupational attainment partly offset the changes in women's occupational attainment predicted by the DFL model.

Table 6 is concerned with a diagnosis of the adequacy of constructed propensity scores. Since model 1 employs a saturated nonparametric model for predicting propensity scores, in which the statistical independence of gender and the covariates is perfect, the diagnosis was made only for models 2 and 3.

(Table 6 about here)

Table 6 presents a test of statistical independence between the gender dummy variable and each covariate before and after the IPT weighting for model 3. Similar results for model 2, where all the *P* values for covariates after the IPT weighting become greater than 0.89, are omitted in Table 6. For the test of independence after the IPT weighting, I employed Clogg and Eliason's method (Clogg and Eliason 1987), which uses the ratio of unweighted to weighted frequencies as "cell weights" for cross-classified data to calculate the chi-square statistics based on unweighted sample counts while testing the independence of the weighted frequencies. However, since cell weights are treated as given by the Clogg-Eliason method rather than as random variables that vary within cells, standard errors are likely to be underestimated and chi-square statistics are likely to be overestimated (Skinner and Vallet 2010). Hence, the Clogg-Eliason method provides a *conservative* test for confirming statistical independence between gender and covariates, which is shown in Table 6 to hold for model 3 after the IPT weighting.

The remaining analysis is concerned with the effect of gender differences in high school types and college majors on gender segregation in occupation. Junior college graduates are classified by the type of high school attended, and those with less than a high school education are treated as a separate single category. Hence, the classification of education in Table 7 distinguishes only three levels of educational attainment. Table 7 presents unweighted sample counts and relative frequency for a classification of educational attainment combined with high school types and college majors by gender based on the data of the 2005 SSM survey.

(Table 7 about here)

Table 7 shows that for certain categories we cannot simply assume that women's relative frequency will become the same as men's for realizing the counterfactual situation for two distinct reasons: men's extreme underrepresentation in two categories, and women's extreme underrepresentation in two other categories. First, there are no men in the sample majoring in home economics or nutrition in college. Making the women's sample proportion equal to the men's is possible, but it leads to a complete ignoring of women in this category. Even though the sample size is not zero, a similar situation exists for graduation from high schools specializing in home economics or nursing. Making the women's relative distribution equal to the men's implies that we are virtually ignoring those people, because the weighted sample size for women will be just three. I decided to omit samples of these two categories from the analysis entirely. A distinct situation exists for engineering high schools, and science and engineering majors in college, in which women are severely underrepresented and women's sample sizes are very small. There are two problems involved in including these categories of people in the analysis. Since the women's sample size is too small, there is a clear lack of corresponding men when the category is combined with age and other covariates, leading to a nonnegligible lack of "common area of support" in propensity score between men and women. Accordingly, I cannot construct a set of propensity scores that attains statistical independence between educational classification and gender when I include those two categories. In addition, their inclusion causes a very small number of women to represent a fairly large number of men in the weighted sample for those categories, thereby leading to unstable results. Hence, I decided to drop samples of those two categories as well from the following analysis to assess the impact of differences in high school type and college major on segregation.

Table 8 presents the results of decomposition analysis for the samples that are retained. In addition to the sample results, the results from two models are presented. Model 3R is a modification of model 3 and includes age, education, marital status, number of children, and work hours and the significant interaction effects included in model 3. Unlike model 3, however, model 3R employs a three-category distinction of educational attainment by combining junior college graduates with high school graduates. This three-category variable for educational attainment is employed for the purpose of comparison with the results of model 4, which employs a further distinction of educational attainment by using the classification of Table 7. In the estimation of propensity scores, the three-category distinction of educational attainment was used for both model 3R and model 4 regarding the category-by-category interaction effects of age and education. The main effects of education differ, however, between the two models because model 4 employs the educational classification of Table 7 while model 3R employs the three-category distinction of educational attainment.

(Table 8 about here)

Table 8 shows for the sample results that the elimination of the four categories of people described above reduces the extent of occupational segregation by 9.6% (0.096 = (.427–.386)/.427). Hence, gender segregation in science and engineering, in which men are strongly overrepresented, and in home economics and nursing, in which women are strongly overrepresented, explains about 10% of gender segregation in occupation. Similar to the results of model 3 in Table 6, results from model 3R in Table 10 indicate that making women's educational attainment, marital status, number of children, and hours of work equal to those of men tends to increase gender segregation in occupation.

