Econometrica, Vol. 64, No. 4 (July, 1996), 957–979

THE EFFECT OF UNIONS ON THE STRUCTURE OF WAGES: A LONGITUDINAL ANALYSIS

BY DAVID CARD[1]

This paper studies the effects of unions on the structure of wages, using an estimation technique that explicitly accounts for misclassification errors in reported union status, and potential correlations between union status and unobserved productivity. The econometric model is estimated separately for five skill groups using a large panel data set formed from the U.S. Current Population Survey. The results suggest that unions raise wages more for workers with lower levels of observed skills. In addition, the patterns of selection bias differ by skill group. Among workers with lower levels of observed skill, unionized workers are positively selected, whereas union workers are negatively selected from among those with higher levels of observed skill.

KEYWORDS: Longitudinal data, unobserved heterogeneity, measurement error, trade unions.

DESPITE A LARGE AND SOPHISTICATED LITERATURE there is still substantial disagreement over the extent to which differences in the structure of wages between union and nonunion workers represent an *effect* of trade unions, rather than a consequence of the nonrandom selection of unionized workers. Over the past decade several alternative approaches have been developed to control for unobserved heterogeneity between union and nonunion workers.[2] One method that has been successfully applied in other areas of applied microeconometrics is the use of longitudinal data to measure the wage gains or losses of workers who change union status. Unfortunately, longitudinal estimators are highly sensitive to measurement error: even a small fraction of misclassified union status changes can lead to significant biases if the true rate of mobility between union and nonunion jobs is low. This sensitivity led Lewis (1986) to essentially dismiss the longitudinal evidence in his landmark survey of union wage effects.

In this paper I present some new evidence on the union wage effect, based on a longitudinal estimator that explicitly accounts for misclassification errors in reported union status. The estimator uses external information on union status misclassification rates, along with the reduced-form coefficients from a multivariate regression of wages on the observed sequence of union status indicators, to isolate the causal effect of unions from any selection biases introduced by a correlation between union status and the permanent component of unobserved wage heterogeneity. Recognizing that unions may raise wages more or less for

[1] Originally prepared for the 1991 Conference of the Econometric Study Group in Bristol, England. I am grateful to Michael Quinn for outstanding research assistance, and to Gary Solon for pointing out an error in an earlier draft. Thanks to Orley Ashenfelter, Henry Farber, Alan Krueger, the editor and two referees for comments.

[2] See Robinson (1989) for a discussion of these approaches and a comparison of the underlying assumptions typically used in each.

workers of different skill levels, and that the selection process into unionized jobs may generate different selection biases for workers with different levels of observed skills, the econometric model is estimated separately for five skill groups using a large panel data set formed from the 1987 and 1988 Current Population Surveys.

Simple cross-sectional estimates of the union-nonunion wage gap are large and positive for workers with lower levels of observed skills (35 percent for workers in the lowest quintile of the distribution of observed skills) and negative for workers with the highest levels of observed skills (-10 percent for workers in the upper quintile of observed skills). Estimates from a measurement-error-corrected longitudinal estimator suggest that this pattern arises from a combination of a larger union wage effect for less-skilled workers and opposing patterns of selection bias for unionized workers from the upper and lower tails of the observed skill distribution. Among workers with lower levels of observable skills, union members are positively selected, leading to a *positive* bias in the OLS union wage gap. Among workers with higher levels of observable skill, on the other hand, union members are negatively selected, leading to a *negative* bias in the OLS union wage gap. Perhaps surprisingly, estimates for a pooled sample indicate essentially no selection bias, suggesting that the opposing selection biases for more- and less-skilled workers approximately "cancel" in the overall workforce. These findings shed some new light on the nature of the union selection process, and suggest that both employer and employee incentives affect the nature of the unobserved differences between union and nonunion workers.

