Wage Differentials in the Presence of Unobserved Worker, Firm, and Match Heterogeneity*

Simon D. Woodcock†
Simon Fraser University
simon_woodcock@sfu.ca

May 2007

Abstract

We consider the problem of estimating and decomposing wage differentials in the presence of unobserved worker, firm, and match heterogeneity. Controlling for these unobservables corrects omitted variable bias in previous studies. It also allows us to measure the contribution of unmeasured characteristics of workers, firms, and worker-firm matches to observed wage differentials. An application to linked employer-employee data shows that decompositions of inter-industry earnings differentials and the male-female differential are misleading when unobserved heterogeneity is ignored.

JEL Codes: J31, C23

Keywords: wage differentials, unobserved heterogeneity, employer-employee data

*This document reports the results of research and analysis undertaken by the U.S. Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This document is released to inform interested parties of ongoing research and to encourage discussion of work in progress. This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD), which is partially supported by the National Science Foundation Grants SES-9978093 and SES-0427889 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging Grant R01~AG018854, and the Alfred P. Sloan Foundation. The views expressed herein are attributable only to the author(s) and do not represent the views of the U.S. Census Bureau, its program sponsors or data providers. Some or all of the data used in this paper are confidential data from the LEHD Program. The U.S. Census Bureau supports external researchers' use of these data through the Research Data Centers (see www.ces.census.gov). For other questions regarding the data, please contact Jeremy S. Wu, Manager, U.S. Census Bureau, LEHD Program, Demographic Surveys Division, FOB 3, Room 2138, 4700 Silver Hill Rd., Suitland, MD 20233, USA. (Jeremy.S.Wu@census.gov http://lehd.dsd.census.gov).

†Correspondence to: Department of Economics, Simon Fraser University, 8888 University Dr., Burnaby, BC V5A 1S6, Canada. I thank Julia Lane, Krishna Pendakur, John Abowd, participants at CAFE 2006, and two anonymous referees for helpful comments and suggestions. This research was partially supported by the SSHRC Institutional Grants program and NSF Grant SES-0339191 to Cornell University.

1 Introduction

It is well documented that there are large, persistent, unexplained wage differentials in most labor markets. Among those that have received the most intense scrutiny are the male-female differential, the black-white differential, the union wage gap, and inter-industry differentials. A variety of explanations have been posited for observed differences between earnings of various groups, ranging from labor market discrimination to unobserved heterogeneity. A vast literature has sought to decompose and explain these differentials using various regression-based methods. However, regression-based estimates are subject to bias in the presence of unobserved heterogeneity – even if unobserved heterogeneity is not the actual cause of the observed differential.

A recent literature based on linked employer-employee data has shown that unobserved characteristics of workers, firms, and worker-firm matches account for the vast majority of wage dispersion. In this paper, we consider the problem of estimating and decomposing wage differentials in the presence of these unobserved characteristics. Our main contribution is to generalize existing regression-based decompositions of wage differentials to account for unobserved worker, firm, and match heterogeneity. Controlling for these unobservables corrects omitted variable bias in previous studies. It also allows us to measure the contribution of unmeasured characteristics of workers, firms, and worker-firm matches to observed wage differentials.

We focus on two recent empirical specifications. The more general of the two is the match effects model of Woodcock (2006). This specification controls for observable and unobservable characteristics of workers (person effects), unmeasured characteristics of their employers (firm effects), and unmeasured characteristics of worker-firm matches (match effects). The match effects model admits decompositions of wage differentials that are robust to unmeasured worker, firm, and match characteristics; and differential sorting of workers across firms and worker-firm matches. The second specification is the special case that arises in the absence of match effects. This is the person and firm effects model of Abowd et al. (1999). This specification is more parsimonious than the match effects model, but may be subject to bias if unobserved match characteristics (e.g., match-specific human capital or match quality) are important determinants of wages.

We use these two specifications to estimate and decompose wage differentials using data from the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) database. We focus on two differentials that have received considerable attention from researchers: the male-female differential and inter-industry differentials. The empirical application delivers a clear message: wage decompositions that fail to control for unobserved worker, firm, and match heterogeneity can be misleading.

Our analysis of inter-industry wage differentials illustrates several important points. We show that regression-adjusted inter-industry wage differentials (i.e., the estimated coefficients on industry indicator variables) that do not control for unobserved person, firm, and match heterogeneity are a weighted average of the omitted effects. Consequently, traditional estimates of inter-industry wage differentials confound "pure" industry differentials (which are a characteristic of firms) with unobserved personal and match heterogeneity. Furthermore, even though match effects make a

negligible contribution to observed differences in average earnings between industries, they are important for correcting bias in estimated person and firm effects. In fact, estimates that omit match effects can be very misleading. For instance, estimates based on the person and firm effects model predict that, on average, highly skilled workers sort into employment in low-paying industries. This result is overturned when the empirical specification controls for match effects.

Our analysis of the male-female differential further illustrates that omitted person, firm, and/or match effects result in misleading inferences. Contrary to a specification that omits these effects, we find that male-female differences in the returns to education narrow the male-female wage differential. We also find that ten percent of the overall difference in average earnings between men and women is attributable to women sorting into lower-paying firms. Of this, roughly one third is due to sorting into lower-paying industries, and the remaining two thirds is attributable to sorting into lower-paying firms within industries.

The remainder of the paper is organized as follows. We begin, in Section 2, with a brief review of traditional approaches to estimating and decomposing wage differentials. Section 3 presents the match effects model, and discusses the estimation of wage differentials in the presence of person, firm, and match effects. We describe the data in Section 4, and present the empirical results in Section 5. Section 6 concludes.

2 Wage Differentials: Traditional Approaches

Traditional methods for estimating wage differentials are straightforward and well known. In general, the objective is to explain the difference in average log wages y (or some other measure of compensation) between members of a group g and a reference group: $\bar{y}_g - \bar{y}_0$. The groups are usually defined by observable characteristics of workers (e.g., sex or race) or firms (e.g., industry or size). In what follows, we call $\bar{y}_g - \bar{y}_0$ the *raw wage differential.*

The simplest approach assumes that wages depend on a vector of observable characteristics x_i that earn the same returns β for all groups. Suppose the log wage of individual i is given by:

$$y_i = x_i'\beta + g_i'\delta + \varepsilon_i \tag{1}$$

where g_i is a vector of indicator variables for group membership, δ is a coefficient vector, and ε_i is statistical error. We call the estimated coefficient vector $\hat{\delta}$ the *regression-adjusted* (for x_i) *wage differential* between groups.

In this framework, the raw wage differential between group g and the reference group can be decomposed as $\bar{y}_g - \bar{y}_0 = (\bar{x}_g - \bar{x}_0)'\hat{\beta} + \left(\hat{\delta}_g - \hat{\delta}_0\right)$. The first term is the component of the raw wage differential explained by differences in characteristics between group g and the reference group, and the second term is the unexplained component. This simple approach is most often adopted to analyze wage differentials due to characteristics of firms or jobs, e.g., inter-industry or occupational wage differentials (Krueger and Summers (1988), Katz and Summers (1989), Groshen (1991), Goux and Maurin (1999), and Abowd et al. (2005)), and firm-size wage differentials (see Oi and Idson

(1999) for a review).

The well-known Oaxaca-Blinder decomposition (Blinder (1973), Oaxaca (1973)) generalizes the preceding by allowing the returns to characteristics to differ between groups. In this case, the raw wage differential is $\bar{y}_g - \bar{y}_0 = \bar{x}_g' \hat{\beta}_g - \bar{x}_0' \hat{\beta}_0$, where x_i now includes an intercept for each group. This can be further decomposed in various ways, most commonly:

$$\bar{y}_g - \bar{y}_0 = (\bar{x}_g - \bar{x}_0)' \hat{\beta}_0 + \bar{x}_g' \left(\hat{\beta}_g - \hat{\beta}_0 \right). \tag{2}$$

The first term in (2) measures the component of the wage differential attributable to differences in characteristics between the two groups, evaluated at the returns of the reference group. The second term measures the component attributable to differences in the returns to characteristics, evaluated at the average characteristics of group g. The first term is often referred to as the explained component. The second term is the unexplained component sometimes attributed to labor market discrimination. This decomposition is usually applied to the analysis of wage differentials due to individual characteristics such as sex or race (see, e.g., Blau and Kahn (2003), or Altonji and Blank (1999) for a summary).

Both of these approaches are subject to bias in the presence of omitted variables that are correlated with observable characteristics (including group membership). When researchers have access to panel data on individuals, it is standard to augment the wage equation with a main effect for each individual, θ_i, that controls for unobserved personal heterogeneity. When researchers have access to panel data on firms, it is likewise standard to include a main effect for each firm, ψ_j, that controls for unobserved firm heterogeneity. In a few recent instances based on longitudinal linked employer-employee data, researchers have estimated wage differentials controlling for both unobserved personal and firm heterogeneity (e.g., Goux and Maurin (1999), and Abowd et al. (2005)). In the next section, we introduce an empirical specification that controls for unobserved worker, firm, and worker-firm match heterogeneity. This framework permits decompositions of wage differentials that include components due to unobserved worker, firm, and match heterogeneity, and corrects bias due to omitted variables along these dimensions.