The results from model 4 show that making women's high school types and college majors equal to those of men yields a somewhat ambiguous overall effect on the extent of segregation. According to the prediction from the DFL method, it has little effect on segregation (a change from 0.463 to 0.466), but the outcome of the matching model indicates a small increase in segregation, from 0.456 to 0.464. On the other hand, if we compare the change in the distribution of occupation from model 3R to model 4, we find outcomes that are consistent between the predictions from the DFL model and the matching model. The further equalization of high school types and college majors reduces the gender gap in the proportion of type-II professionals but increases the gender gap in the proportion of clerical workers, and those two changes largely offset each other in yielding the overall extent of segregation. The major remaining gap in high school types and college majors comes from a larger proportion of high school graduates from general high schools and a smaller proportion of college graduates with social science majors among women than among men. Somewhat unexpectedly, women with social science majors in college have a much higher probability of becoming clerical workers (62%) than women who graduated from general high schools (33%), while the probability of becoming type-II professionals is slightly lower for the former (18%) than for the latter (20%). Hence, the results in model 4 reflect female college graduates' strong tendency to become clerical workers and the consequent increase in clerical occupation among women when the women's proportion of college graduates with social science majors is equated with men's, and a consequent reduction in the proportion of women in the type-II profession.

IV. CONCLUSION

This paper introduced a new method for the decomposition analysis of segregation. The juxtaposition of the model based on the supply-driven DFL method and the matching model, where the occupational size is constrained by demand, will be useful in predicting the range of possible outcomes under a given counterfactual situation. The semiparametric approach employed for the method will also generate robust estimates for the decomposition.

The illustrative analysis of the decomposition of occupational segregation by gender in Japan showed a paradoxical result that the gender equalization of human capital and labor supply characteristics such as educational attainment and hours of work will increase the extent of gender segregation in occupation. It is clear that the role that gender differences in human capital and labor supply play in generating the gender gap in occupational attainment is quite different from the role that gender differences in human capital and labor supply play in generating the gender gap in income, because we usually observe that equalizing the gender gap in human capital and work hours leads to a smaller gender gap in income. It is clear we need further research on the role that gender plays in occupational opportunities and attainment, and in the consequent gender segregation in occupation.

REFERENCES

Barsky, Robert, John Bound, Kerwin K. Charles, and Joseph P. Lupton. 2002. “Accounting for Black-White Wealth Gap: A Nonparamaric Approach.” *Journal of the American Statistical Association* 93: 663-73.

Bergsma, W.P. and T. Rudas. 2002. “Marginal Models for Categorical Data.” *Annals of Statistics* 30: 140–59.

Blackburn, R.M. and C. Marsh. 1991. “Education and Social Class: Revisiting the 1955 Education Act with Fixed Marginals.” *British Journal of Sociology* 42: 507–36.

Blinder, Alan S. 1973. “Wage Discrimination: Reduced Form and Structural Variables.” *Journal of Human Resources* 8: 436-55.

Charles, Maria. and David B. Grusky. 1996. “Models for Describing the Underlying Structure of Sex Segregation.” *American Journal of Sociology* 100: 931‑71.

Charles, Maria. and David B. Grusky. 2004. *Occupational Ghettos: The Worldwide Segregation of Men and Women*: Palo Alto, CA: Stanford University Press.

Clogg, Clifford C. and Scott. R. Eliason. 1987. “Some Common Problems in Loglinear Analysis.” *Sociological Methods & Research* 16: 8-44.

DiNardo, J., N. Fortin, and T. Lemieux 1996. “Labor Market Institution and the Distribution of Wages.” *Econometrica* 64:1001–44.

Fortin, N., T. Lemieux, and S. Firpo. 2011. “Decomposition Methods in Econometrics.” Pp. 1–102 in *Handbook of Labor Economics*, Volume 4a, edited by O. Ashenfelter and D. Card. New York: Elsevier B.V.

Lieberson, Stanley. 1976. “Rank-Sum Comparisons between Groups.” *Sociological Methodology* 7: 276–91.

Morgan, Stephen L. and Christopher Winship, 2007. *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press.

Oaxaca, Ronald L. 1973. “Male-Female Wage Differentials in Urban labor Markets.” *International Economic Review* 14: 693-709.

Reardon, Sean F. and Glenn Firebaugh. 2002. “Measures of Multigroup Segregation.” *Sociological Methodology* 32: 33‐67.

Reardon, Sean F. and David O’Sullivan. 2004. “Measures of Spatial Segregation.” *Sociological Methodology* 34: 121–62.

Robins, James M. 1998. “Marginal Structural Models.” Pp. 1–10 in *Proceedings of the American Statistical Association*, *Section on Bayesian Statistical Science*.

Rubin, Donald B. 1985. “The Use of Propensity Scores in Applied Bayesian Inference”. Pp. 463-72 in J. M. Bernardo, M. H. De Groot, D. V. Lindley and A. F. M. Smith (eds.), *Bayesian Statistics,* vol. 2. North-Holland: Elsevier.

Skinner, Chis J. and Louis-André Vallet. 2010. “Fitting Log-linear Models to Contingency Tables from Surveys with Complex Sampling Designs: An Investigation of the Clogg-Eliason Approach.” *Sociological Methods & Research* 39: 83-108.

Winship, Christoper. 1977. “A Reevaluation of the Indices of Residential Segregation.” *Social Forces* 55: 1058–66.

Yamaguchi, Kazuo. 2011. “A Macrocounterfactual Analysis of Group Differences: An Application to the Gender Wage Gap in Japan.” *Sociological Methodology* 41:223-55..

Yamaguchi, Kazuo. 2015. “Decomposition of Gender or Racial Inequality with Endogenous Intervening Covariates: An Extension of the DiNardo-Fortin-Lemieux Method.” *Sociological Methodology* 45: 388-428.

Table 1. Cross-Classified Proportions of Gender and a Covariate

	Men: $X = 1$	Women: $X = 0$
$V = 1$	Q	S
$V = 0$	R	T

$Q+R+S+T=1$

Table 2. Gender Segregation in Occupation

	Japan, 2005		Japan,1995		U.S.A., 2010	
	Men	Women[1]	Men	Women[1]	Men	Women
Sample size	1,261	1,187	1,468	1,084	81,323,085	72,714,395
	Composition		composition		Composition	
1. Professional, type I	0.116	0.018*	0.094	0.019*	0.156	0.127
2. Professional, type II	0.041	0.196*	0.048	0.142*	0.043	0.208
3. Managerial/administrative	0.100	0.007*	0.101	0.012*	0.108	0.075
4. Clerical	0.167	0.329*	0.218	0.345*	0.070	0.219
5. Sales	0.131	0.104*	0.093	0.113	0.106	0.120
6. Manual excluding service[2]	0.305	0.159*	0.294	0.197*	0.255	0.047
7. Manual service	0.026	0.136*	0.031	0.132*	0.117	0.158
8. Other[3]	0.114	0.050*	0.121	0.040*	0.145	0.046
Index of dissimilarity	0.427		0.343		0.369	

[1]The asterisk indicates, for Japanese data, significant gender difference.

[2]This category includes production workers, construction workers, and skilled manual workers.

[3]This category includes protective service workers, agricultural workers, and occupations not classifiable.

Table 3. Gender Inequality in Education and Employment Status[1]

(2005 SSM)

	Men	Women[2]
Sample size	1,261	1,187
	Composition	
1. Proportion who are 4-year college graduates	0.377	0.148*
2. Proportion who are irregular workers	0.078	0.520*
3. Composition by hours of work per week		
Less than 35 hours	0.024	0.333*
35-40 hours	0.323	0.398*
41-49 hours	0.235	0.161*
50-59 hours	0.192	0.062*
60 or more hours	0.211	0.032*
Missing	0.015	0.014
Total	1.000	1.000

[1]The results of the 2005 SSM survey.
[2]The asterisk indicates a significant gender difference.