3 The Match Effects Model

The Woodcock (2006) match effects model is:

$$y_{ijt} = \mu + x_{ijt}' \beta + \theta_i + \psi_j + \phi_{ij} + \varepsilon_{ijt} \tag{3}$$

where y_{ijt} is log compensation of worker i at firm j in period t; μ is the grand mean; x_{ijt} is a vector of time-varying observable characteristics that earn returns β; θ_i is a person effect that measures the returns to time-invariant personal characteristics; ψ_j is a firm effect that measures the returns to time-invariant firm characteristics; ϕ_{ij} is a match effect that measures the returns to characteristics of the worker-firm match; and ε_{ijt} is stochastic error.

The person, firm, and match effects may include both observed and unobserved components. Here, we consider the case where:

$$\theta_i = \alpha_i + u_i'\eta \tag{4}$$

where u_i is a vector of time-invariant observable personal characteristics that earn returns η; and α_i is the unobserved component of the person effect.

In general, the person effect will measure persistent differences in compensation between individuals, conditional on observable characteristics, firm effects, and match effects. It is intuitive, even in the absence of a formal economic model, to interpret the portable component of compensation $x_{ijt}'\beta + \theta_i$ as the returns to general human capital.

The firm effect measures persistent differences in compensation between firms, conditional on measured and unmeasured characteristics of workers and match effects. Persistent differences in compensation could arise for a variety of reasons, including productivity differences between firms, firm-specific human capital, product market conditions, monopsony power, compensating differentials, or firm-specific compensation policies. Generally, some labor market friction is necessary for inter-firm compensation differences to persist in equilibrium.

The match effect measures the returns to time-invariant characteristics of worker-firm matches. It is intuitive to interpret this term as the return to match-specific human capital, or the value of production complementarities between the worker and firm. These have similar implications in most instances.

Let N^* denote the total number of observations; N is the number of individuals; J is the number of firms; $M \leq NJ$ is the number of worker-firm employment matches; k is the number of time-varying covariates; and q is the number of time-invariant observable individual characteristics. We rewrite the match effects model in matrix notation:

$$y = \mu + X\beta + D\theta + F\psi + G\phi + \varepsilon \tag{5}$$

$$\theta = \alpha + U\eta \tag{6}$$

where y is the $N^* \times 1$ vector of log compensation; μ is now the $N^* \times 1$ mean vector; X is the $N^* \times k$ matrix of time-varying covariates; β is a $k \times 1$ parameter vector; D is the $N^* \times N$ design matrix of the person effects; θ is the $N \times 1$ vector of person effects; F is the $N^* \times J$ design matrix of the firm effects; ψ is the $J \times 1$ vector of firm effects; G is the $N^* \times M$ design matrix of the match effects; ϕ is the $M \times 1$ vector of match effects; α is the $N \times 1$ vector of unobserved components of the person effect; U is the $N \times q$ matrix of time-invariant individual characteristics; η is a $q \times 1$ parameter vector; and ε is the $N^* \times 1$ error vector.

A special case arises in the absence of match effects. This is the person and firm effects model of Abowd et al. (1999). This specification implies M linear restrictions ($\phi_{ij} = 0$) on the match effects model. Woodcock (2006) finds the data reject these restrictions. We arrive at the same conclusion in the empirical application of Section 5.

3.1 Wage Decompositions

Before discussing identification and estimation of the match effects model, we first illustrate how it contributes to the estimation of wage differentials. First, it corrects bias in the estimated coefficients due to omitted person, firm, and/or match effects. We discuss bias due to omitted effects in Section 3.2. Second, it provides a general decomposition of raw wage differentials into components attributable to differences in observable characteristics, differences in the returns to observable characteristics, and differences in average person, firm, and match effects.

Suppose we are interested in the raw wage differential between group g and a reference group. As in the Oaxaca-Blinder decomposition, we allow the returns to observable characteristics to differ between groups. However, unlike the case where person, firm, and match effects are omitted, it is cumbersome to estimate separate regression models for the two groups. This is because person effects are common to all of an individual's employment spells, and firm effects are common to all of its employees. Consider, for example, the male-female wage differential. Firm j's firm effect, ψ_j, is the same for all of its employees – including men and women. Estimating separate regressions for men and women would therefore imply J cross-equation restrictions (one for each firm effect). It is simpler in practice to estimate a single equation and allow coefficients to vary across groups by interacting observable characteristics with indicator variables for group membership.[1]

When wages are given by the match effects model (3), the raw wage differential between group g and a reference group is:

$$\bar{y}_g - \bar{y}_0 = \left(\bar{x}_g'\hat{\beta}_g - \bar{x}_0'\hat{\beta}_0\right) + \left(\bar{\theta}_g - \bar{\theta}_0\right) + \left(\bar{\psi}_g - \bar{\psi}_0\right) + \left(\bar{\phi}_g - \bar{\phi}_0\right) \tag{7}$$

where overbars indicate sample means, subscripts denote groups, and $\hat{\beta}_g$ and $\hat{\beta}_0$ are estimated elements of β corresponding to group g and the reference group, respectively. The first term in (7) is the component of the raw wage differential attributable to observable characteristics x_{ijt}. Just like the Oaxaca-Blinder decomposition, this can be further decomposed into components attributable to differences in characteristics between groups, and differences in returns to characteristics, e.g.,

$$\bar{x}_g'\hat{\beta}_g - \bar{x}_0'\hat{\beta}_0 = \left(\bar{x}_g - \bar{x}_0\right)'\hat{\beta}_0 + \bar{x}_g'\left(\hat{\beta}_g - \hat{\beta}_0\right). \tag{8}$$

The second term in (7) is the component of the raw wage differential attributable to differences in person effects between groups. It measures the contribution of time-invariant individual characteristics – both observed and unobserved – to the raw wage differential. We further decompose

[1] A single equation restricts the error variance to be the same for all groups. Since we control for unobserved person, firm, and match heterogeneity, this restriction is likely to be satisfied in most instances. An alternative to the approach taken here would be to estimate separate equations for each group and redefine the unobserved components of person, firm, and/or match effects to vary across groups, e.g., separate firm effects for men and women. There are two drawbacks to this approach. One is the increase in computational burden. The second is that the means of unobserved effects are not separately identified from the overall intercept. Hence we can not separately identify the difference between average male and female person, firm, and match effects from the difference between male and female intercepts.

this component as:

$$\bar{\theta}_g - \bar{\theta}_0 = (\bar{u}_g - \bar{u}_0)' \hat{\eta}_0 + \bar{u}_g' \left(\hat{\eta}_g - \hat{\eta}_0\right) + (\bar{\alpha}_g - \bar{\alpha}_0) \tag{9}$$

so that the first term in (9) is the component due to differences in time-invariant personal characteristics between groups, the second term is the component due to differences in the returns to time-invariant personal characteristics, and the third term is the component due to differences in unobserved personal characteristics.

The final two terms in (7) are the components of the raw wage differential attributable to differences in firm effects and match effects between groups. These measure the extent to which raw wage differentials are explained by differential sorting into high- and low-paying firms and worker-firm matches.

The preceding discussion has focused on generalizing the Oaxaca-Blinder decomposition. The match effects model is also useful for estimating wage differentials in the simple case where returns are the same for both groups, i.e., in simple models like (1) where wage differentials are measured by differences in regression intercepts. In this case, the primary benefit of the match effects model is to correct bias in the estimated coefficients, including coefficients on the indicator variables for group membership. More subtly, however, when group membership is a characteristic of workers, firms, or worker-firm matches, the "pure" regression-adjusted differential is the appropriate aggregation of person, firm, or match effects. We now illustrate this for the case of inter-industry wage differentials.

Industry is a characteristic of the firm. Hence, in the presence of firm effects, the "pure" industry effect (as defined by Abowd et al. (1999)) is the correct aggregation of firm effects.[2] The pure industry effect is defined as the one that corresponds to including indicator variables for industry in (3). In this case, we define the remainder of the firm effect as a deviation from industry effects. We now have the augmented regression equation:

$$y_{ijt} = \mu + x_{ijt}'\beta + \theta_i + \kappa_{\mathcal{K}(j)} + \left(\psi_j - \kappa_{\mathcal{K}(j)}\right) + \phi_{ij} + \varepsilon_{ijt} \tag{10}$$

where κ_k is the pure industry effect for industry k, and the function $\mathcal{K}(j) = k$ indicates the industry classification of firm j. In matrix notation,

$$y = \mu + X\beta + D\theta + FA\kappa + (F\psi - FA\kappa) + G\phi + \varepsilon \tag{11}$$

where A is the $J \times K$ matrix that classifies each firm into one of K industries, and κ is the $K \times 1$ vector of pure industry effects. Equation (11) simply defines an orthogonal decomposition of firm effects into industry effects $FA\kappa$, and deviations from industry effects $F\psi - FA\kappa = M_{FA}F\psi$, where $M_Z \equiv I - Z\left(Z'Z\right)^{-1}Z'$ projects onto the column null space of a matrix Z. In this case, the pure industry effects are defined as:

$$\kappa \equiv \left(A'F'FA\right)^{-1} A'F'F\psi.$$

[2] This discussion follows Abowd et al. (1999) and Abowd, Kramarz, and Woodcock (forthcoming), who discuss inter-industry differentials in the presence of person and firm effects.

Hence the pure industry effect for industry k is the duration-weighted average of firm effects:

$$\kappa_k = \sum_{i=1}^{N} \sum_{t=t_i^1}^{T_i} \frac{\mathbf{1}\left(\mathcal{K}\left(\mathcal{J}\left(i,t\right)\right)=k\right)\psi_{\mathcal{J}(i,t)}}{N_k}$$

where $t_i^1, t_i^2, ..., T_i$ denote the periods that person i appears in the sample, the function $\mathcal{J}\left(i,t\right)=j$ indicates the firm j at which worker i was employed in period t, N_k is the number of observations on industry k, and $\mathbf{1}\left(A\right)$ is the indicator function taking value one if A is true and zero otherwise.

The preceding illustrates how we can estimate pure regression-adjusted differentials in the presence of person, firm, and match effects. We need not even include indicator variables for the groups. The pure regression-adjusted wage differential for groups defined by a firm characteristic (such as industry) is simply the duration-weighted average of firm effects in each group. Likewise, the pure regression-adjusted wage differential for groups defined by personal characteristics (e.g., sex or race) or match characteristics (e.g., occupation) is the analogous duration-weighted average of person or match effects, respectively, in each group. We take this approach to estimate inter-industry wage differentials in Section 5.

3.2 Biases Due to Omitted Effects

Abowd et al. (1999) discuss bias due to omitted person and/or firm effects. Woodcock (2006) discusses bias due to omitted match effects. Here, we summarize the latter discussion and derive the bias when all three effects are omitted. These bias expressions help to contextualize the empirical results of Section 5.

3.2.1 Omitted Person, Firm, and Match Effects

When wages are determined according to (3) but the estimated equation excludes the person, firm, and match effects, the estimated returns to time-varying observables, β^*, are biased. In particular, the least squares estimator in the mis-specified model satisfies:

$$E\left[\beta^*\right] = \beta + \left(X'X\right)^{-1} X'\left(D\theta + F\psi + G\phi\right). \tag{12}$$

That is, the estimated returns to observable characteristics equal the true vector of returns, plus an omitted variable bias that we can interpret as the estimated coefficients in an auxiliary regression of the omitted effects on X. The sign and magnitude of the bias depends on the covariance between X and the omitted effects.

To illustrate the bias due to omitted person, firm, and match effects, we return to our example of inter-industry wage differentials. If our estimating equation includes indicator variables for industry, but excludes the remainder of the firm effect, person effects, and match effects, the

estimated industry effects in the mis-specified model satisfy:

$$E\left[\kappa^*\right] = \kappa + \left(A'F'M_XFA\right)^{-1} A'F'M_X\left(D\theta + M_{FA}F\psi + G\phi\right) \tag{13}$$

which, after some algebra, equals

$$E\left[\kappa^*\right] = \left(A'F'M_XFA\right)^{-1} A'F'M_X\left(D\theta + F\psi + G\phi\right). \tag{14}$$

Equation (14) shows that the mis-specified industry effects are the sum of employment-duration-weighted average person, firm, and match effects, given X, in each industry.

In the special case where the design of the industry effects, FA, is orthogonal to X, D, and G, so that $A'F'M_XFA = A'F'FA$, $A'F'M_XD = A'F'D$, $A'F'M_XF = A'F'F$, and $A'F'M_XG = A'F'G$, estimated industry effects in the mis-specified model are exactly the sum of the duration-weighted average person, firm, and match effects. That is, the estimated wage differential for industry k satisfies:

$$\begin{aligned} E\left[\kappa_k^*\right] &= \sum_{i=1}^{N}\sum_{t=t_i^1}^{T_i} \frac{\mathbf{1}\left(\mathcal{K}\left(\mathcal{J}\left(i,t\right)\right)=k\right)\left(\theta_i + \psi_{\mathcal{J}(i,t)} + \phi_{i\mathcal{J}(i,t)}\right)}{N_k} \\ &= \kappa_k + \sum_{i=1}^{N}\sum_{t=t_i^1}^{T_i} \frac{\mathbf{1}\left(\mathcal{K}\left(\mathcal{J}\left(i,t\right)\right)=k\right)\left(\theta_i + \phi_{i\mathcal{J}(i,t)}\right)}{N_k} \end{aligned} \tag{15}$$

Hence estimated inter-industry wage differentials that omit person, firm, and match effects confound pure inter-industry wage differentials with industry-average person effects and match effects.

3.2.2 Omitted Match Effects

We now consider the case where wages are determined according to equation (3) but the estimated equation excludes match effects only, i.e., the Abowd et al. (1999) person and firm effects model. When match effects are omitted, the estimated parameters β^{**}, θ_i^{**}, and ψ_j^{**} are biased. Specifically, least squares estimates of the mis-specified model satisfy

$$\begin{aligned} E\left[\beta^{**}\right] &= \beta + \left(X'M_{[D\ F]}X\right)^{-1} X'M_{[D\ F]}G\phi \\ E\left[\theta^{**}\right] &= \theta + \left(D'M_{[X\ F]}D\right)^{-} D'M_{[X\ F]}G\phi \\ E\left[\psi^{**}\right] &= \psi + \left(F'M_{[X\ D]}F\right)^{-} F'M_{[X\ D]}G\phi \end{aligned} \tag{16}$$

where A^- denotes a generalized inverse of A.[3]

In expectation, the estimated returns to time-varying observable characteristics, β^{**}, equal the

[3]For simplicity, we assume X has full column rank k. However D, F, and G do not, in general, have full column rank without additional identifying restrictions, e.g., exclusion of one column per connected group of workers and firms. See Searle (1987, Ch. 5) for a general statistical discussion of connected data, or Abowd et al. (2002) for a discussion in the context of linked employer-employee data.

true vector of returns plus an employment-duration weighted average of the omitted match effects, conditional on the design of the person and firm effects. The sign and magnitude of the bias depends on the conditional covariance between X and G, given D and F.

There is a simple relationship between D, F, and G that implies estimated person and firm effects are biased when match effects are omitted, except in the special case where $\phi_{ij} = 0$ for all matches. This is quite intuitive: the design of the person effects contains information on worker identities ("who you are"), the design of the firm effects contains information on firm identities ("where you work"), and the design of match effects contains information on match identities ("who you are and where you work"). Consequently, the design of the match effects is always correlated with the design of person and firm effects.[4] Hence if match effects are nonzero, estimated person and firm effects are always biased by their omission. This can be seen in the bias expression (16). Note the bias arises because of correlation between D, F, and G, and not because of correlation between the match effects and the person and firm effects themselves.

The expected value of the estimated person effects in the mis-specified model, θ^{**}, equal the true vector of person effects plus the employment-duration-weighted average of omitted match effects, conditional on observable time-varying characteristics and the design of the firm effects. In the simplest case where X and F are orthogonal to D and G, so that $D'M_{[X\ F]}D = D'D$ and $D'M_{[X\ F]}G = D'G$, the omitted variable bias is a vector of employment duration-weighted average match effects, so that

$$E\left[\theta_i^{**}\right] = \theta_i + \frac{1}{T_i}\sum_{t=t_i^1}^{T_i} \phi_{i\mathcal{J}(i,t)}. \tag{17}$$

The omitted variable bias in ψ^{**} is likewise the employment-duration-weighted average of omitted match effects, conditional on X and D. If X and D are orthogonal to F and G, so that $F'M_{[X\ D]}F = F'F$ and $F'M_{[X\ D]}G = F'G$, the omitted variable bias in ψ^{**} is a vector of employment duration-weighted average match effects, so that

$$E\left[\psi_j^{**}\right] = \psi_j + \frac{1}{N_j}\sum_{i=1}^{N}\sum_{t=t_i^1}^{T_i} \mathbf{1}\left(\mathcal{J}\left(i,t\right) = j\right)\phi_{i\mathcal{J}(i,t)} \tag{18}$$

where N_j is the total number of observations on firm j. It follows from (18) that when match effects are omitted, pure inter-industry differentials are confounded with omitted match effects.

The preceding illustrates that if match effects are nonzero, the person and firm effects model attributes variation to person and firm effects that is actually due to omitted match effects. The returns to observable characteristics are also biased if workers with certain characteristics (e.g., more education or experience) sort into better employment matches than others. Consequently, estimated wage differentials are confounded with omitted match effects.