Table 4. Main Results for the Decomposition of Segregation

	Occupational segregation: Sample ID=0.427	
	DFL	Matching
Model 1	0.443	0.430
Model 2	0.450	0.440
Model 3	0.493	0.481

Model 1: age × education.

Model 2: age × education; age × marital status; marital status × number of children.

Model 3: Variables in model 2 plus hours of work per week, marital status × less than 35 hours of work, and college graduation × sixty or more hours of work.

Table 5. Detailed Decomposition Results for Models 2 and 3

	Sample		Model 2			Model 3		
			DFL	Matching		DFL	Matching	
	Men	Women	Women	Men	Women	Women	Men	Women
	Composition.		Composition			Composition		
1. Professional, type I	0.116	0.018	0.041	0.097	0.030	0.024	0.110	0.025
2. Professional, type II	0.041	0.196	0.238	0.033	0.204	0.291	0.017	0.221
3. Managerial/ Administrative	0.100	0.007	0.008	0.100	0.008	0.011	0.097	0.010
4. Clerical	0.167	0.330	0.320	0.167	0.329	0.308	0.164	0.333
5. Sales	0.131	0.104	0.083	0.143	0.091	0.085	0.135	0.099
6. Manual, excluding service	0.305	0.159	0.139	0.315	0.148	0.126	0.318	0.144
7. Manual service	0.026	0.136	0.126	0.027	0.134	0.129	0.026	0.135
8. Other	0.114	0.050	0.045	0.117	0.047	0.022	0.131	0.032
Index of dissimilarity	0.427		0.450	0.440		0.493	0.481	

Table 6. Diagnostic Tests of Statistical Independence between Gender and Covariates after the IPT Weighting

Covariates	Before the IPT Weighting				After the IPT weighting: Model 3		
	L^2	*df*	*P*		L^2	*df*	*P*
Age	14.0	6	0.030		4.17	6	0.654
Education	282.2	3	0.000		0.50	3	0.919
Marital Status	6.3	1	0.012		0.07	1	0.786
Number of Children	14.7	3	0.002		4.02	3	0.259
Hours of work/week	686.8	5	0.000		4.86	5	0.433

Table 7. Gender Differences in Specialization in Educational Attainment[1]

		Men		Women	
Educational Attainment	High school type or college major	N	%	N	%
Middle school	N.A.	102	8.1	61	5.1
High school	General	291	23.1	629	53.0
	Engineering	221	17.5	13	1.1
	Commerce	85	6.7	191	16.1
	Agricultural	62	4.9	18	1.5
	Home economics & Nursing	3	0.2	87	7.3
	Other	22	1.7	12	1.0
College	Education	24	1.9	29	2.4
	Humanities	36	2.9	52	4.4
	Social sciences	215	17.0	36	3.0
	Sciences & engineering	146	11.6	10	0.8
	Medicine & health	18	1.4	15	1.3
	Home economics & nutrition	0	0.0	20	1.7
	Other	37	2.9	14	1.2
Total		1,262	100.0	1,187	100.0
Index of Dissimilarity		0.501			

[1]The results from the 2005 SSM survey.

Table 8. Results with a Detailed Classification of Educational Attainment

N = 890 for men, N = 1,049 for women

	Sample		Model 3R			Model 4		
			DFL	Matching		DFL	Matching	
	Men	Women	Women	Men	Women	Women	Men	Women
	Composition.		Composition			Composition		
1. Professional, type I	0.066	0.015	0.020	0.060	0.020	0.016	0.064	0.016
2. Professional, type II	0.049	0.185	0.297	0.016	0.213	0.227	0.040	0.192
3. Managerial /administrative	0.101	0.008	0.013	0.095	0.012	0.007	0.102	0.007
4. Clerical	0.197	0.340	0.313	0.192	0.345	0.377	0.166	0.367
5. Sales	0.160	0.107	0.079	0.171	0.097	0.079	0.172	0.097
6. Manual, excluding service	0.286	0.156	0.121	0.305	0.140	0.126	0.299	0.146
7. Manual service	0.029	0.136	0.127	0.030	0.135	0.137	0.027	0.138
8. Other	0.112	0.052	0.030	0.131	0.037	0.031	0.131	0.036
Index of dissimilarity	0.386		0.463	0.456		0.466	0.464	