Finally, we note that when $\phi = 0$ (i.e., the data generating process does not include match

[4] Formally, the column of G corresponding to the match between worker i and firm j is the elementwise product of the i^{th} column of D and the j^{th} column of F.

effects) estimating equation (3) does not introduce bias into any of the estimated effects. We omit the proof of this claim, but it is available on request. The proof replicates the usual result that there is no bias from including irrelevant explanatory variables in regression models. There is, of course, an attendant loss of efficiency. This may be large if the number of worker-firm matches is large relative to the number of workers and firms.

3.3 Identification and Estimation

We now discuss identification and estimation of the match effects model. We assume throughout that errors have zero conditional mean and are spherical:

$$E\left[\varepsilon_{ijt}|i,j,t,x_{ijt}\right] = 0 \tag{19}$$

$$E\left[\varepsilon_{ijt}\varepsilon_{mns}|i,j,t,m,n,s,x_{ijt},x_{mns}\right] = \begin{cases} \sigma_\varepsilon^2 \text{ for } i=m,\ j=n,\ t=s \\ 0 \text{ otherwise.} \end{cases} \tag{20}$$

Assumption (20) can be relaxed, but doing so complicates estimation.[5]

Assumptions (19) and (20) are standard for linear regression models. However, they are insufficient to identify all parameters of the match effects model. The simplest way to see this is to consider estimating the model in two steps. Applying standard results for partitioned regression, the least squares estimator of β is:

$$\hat{\beta} = \left(X' M_{[D\ F\ G]} X\right)^{-1} X' M_{[D\ F\ G]} y. \tag{21}$$

Some algebra verifies that $M_{[D\ F\ G]}$ takes deviations from match-specific means.[6] So we can easily recover $\hat{\beta}$ from the regression of y_{ijt} on x_{ijt}, both in deviations from match-specific means. Note this simple method to recover the least squares estimate of β is only valid when the model includes match effects.[7]

Having estimated β, the second step is to decompose $y - X\hat{\beta}$ into person effects, firm effects, match effects, and residuals. Intuitively, the identification problem here is to distinguish "good" workers and firms (i.e., those with larger person/firm effects) from "lucky" ones (i.e., those with large match effects).[8] In principle, we can estimate the person, firm, and match effects by fixed or random effects methods. Woodcock (2006) provides a comprehensive discussion of various ap-

[5] See Woodcock (2005a) for an application of the person and firm effects model with non-spherical errors.

[6] $M_{[D\ F\ G]}$ projects onto the column null space of $[D\ F\ G]$. It is a block diagonal matrix with N^* rows and columns, where the M diagonal blocks correspond to each of the M worker-firm matches. The ij^{th} diagonal block is zero if worker i never works at firm j. Otherwise, it is the $T_{ij} \times T_{ij}$ submatrix $M^{ij}_{[D\ F\ G]} = I_{T_{ij}} - \frac{1}{T_{ij}} \iota_{T_{ij}} \iota'_{T_{ij}}$ where $T_{ij} = \sum_{t=t_i^1}^{T_i} \mathbf{1}\left(\mathcal{J}\left(i,t\right)=j\right)$ is the duration of the match between worker i and firm j; I_A is the identity matrix of order A; and ι_A is an $A \times 1$ vector of ones. Each $M^{ij}_{[D\ F\ G]}$ takes deviations from means in the match between worker i and firm j.

[7] That is, whereas $M_{[D\ F\ G]}$ takes deviations from match match-specific means, $M_{[D\ F]}$ does not.

[8] Under a human capital interpretation of (3), person and firm effects reflect the returns to general and firm-specific human capital, respectively. Match effects likewise reflect the returns to match-specific human capital. This could reflect luck (e.g., finding a good match) or production complementarities.

proaches. We briefly summarize the main points here.

Fixed effect estimators are popular among economists, primarily because they are perceived to embody fewer assumptions about the relationship between observables and unobservables than mixed (random) effect estimators. Unfortunately, they are poorly suited to estimating the match effects model. In fact, they present a fundamental identification problem here, because the fixed effect formulation of the match effects model is over-parameterized. There are $N + J + M + 1$ person effects, firm effects, match effects, and a constant term to estimate, but only M worker-firm matches ("cell means") from which to estimate them.[9] Alternately put, the only estimable functions of $\theta_i, \psi_j, \phi_{ij}$ and μ in equation (3) are the M population cell means $\mu_{ij\cdot} = \mu + \theta_i + \psi_j + \phi_{ij}$ (Searle, 1987 p. 331).[10] That is, the cell means are always identified, but decompositions of the cell means into the various effects require additional (ancillary) assumptions. By their very nature, however, such ancillary assumptions are arbitrary and untestable, and parameter estimates are not invariant to the choice of identifying assumptions.

Because of these identification problems, we take a different approach here. We treat the unobserved components α_i, ψ_j, and ϕ_{ij} as random effects. Woodcock (2006) calls this a hybrid mixed effects estimator. It differs from a traditional mixed (random) effect estimator because β is estimated under the minimal identifying assumptions (19) and (20) required for least squares. As a consequence, the hybrid mixed effect estimator does not impose the usual assumption that the random effects have zero conditional mean given x_{ijt}. The identifying assumptions are:[11]

$$E\left[\alpha_i | u_i\right] = E\left[\psi_j | u_i\right] = E\left[\phi_{ij} | u_i\right] = 0 \tag{22}$$

$$Cov\left[\begin{array}{c|c} \alpha_i & \\ \psi_j & u_i \\ \phi_{ij} & \end{array}\right] = \left[\begin{array}{ccc} \sigma_\alpha^2 & 0 & 0 \\ 0 & \sigma_\psi^2 & 0 \\ 0 & 0 & \sigma_\phi^2 \end{array}\right]. \tag{23}$$

These are weaker than the identifying assumptions of a traditional mixed (random) effect model, for which (22) and (23) would also condition on x_{ijt}.

Estimating the hybrid mixed model in fact proceeds in three steps. In the first step, we estimate

[9] The term "cell mean" is adopted from the statisical literature on estimation of the two-way crossed classification with interaction, of which the match effects model is an example. It arises from representing the data as a table with rows defined by the levels of i (workers), and columns defined by the levels of j (firms). The entry in row i and column j is the mean earnings of worker i at firm j, or "cell mean."

[10] In practice, there are only M estimable functions of the person, firm, and match effects, the overall constant, and a set of "group means" for groups of connected observations in the sample. When the sample consists of $\mathcal{G}$ connected groups of observations, the number of estimable functions of the other effects is reduced by a corresponding amount. We abstract from these considerations in the main text, and presume the sample consists of a single connected group. See Abowd et al. (2002) for further discussion of connectedness, including a graph-theoretic algorithm for determining connected groups of observations.

[11] Equation (22) may appear restrictive. However note that a fixed effect estimator would also impose orthogonality between observable and unobserved components of the person effect. The assumption in (23) that random effects are uncorrelated with one another is standard. It is technically feasible though computationally burdensome to allow non-zero correlation between random effects, e.g., between α_i and ψ_j. For an example, see Conway and Houtenville (2001) for a model of migration flows with cross-correlated random effects. We leave such considerations for future research.

β by least squares, so that $\hat{\beta}$ is given by the "within" estimator (21). In the second step we estimate the variance of the random effects $(\sigma^2_\alpha, \sigma^2_\psi, \sigma^2_\phi)$ and errors (σ^2_ε) by Restricted Maximum Likelihood (REML) on $y - X\hat{\beta}$.[12] Finally, conditional on $\hat{\beta}$ and the REML estimates, we solve the Henderson et al. (1959) mixed model equations:

$$\begin{bmatrix} U'U & U'D & U'F & U'G \\ D'U & D'D + \left(\tilde{\sigma}^2_\varepsilon/\tilde{\sigma}^2_\alpha\right) I_N & D'F & D'G \\ F'U & F'D & F'F + \left(\tilde{\sigma}^2_\varepsilon/\tilde{\sigma}^2_\psi\right) I_J & F'G \\ G'U & G'D & G'F & G'G + \left(\tilde{\sigma}^2_\varepsilon/\tilde{\sigma}^2_\phi\right) I_M \end{bmatrix} \begin{bmatrix} \tilde{\eta} \\ \tilde{\alpha} \\ \tilde{\psi} \\ \tilde{\phi} \end{bmatrix} = \begin{bmatrix} U' \\ D' \\ F' \\ G' \end{bmatrix} \left(y - X\hat{\beta}\right). \tag{24}$$

for estimates of the remaining parameters: $\tilde{\eta}, \tilde{\alpha}, \tilde{\psi}$, and $\tilde{\phi}$.

The hybrid mixed effect estimator has the following properties. $\hat{\beta}$ is consistent and the BLUE of β given the minimal assumptions (19) and (20) on ε. Given the additional stochastic assumptions (22) and (23), $\tilde{\eta}$ is consistent and the BLUE of η, and $\left(\tilde{\alpha}, \tilde{\psi}, \tilde{\phi}\right)$ are Best Linear Unbiased Predictors (BLUPs) of the random effects.[13] Furthermore, we see from (24) that the least squares estimator is a special case as $\left(\tilde{\sigma}^2_\varepsilon/\tilde{\sigma}^2_\alpha\right) \to 0$, $\left(\tilde{\sigma}^2_\varepsilon/\tilde{\sigma}^2_\psi\right) \to 0$, and $\left(\tilde{\sigma}^2_\varepsilon/\tilde{\sigma}^2_\phi\right) \to 0$.

Estimating the person and firm effects model is more straightforward. This is because the collection of M restrictions $\phi_{ij} = 0$ is generally sufficient to identify the least squares estimator of all remaining model parameters.[14] Here, the primary hindrance to estimation is computational: directly solving the least squares normal equations implies inverting a cross-products matrix with $k + N + J + 1$ rows and columns – typically a very large number. Abowd et al. (2002) present a conjugate gradient algorithm to directly minimize the sum of squared residuals without inverting this cross-products matrix. We use this algorithm to compute least squares estimates of the person and firm effects model.

4 Data

Identifying the person, firm, and match effects requires longitudinal data on employers and employees. We use data from the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) database. These data span thirty-seven states that represent the majority of US employ-

[12] REML is often described as maximizing the part of the likelihood that is invariant to the values of the fixed effects. It is akin to partitioned regression. It is maximum likelihood on linear combinations of y under normality. The linear combinations $K'y$ are chosen so that $K'(X\beta + U\eta) = 0$ for all values of β and η, which implies $K'[X \;\; U] = 0$. Thus K' projects onto the column null space of $[X \;\; U]$ and is of the form $K' = C'M_{[X \;\; U]}$ for arbitrary C'. The REML estimator has many attractive properties: estimates are invariant to the value of (β, η), consistent, asymptotically normal, and asymptotically efficient in the Cramer-Rao sense. We compute REML estimates using the Average Information algorithm of Gilmour et al. (1995).

[13] BLUPs are *best* in the sense of minimizing the mean square error of prediction among linear unbiased estimators, and *unbiased* in the sense $E[\tilde{\alpha}] = E[\alpha]$, $E\left[\tilde{\psi}\right] = E[\psi]$, and $E\left[\tilde{\phi}\right] = E[\phi]$. See Robinson (1991).

[14] In a balanced panel of full-time full-year workers, it is not possible to separately identify person effects, time effects, and experience effects using a fixed effect estimator except through functional form restrictions. This is well known, and is true whether or not the model includes firm and match effects. In the application of Section 5, all three effects are identified because the panel is unbalanced and because of work history interruptions.

ment. We use data from two participating states (whose identity is confidential) that are broadly representative of the LEHD database.[15]

The LEHD data are administrative, constructed from Unemployment Insurance (UI) system employment reports. These are collected by each state's Employment Security agency to manage the unemployment compensation program. Employers are required to report total payments to all employees on a quarterly basis. These payments (earnings) include gross wages and salary, bonuses, stock options, tips and gratuities, and the value of meals and lodging when these are supplied (Bureau of Labor Statistics (1997, p. 44)).

The coverage of UI data varies slightly from state to state, though the Bureau of Labor Statistics (1997, p. 42) claims that UI coverage is "broad and basically comparable from state to state" and that "over 96 percent of total wage and salary civilian jobs" were covered in 1994. See Abowd et al. (2006) for further details. With the UI employment records as its frame, the LEHD data comprise the universe of employment at firms required to file UI reports.

Individuals and firms are uniquely identified in the data. The UI employment records contain only limited information: identifiers and earnings. The LEHD database integrates these with internal Census Bureau data to obtain demographic and firm characteristics, including sex, race, date of birth, industry, and geography.

Though the underlying data are quarterly, they are aggregated to the annual level for estimation. The full sample consists of over 49 million annualized employment records on full-time workers between 25 and 65 years of age who were employed at private-sector non-agricultural firms between 1990 and 1999.

Solving the mixed model equations (24) is computationally intensive. All our estimates are therefore based on a subsample. Sampling from linked employer-employee data is nontrivial because employment histories must be sufficiently connected to precisely estimate the person, firm, and match effects. Thus we take a ten percent subsample of individuals employed in 1997 using the dense sampling algorithm of Woodcock (2005b). This algorithm ensures that each worker is connected to at least five others by a common employer, but is otherwise representative of the population of individuals employed in 1997. That is, all individuals employed in 1997 have an equal probability of being sampled.[16] The dense subsample consists of the full work history of each sampled individual.

Table 1 presents characteristics of the samples (see Appendix Table 1 for variable definitions). The sample of individuals employed in 1997 is largely representative of the full sample of observations. Differences indicate that individuals employed in 1997 have a slightly stronger labor force attachment than the sample of individuals ever employed between 1990 and 1999: males are slightly

[15] As discussed below, computational complexities dictate that we restrict our analysis to a subset of observations. In a small sample drawn from all thirty-seven states, work histories are not sufficiently connected to estimate the person, firm, and match effects precisely. Hence we focus on two representative states instead.

[16] The dense subsample is constructed by sampling firms with probabilities proportional to employment in a reference period (1997), and then sampling workers within firms with probabilities inversely proportional to firm employment. A minimum of 5 employees are sampled from each firm. By careful choice of sampling probabilities, all workers employed in the reference period have an equal probability of being sampled, and each sampled worker is connected to at least 5 others by a common employer.

over-represented, as are individuals with higher educational attainment and individuals who work four full quarters in an average calendar year. The dense subsample has characteristics virtually identical to the sample of all individuals employed in 1997.

5 Results

Table 2 presents the estimated variance of log earnings components. These are given for three different specifications. Column 1 reports estimates for a baseline specification that includes observable characteristics only: sex, race, education (5 categories), a quartic in experience; and indicators for the number of quarters worked in the calendar year, industry (SIC Major Division), and year. All characteristics other than industry are interacted with sex. We do not interact industry with sex because this allows the most straightforward comparison with specifications that include firm effects.[17] Column 2 reports estimates for the person and firm effects model, and column 3 gives estimates for the match effects model.

Comparing estimates from the three specifications, we see that controlling for additional components of unobserved heterogeneity increases the proportion of variation explained by the model and reduces the proportion attributed to observable characteristics. This is not surprising. Person effects exhibit the greatest variation (0.291 and 0.198 squared log points in the person and firm effects model and match effects model, respectively). The match effects model estimates greater dispersion in firm effects than the person and firm effects model does (0.102 versus 0.080 squared log points). There is considerable variation in match effects also (0.079 squared log points) – more than in the returns to all observable characteristics (0.056 squared log points in the match effects model). Estimates from the match effects model imply that a one standard deviation increase in the person effect increases earnings by 0.44 log points, a one standard deviation increase in the firm effect increases earnings by 0.32 log points, and a one standard deviation increase in the match effect increases earnings by 0.28 log points. Hence all three effects contribute considerable variation to log earnings.

Column 3 of Table 2 also reports the p-value of a formal test for the presence of match effects. Since we treat match effects as random, the null of match effects is $H_0 : \sigma^2_\phi = 0$. Because the null hypothesis places σ^2_ϕ on the boundary of the parameter space, the likelihood ratio test statistic has a non-standard asymptotic distribution. Stram and Lee (1994) show its asymptotic distribution is a 50:50 mixture of a χ^2_0 and a χ^2_1. We easily reject the null of no match effects at conventional significance levels.[18]

[17] Firm effects are common to all employees and therefore do not vary by sex. Since pure industry effects are the aggregation of firm effects (Section 3.1), comparing estimated industry effects between specifications with and without firm effects is most direct when industry effects are the same for both sexes.

[18] The test statistic exceeds 35,000. An alternate test is also available based on a fixed effect estimator. Although fixed effect estimates of the person, firm, and match effects are not separately identified without ancillary assumptions, their sum is always identified. Hence we can compute fixed effect residuals for models with and without match effects and test the null hypothesis $H_0 : \phi_{ij} = 0$ for each i, j pair in the data. This is a test of $M - N - J = 323,477$ linear restrictions. We easily reject the null of no match effects by this test also (the Wald statistic exceeds 1.4 million).

We use these three specifications to illustrate the estimation of wage differentials in the presence and absence of person, firm, and match effects. We consider two often investigated wage differentials: inter-industry differentials and the male-female differential. In each case, we decompose wage differentials following equations (7)-(9) using parameter estimates, including BLUPs in the case of the match effects model, from these three specifications.

5.1 Inter-Industry Differentials

Table 3 presents decompositions of inter-industry earnings differentials for SIC Major Divisions. Most studies of inter-industry differentials are based on more detailed industrial definitions than this. However, our analysis of aggregated inter-industry differentials is sufficient to illustrate the consequences of omitted person, firm, and/or match effects.[19]

Column 1 in panel A gives the raw inter-industry log earnings differentials: the difference between average log earnings in each industry and the overall mean of log earnings. There is considerable earnings variation between industries: the weighted standard deviation (WSD) of raw inter-industry differentials is 0.131 log points.[20]

Column 2 of panel A reports regression-adjusted inter-industry earnings differentials for our baseline specification that excludes person, firm, and match effects. The estimates are normalized to have zero mean when weighted by employment shares. This normalization makes the regression-adjusted differentials directly comparable to raw differentials and to our estimated pure inter-industry differentials (estimated firm effects are also normalized to have zero mean). In general, the regression-adjusted differentials are smaller in absolute value than the raw differentials, suggesting that observable characteristics explain much of the observed differences in log earnings between industries. The WSD of regression-adjusted differentials is 0.105 log points.[21] This implies approximately 20 percent of the weighted variance (WSD squared) of raw differentials is explained by inter-industry differences in observable characteristics.

Panels B and C decompose the raw inter-industry differentials according to (7). The decomposition in panel B is based on the person and firm effects model, and panel C is based on the match effects model. All components are normalized to have zero mean in the estimation sample. They can therefore be interpreted as log point deviations (or approximately as percentage deviations) from the overall mean of earnings.

Both the person and firm effects model and the match effects model attribute approximately 19 percent of the weighted variance of raw differentials to inter-industry differences in observable

[19] Most authors study disaggregated indutries because estimates may be subject to bias if compensation policies differ between sub-industries within the aggregates. Pure industry effects are not subject to aggregation bias because they are based on firm-level estimates (firm effects).

[20] The weighted standard deviation of raw differentials is $WSD = \sqrt{\sum_k s_k \left(\bar{y}_k - \bar{y}\right)^2}$ where k indexes industries, s_k is industry k's employment share, $\bar{y}_k$ is average log earnings in industry k, and $\bar{y}$ is the overall mean of log earnings.

[21] The WSD of regression-adjusted differentials is $WSD = \sqrt{\sum_k s_k \left(\hat{\delta}_k - \bar{\hat{\delta}}\right)^2}$ where $\hat{\delta}_k$ is the regression-adjusted differential in industry k, and $\bar{\hat{\delta}}$ is the employment-share weighted average of regression-adjusted differentials.

characteristics ($x'_{ijt}\beta + u'_i\eta$, column 1).[22] In all industries except FIRE, the observable component has the same sign as the raw differential but is smaller in magnitude.

Unobserved personal characteristics (α_i, column 2) and observable characteristics tend to make opposing contributions to the raw differentials. Estimates of the unobserved personal component from the person and firm effects model and the match effects model generally have the same sign, but estimates that exclude match effects are larger in absolute value. Consequently, the person and firm effects model attributes about 15 percent of the weighted variance of raw differentials to inter-industry differences in unobserved personal characteristics, versus 8 percent for the match effects model.

Column 3 presents the component due to all personal characteristics, both observed and unobserved ($x'_{ijt}\beta + \theta_i$, less time effects). The person and firm effects model attributes 34 percent of the weighted variance of raw differentials to inter-industry differences in personal characteristics, versus 28 percent for the match effects model.

Column 4 presents the component due to firm effects, i.e., the pure inter-industry earnings differentials. The pure inter-industry differentials contribute the lion's share of the weighted variance of the raw differentials: 64 percent in the person and firm effects model and 72 percent in the match effects model. In all industries, estimates based on the person and firm effects model and the match effects model have the same sign. Again, estimates based on the person and firm effects model are generally larger in absolute value than those based on the match effects model.

Although estimates based on the person and firm effects model and the match effects model are similar in many respects, there are some striking differences. Notably, the person and firm effects model predicts negative sorting of workers across industries: the correlation between the component due to personal characteristics (column 3) and firm effects (column 4) is negative (-0.10). However, the match effects model overturns this result: here the correlation between industry-average personal characteristics and firm effects is strongly positive (0.60). As a consequence, the two specifications give very different interpretations of the source of inter-industry earnings differences. For instance, the person and firm effects model suggests the large raw differential in the mining industry (0.194 log points) is the result of "low-wage" workers (the component due to personal characteristics is -0.135) employed in very "high-wage" firms (the component due to firm effects is 0.352). The match effects model, in contrast, attributes the differential to a combination of high-wage workers and high-wage firms, since both components are positive. This difference illustrates that ignoring match effects can result in misleading inferences about the nature of inter-industry earnings differentials – despite the fact there is negligible inter-industry variation in average match effects (column 5).[23]

[22] We decompose the weighted variance of raw differentials as follows: $\sum_k s_k (\bar{y}_k - \bar{y})^2 = \sum_k s_k (\bar{y}_k - \bar{y}) \left((\bar{x}_k - \bar{x}) \hat{\beta} + \bar{\theta}_k + \bar{\psi}_k + \bar{\phi}_k + \bar{e}_k \right)$ where $\bar{\theta}_k, \bar{\psi}_k$, and $\bar{\phi}_k$ are normalized to have zero weighted mean in the estimation sample. Dividing both sides of the equality by $\sum_k s_k (\bar{y}_k - \bar{y})^2$ gives a proportionate decomposition of the weighted variance into components that reflect the contribution of inter-industry differences in observable characteristics, person effects, firm effects, match effects, and residuals.

[23] Differences between the person and firm effects model and the match effects model appear to be the consequence of controlling for match effects, rather than differences between fixed and random effects estimation. That is, random

Finally, column 6 presents the component due to all unobservables: $\alpha_i + \psi_j$ in the person and firm effects model, and $\alpha_i + \psi_j + \phi_{ij}$ in the match effects model. As noted in Section 3.2.1, regression-adjusted differentials that do not control for unobserved worker, firm, and/or match characteristics are simply the duration-weighted average of the omitted effects, adjusted for X. Consequently, entries in column 6 correspond very closely to the regression-adjusted differentials in column 2 of panel A.[24]

5.2 The Male-Female Differential

We now consider a detailed decomposition of the male-female earnings differential. This is presented in Table 4. Following equations (7)-(9), we decompose the raw difference between the average earnings of women and men (-0.36 log points) into the component due to differences in observable characteristics, the component due to differences in returns to observable characteristics, and components due to unobservables.

The baseline specification (column 1) controls for observable characteristics only. Estimates in this column are very similar to others' findings, e.g., Altonji and Blank (1999). Columns 2 and 3 present the decomposition for the person and firm effects model and the match effects model, respectively. All three specifications agree that differences in observable characteristics contribute little (0.02 log points or less) to the raw differential. This is not surprising, given the minimal differences between male and female characteristics in Table 1.

There is considerable disagreement between specifications, however, regarding the contribution of differences in returns. This disagreement is primarily manifested in the estimated returns to experience and education. The baseline specification attributes the vast majority of the raw wage differential (-0.264 log points) to differences in returns to observable characteristics. Of this, lower returns to experience are the largest component (-0.307 log points), and lower returns to education widen the differential by a further -0.021 log points. In contrast, the person and firm effects model attributes very little of the differential to differences in returns to observable characteristics. This is due to a much smaller differential in the returns to experience (-0.172 log points) and an offsetting positive differential in the returns to education (0.106 log points). The match effects model estimates a similar differential in the returns to experience (-0.155 log points), but a smaller positive differential in the returns to education (0.024 log points). Differences between these two specifications reflect Woodcock's (2006) finding that the person and firm effects model over-estimates the returns to education and experience: more educated and more experienced workers sort into better worker-firm matches on average, and the returns to sorting are attributed to education and experience when match effects are omitted.

The person and firm effects model and the match effects model both attribute a sizable component of the overall earnings differential to employment at lower-paying firms. In the person and firm

effects estimates of the person and firm effects model are very similar to the fixed effect estimates presented here. These are available on request.

[24] They are not exactly equal because of covariation between unobservables and X.

effects specification, employment at firms with lower average firm effects reduces female earnings by 0.069 log points compared to males. This is nearly 20 percent of the raw differential. Controlling for unobserved match heterogeneity reduces this component by almost half.

The suggestion that a sizable component of the male-female earnings differential is due to employment in lower-paying firms is intriguing. To better understand this finding, we further decompose the component due to firm effects into a component that reflects differences in male-female sorting across industries, and a component that reflects differential sorting across firms within industries:

$$\bar{\psi}_f - \bar{\psi}_m = \sum_{k=1}^{K} \left(s_f^k - s_m^k \right) \bar{\psi}_m^k + \sum_{k=1}^{K} s_f^k \left(\bar{\psi}_f^k - \bar{\psi}_m^k \right)$$

where $\bar{\psi}_f$ and $\bar{\psi}_m$ are the average firm effects of females and males, respectively; $k = 1, ..., K$ indexes industries (SIC Major Division); s_f^k and s_m^k are the employment shares of females and males, respectively, in industry k; and $\bar{\psi}_f^k$ and $\bar{\psi}_m^k$ are the average firm effects of females and males, respectively, in industry k. The first term measures the returns to differential inter-industry sorting, evaluated at the male industry-average firm effects (i.e., the male pure industry effects). The second term measures the returns to differential intra-industry sorting between firms, evaluated at the female employment shares.

Of the -0.069 log point earnings differential attributed to employment in lower-paying firms, the person and firm effects model attributes about equal proportions to employment in lower-paying industries and employment in lower-paying firms within industries. The match effects model, on the other hand, attributes only -0.011 log points to sorting into lower-paying industries, versus -0.026 log points to sorting into lower-paying firms within industries. However, both specifications agree that the male-female earnings differential is partly due to industrial segregation (inter-industry sorting), and partly due to employment at lower-paying firms within industries.

Finally, a large component of the earnings differential remains unexplained in all specifications. This is the component attributed to differences between male and female regression intercepts. In the baseline model, this measures the differential for the reference category of all categorical variables (whites with less than high school education, who worked four full quarters in 1990). The male and female means of α_i and ϕ_{ij} are not separately identified from the intercept, so these too are reflected in the difference between male and female intercepts in the person and firm effects model and the match effects model. Large differences between the unexplained component in our baseline specification and the other specifications suggest unobserved personal and match heterogeneity are important contributors to the raw male-female differential.

6 Conclusion

The empirical application demonstrates that decompositions of wage differentials that do not control for person, firm, and match effects can be misleading. It is not sufficient to control for person and firm effects only, because the estimated returns to observable characteristics and the estimated

person and firm effects are biased by the omission of match effects. This is despite the fact we found no substantial direct contribution of match effects to inter-industry or male-female earnings differentials.

Our analyses of inter-industry and male-female log earnings differentials suggest that how workers sort into firms and industries is an important component of observed earnings differentials. However, our application only considered highly aggregated industrial definitions. Because these may be composed of fairly heterogeneous sub-industries, a detailed investigation of less aggregated inter-industry and intra-industry differentials is warranted.

References

Abowd, J. M., R. H. Creecy, and F. Kramarz (2002). Computing person and firm effects using linked longitudinal employer-employee data. Mimeo.

Abowd, J. M., F. Kramarz, P. Lengermann, and S. Roux (2005). Persistent inter-industry wage differences: Rent sharing and opportunity costs. Mimeo.

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*(2), 251–334.

Abowd, J. M., F. Kramarz, and S. D. Woodcock (forthcoming). Econometric analyses of linked employer-employee data. In L. Matyas and P. Syvestre (Eds.), *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory*. Kluwer. 3rd. ed.

Abowd, J. M., B. E. Stephens, L. Vilhuber, F. Andersson, K. L. McKinney, M. Roemer, and S. D. Woodcock (2006). The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators. LEHD Technical Paper TP-2006-01, U.S. Census Bureau.

Altonji, J. G. and R. M. Blank (1999). Race and gender in the labor market. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics Volume 3C*, pp. 3143–3259. Amsterdam: Elsevier Science.

Blau, F. D. and L. M. Kahn (2003). Understanding international differences in the gender pay gap. *Journal of Labor Economics 21*(1), 106–144.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural variables. *Journal of Human Resources 8*, 436–455.

Bureau of Labor Statistics (1997). *BLS Handbook of Methods*. U.S. Department of Labor.

Conway, K. S. and A. J. Houtenville (2001). Elderly migration flows and state government policy - evidence from the 1990 Census migration flows. *National Tax Journal LIV*(1), 103–123.

Gilmour, A. R., R. Thompson, and B. R. Cullis (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics 51*, 1440–1450.

Goux, D. and E. Maurin (1999). Persistence of interindustry wage differentials: A reexamination using matched worker-firm panel data. *Journal of Labor Economics 17*(3), 492–533.

Groshen, E. L. (1991). Sources of intra-industry wage dispersion: How much do employers matter? *Quarterly Journal of Economics 106*(3), 869–884.

Henderson, C., O. Kempthorne, S. Searle, and C. V. Krosigk (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*(2), 192–218.

Katz, L. F. and L. H. Summers (1989). Industry rents: Evidence and implications. *Brookings Papers on Economic Activity: Microeconomics*, 209–290.

Krueger, A. B. and L. H. Summers (1988). Efficiency wages and the inter-industry wage structure. *Econometrica 56*(2), 259–293.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review 14*, 693–709.

Oi, W. Y. and T. L. Idson (1999). Firm size and wages. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics Volume 3B*, pp. 2165–2214. Amsterdam: Elsevier Science.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science 6*(1), 15–32.

Searle, S. R. (1987). *Linear Models for Unbalanced Data*. New York: John Wiley and Sons.

Stram, D. O. and J. W. Lee (1994). Variance component testing in the longitudinal mixed effects model. *Biometrics 50*, 1171–1177.

Woodcock, S. D. (2005a). Heterogeneity and learning in labor markets. Mimeo.

Woodcock, S. D. (2005b). Sampling connected histories from longitudinal linked data. Mimeo.

Woodcock, S. D. (2006). Match effects. Mimeo.

TABLE 1
SUMMARY STATISTICS
(Sample Proportions Unless Otherwise Stated)

	FULL SAMPLE		ALL INDIVIDUALS EMPLOYED IN 1997		TEN PERCENT DENSE SUBSAMPLE	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Demographic Characteristics						
Male	.56	.50	.58	.49	.57	.50
Age (Years)	40.6	10.2	40.3	9.6	40.3	9.6
Men						
Nonwhite	.21	.57	.20	.55	.20	.56
Race Missing	.04	.25	.03	.24	.03	.24
Less Than High School	.12	.45	.11	.43	.11	.43
High School	.30	.67	.30	.65	.29	.66
Some College	.23	.60	.23	.59	.23	.59
Associate or Bachelor's Degree	.25	.62	.25	.61	.25	.62
Graduate or Professional Degree	.10	.42	.11	.42	.11	.42
Women						
Nonwhite	.24	.69	.24	.71	.25	.72
Race Missing	.02	.22	.02	.22	.02	.22
Less Than High School	.09	.45	.09	.45	.09	.44
High School	.31	.78	.30	.79	.30	.78
Some College	.25	.71	.25	.73	.25	.72
Associate or Bachelor's Degree	.26	.72	.27	.75	.27	.75
Graduate or Professional Degree	.08	.42	.09	.44	.09	.44
Work History Characteristics						
Real Annualized Earnings (1990 Dollars)	41,107	38,849	43,183	39,324	43,528	38,782
Men						
Labor Market Experience (Years)	11.8	13.1	11.9	12.7	11.8	12.7
Worked 0 Full Quarters in Calendar Year	.08	.36	.06	.32	.06	.32
Worked 1 Full Quarter in Calendar Year	.15	.49	.12	.44	.12	.44
Worked 2 Full Quarters in Calendar Year	.13	.47	.12	.44	.12	.44
Worked 3 Full Quarters in Calendar Year	.14	.48	.13	.46	.14	.47
Worked 4 Full Quarters in Calendar Year	.50	.80	.56	.81	.57	.00
Women						
Labor Market Experience (Years)	9.5	13.0	9.0	12.5	9.2	12.6
Worked 0 Full Quarters in Calendar Year	.07	.39	.06	.36	.05	.35
Worked 1 Full Quarter in Calendar Year	.14	.54	.11	.50	.11	.50
Worked 2 Full Quarters in Calendar Year	.13	.53	.12	.51	.11	.50
Worked 3 Full Quarters in Calendar Year	.14	.55	.13	.54	.13	.54
Worked 4 Full Quarters in Calendar Year	.52	.96	.58	1.02	.59	1.01
Year						
1990	.09	.29	.07	.26	.07	.26
1991	.09	.29	.08	.27	.08	.27
1992	.09	.29	.08	.27	.08	.28
1993	.10	.29	.09	.28	.09	.28
1994	.10	.30	.10	.29	.10	.29
1995	.10	.30	.10	.31	.10	.31
1996	.10	.31	.11	.32	.11	.32
1997	.11	.31	.14	.35	.14	.34
1998	.11	.31	.12	.32	.12	.32
1999	.11	.31	.11	.31	.11	.31
Number of Observations (N*)	49,291,205		37,688,492		3,652,544	
Number of Workers (N)	9,272,529		5,235,887		503,179	
Number of Firms (J)	573,307		476,745		121,227	
Number of Worker-Firm Matches (M)	15,309,134		9,889,502		947,883	
Number of Connected Groups	84,748		46,829		1,460	

TABLE 2
VARIANCE OF ESTIMATED COMPONENTS OF LOG EARNINGS

	OBSERVABLE CHARACTERISTICS ONLY* (1)	PERSON AND FIRM EFFECTS MODEL* (2)	MATCH EFFECTS MODEL† (3)
Variance of Log Real Annualized Earnings (y)	.410	.410	.410
Variance of Time-Varying Covariates (Xβ)	.068	.030	.017
Variance of Pure Person Effect (θ)		.291	.198
Time-Invariant Covariates (Uη)	.065	.044	.039
Unobserved Heterogeneity (α)		.247	.159
Variance of Firm Effect (ψ)	.010‡	.080	.102
Variance of Match Effect (ϕ)			.079
Error Variance (ε)	.310	.055	.036
H_0: No Match Effects (p-value)			< .00001
R^2	.243	.889	.933
Model Degrees of Freedom	3,652,503	3,029,559	3,652,500

Source: Author's calculations based on LEHD data.

* Values are sample variances of the estimated effects. The estimated error variance is corrected for degrees of freedom.

† Values in rows labeled y, Xβ, Uη are sample variances. Values in rows labeled θ, ψ, ϕ, ε are REML estimates of variance components.

‡ Sample variance of estimated industry effects.

TABLE 3
DECOMPOSITION OF INTER-INDUSTRY LOG EARNINGS DIFFERENTIALS

A. BASELINE MODEL

	Raw Differential (1)	Adjusted for Observables (2)
Mining	.194	.048
	(.005)	(.006)
Construction	.124	.035
	(.001)	(.001)
Manufacturing	.026	.000
	(.001)	(.001)
TCEGSS	.200	.135
	(.001)	(.001)
Wholesale Trade	.076	.030
	(.001)	(.001)
Retail Trade	-.328	-.272
	(.001)	(.001)
FIRE	.121	.155
	(.001)	(.001)
Services	-.041	-.009
	(.001)	(.001)
WSD	.131	.105

B. DECOMPOSITION OF RAW DIFFERENTIALS: PERSON AND FIRM EFFECTS MODEL

	All Observables (1)	Unobserved Person Effects (α) (2)	All Personal Characteristics (3)	Firm Effects (φ) (4)	All Unobservables (6)
Mining	.149	-.293	-.135	.352	.059
	(.000)	(.004)	(.000)	(.003)	(.005)
Construction	.086	-.025	.061	.064	.040
	(.000)	(.001)	(.000)	(.001)	(.001)
Manufacturing	.014	-.083	-.068	.095	.012
	(.000)	(.000)	(.000)	(.000)	(.001)
TCEGSS	.051	.022	.073	.126	.148
	(.000)	(.001)	(.000)	(.001)	(.001)
Wholesale Trade	.034	.001	.036	.039	.039
	(.000)	(.001)	(.000)	(.001)	(.001)
Retail Trade	-.051	-.078	-.130	-.191	-.269
	(.000)	(.001)	(.000)	(.001)	(.001)
FIRE	-.033	.109	.076	.044	.153
	(.000)	(.001)	(.000)	(.000)	(.001)
Services	-.019	.046	.025	-.069	-.023
	(.000)	(.000)	(.000)	(.000)	(.001)
Proportion of WSD[2]	.193	.149	.345	.644	.792

Source: Author's calculations based on LEHD data.
Notes: WSD is inter-industry weighted standard deviation of log earnings. TCEGSS abbreviates Transportation, Communications, Electric, Gas, and Sanitary Services. FIRE abbreviates Finance, Insurance, and Real Estate. All estimates are normalized to have zero mean when weighted by employment shares. Column 3 of panels B and C equals the sum of columns 1 and 2, less year effects. Column 6 equals the sum of columns 2, 4, and 5. Standard errors are in parentheses.

TABLE 3 CONTINUED
DECOMPOSITION OF INTER-INDUSTRY LOG EARNINGS DIFFERENTIALS

C. DECOMPOSITION OF RAW DIFFERENTIALS: MATCH EFFECTS MODEL

	All Observables (1)	Unobserved Person Effects (α) (2)	All Personal Characteristics (3)	Firm Effects (φ) (4)	Match Effects (ϕ) (5)	All Unobservables (6)
Mining	.134	-.063	.089	.120	-.011	.046
	(.002)	(.002)	(.002)	(.002)	(.001)	(.004)
Construction	.113	.011	.124	.000	-.001	.011
	(.001)	(.001)	(.001)	(.000)	(.000)	(.001)
Manufacturing	.011	-.023	-.010	.046	-.009	.015
	(.001)	(.000)	(.001)	(.000)	(.000)	(.001)
TCEGSS	.055	-.006	.051	.153	-.004	.144
	(.000)	(.001)	(.000)	(.001)	(.000)	(.001)
Wholesale Trade	.045	.011	.058	.016	.003	.030
	(.000)	(.001)	(.000)	(.000)	(.000)	(.001)
Retail Trade	-.034	-.043	-.078	-.244	-.004	-.291
	(.001)	(.001)	(.001)	(.000)	(.000)	(.001)
FIRE	-.030	.036	.007	.112	.002	.149
	(.001)	(.001)	(.001)	(.000)	(.000)	(.001)
Services	-.030	.015	-.017	-.032	.006	-.011
	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)
Proportion of WSD^2	.190	.082	.279	.724	.004	.803

Source: Author's calculations based on LEHD data.

Notes: WSD is inter-industry weighted standard deviation of log earnings. TCEGSS abbreviates Transportation, Communications, Electric, Gas, and Sanitary Services. FIRE abbreviates Finance, Insurance, and Real Estate. All estimates are normalized to have zero mean when weighted by employment shares. Column 3 of panels B and C equals the sum of columns 1 and 2, less year effects. Column 6 equals the sum of columns 2, 4, and 5. Standard errors are in parentheses.

TABLE 4
DECOMPOSITION OF MALE-FEMALE LOG EARNINGS DIFFERENTIALS

	BASELINE MODEL (1)	PERSON AND FIRM EFFECTS MODEL (2)	MATCH EFFECTS MODEL (3)
A. Component Due to Differences in Observable Characteristics			
Education	-.001	-.001	-.001
	(.000)	(.000)	(.000)
Race	-.015	-.015	-.015
	(.000)	(.000)	(.000)
Labor Force Experience	.002	.001	-.005
	(.000)	(.000)	(.000)
Time Effects	.000	.001	.001
	(.000)	(.000)	(.000)
Quarters Worked	.003	.000	.000
	(.000)	(.000)	(.000)
Industry	-.002		
	(.000)		
Subtotal: Differences in Characteristics	**-.013**	**-.014**	**-.020**
	(.000)	(.000)	(.000)
B. Component Due to Differences in Returns to Observable Characteristics			
Education	-.021	.106	.024
	(.006)	(.001)	(.002)
Race	.063	.054	.057
	(.000)	(.000)	(.001)
Labor Force Experience	-.307	-.172	-.155
	(.007)	(.000)	(.005)
Time Effects	.000	.000	.000
	(.000)	(.000)	(.000)
Quarters Worked	.002	-.003	-.003
	(.000)	(.000)	(.000)
Industry	-		
Subtotal: Differences in Returns	**-.264**	**-.014**	**-.076**
	(.001)	(.001)	(.001)
C. Component Due to Differences in Unobservables			
Unobserved Person Effects (α)		.000	.008
		(.001)	(.000)
Firm Effects (φ)		-.069	-.036
		(.000)	(.000)
Due to Inter-Industry Sorting		-.033	-.011
		(.000)	(.000)
Due to Intra-Industry Sorting		-.036	-.026
		(.000)	(.000)
Match Effects (ϕ)			.000
			(.000)
Subtotal: Differences in Unobservables		**-.070**	**-.028**
		(.001)	(.001)
Unexplained (Difference in Intercepts)	**-.083**	**-.261**	**-.236**
	(.007)	(.002)	(.005)
Total Male-Female Earnings Differential	**-.360**	**-.360**	**-.360**
	(.001)	(.001)	(.001)

Source: Author's calculations based on LEHD data.
Notes: Standard errors are in parentheses. There is no industry component in Column 1 of Panel B because industry effects are not interacted with sex in the baseline model.

APPENDIX TABLE 1
VARIABLE DEFINITIONS

VARIABLE	DEFINITION
Log Real Annualized Earnings	Natural logarithm of real annualized earnings, in 1990 dollars. Annualized earnings equal the annual average of reported UI earnings in full quarters* of employment, multiplied by four. Deflated using the CPI.
Education	Dummy variables for educational attainment, interacted with sex. Categories are Less than High School (reference), High School, Some College or Vocational Training, College Degee, and Graduate or Professional Degree.
Race	Dummy variables for Nonwhite and Race Missing, interacted with sex.
Labor Force Experience	Quartic in experience and dummy variable for negative potential experience in first quarter of employment, all interacted with sex. Potential experience is start-of-quarter age minus years of education minus six in the first quarter an individual appears in the sample. Experience increases by 0.25 in each subsequent quarter of employment.
Time Effects	Dummy variables for year, interacted with sex. Reference category is 1990.
Quarters Worked	Dummy variables for the number of full quarters* worked during the calendar year, interacted with sex. Reference category is four full quarters.
Industry	Dummy variables for SIC Major Division. Reference category is Manufacturing.
Sex	Dummy variable for female.

* An individual is defined to have worked a full quarter at firm j in quarter q if she was employed at firm j in quarter q-1 and quarter q+1.