1. A CORRELATED RANDOM EFFECT MODEL WITH MISCLASSIFICATION ERRORS

This section outlines a longitudinal estimation technique for identifying the relative wage effect of unions in the presence of unobserved heterogeneity between union and nonunion workers and misclassification errors in measured union status.[3] As a starting point it is useful to consider the effects of measurement error in a model with no correlation between union status and unobserved productivity components. Let w_i represent the logarithm of wages of individual i in some time period and let u_i^* represent an indicator variable for the true union status of i in that period. Assume that wages are determined by

$$w_i = a + \beta x_i + \delta u_i^* + \epsilon_i, \tag{1}$$

where x_i is a vector of observed covariates, δ represents the (causal) effect of unions on the level of wages, and ϵ_i is an unobserved wage component with $E(\epsilon_i) = E(\epsilon_i x_i) = E(\epsilon_i u_i^*) = 0$.[4]

[3] Jakubson (1990) presents a similar model. Unlike Jakubson, I assume that external information is available on the misclassification rates of union status.

[4] Note that the union wage effect is assumed to be constant across individuals. In the empirical work later in the paper this assumption is relaxed by estimating separate models by skill group.

Actual union status u_i^* is unobserved: instead, an indicator u_i is observed that is only imperfectly correlated with u_i^*. Throughout this paper I assume that the process generating observed union status is of a particularly simple form, with a constant probability q_1 of observing $u_i = 1$ when $u_i^* = 1$, and a constant probability $q_0 < q_1$ of observing $u_i = 1$ when $u_i^* = 0$.[5] These assumptions, together with the assumption that u_i^* is orthogonal to ϵ_i, imply that observed union status is orthogonal to ϵ_i.[6] Letting π denote the true fraction of union workers in the population, the observed union rate is $p = q_1\pi + q_0(1-\pi)$.

To determine the relationship between wages and observed union status, consider the auxiliary regression

$$u_i^* = \gamma_0 + \gamma_1 u_i + \gamma_x x_i + \eta_i, \tag{2}$$

where $E(u_i\eta_i) = E(x_i\eta_i) = 0$. Equations (1) and (2) imply that wages are related to observed union status and the observed covariates by

$$w_i = (a + \delta\gamma_0) + \delta\gamma_1 u_i + (\beta + \delta\gamma_x)x_i + e_i, \tag{3}$$

where $e_i = \delta\eta_i + \epsilon_i$ is orthogonal to u_i. An OLS regression of wages on observed union status yields a consistent estimate of the product $\delta\gamma_1$, rather than the union wage effect δ. The attenuation coefficient γ_1 is

$$\gamma_1 = \text{cov}[u_i^*, u_i|x_i]/\text{var}[u_i|x_i],$$

where $\text{cov}[u_i^*, u_i|x_i]$ denotes the covariance of u_i^* and u_i, partialling out the effect of x_i, and var $[u_i|x_i]$ denotes the variance of u_i, partialling out x_i.

In the absence of any observed covariates,

$$\begin{aligned}\gamma_1 = \gamma_1^o &= \text{cov}[u_i^*, u_i]/\text{var}[u_i] \\ &= [E(u_i^* u_i) - p\pi]/p(1-p) \\ &= \pi/p \cdot (q_1 - p)/(1-p),\end{aligned}$$

which is less than 1 if $q_1 < 1$ and $\pi \approx p$. More generally, denote the linear projection of true union status on x_i by

$$u_i^* = \pi + c(x_i - \bar{x}) + v_i,$$

where $\bar{x}$ represents the mean of x_i and $E(v_i) = E(x_i v_i) = 0$. Assuming that $P(u_i = 1|u_i^*, x_i)$ depends only on u_i^* (i.e., that the misclassification rates are constant across individuals) the implied linear projection of *observed* union status on x_i is

$$u_i = p + (q_1 - q_0)\cdot c(x_i - \bar{x}) + v_i'$$

[5] q_0 is the "false positive" rate, while $(1 - q_1)$ is the "false negative" rate.

[6] Under the assumed measurement model $u_i = D_{i1}\cdot u_i^* + D_{i0}\cdot(1 - u_i^*)$, where D_{i1} is a random indicator variable with mean q_1, D_{i0} is a random indicator variable with mean q_0, and D_{i1} and D_{i0} are mutually independent and independent of u_i^* and x_i. Thus $E(\epsilon_i u_i) = q_0 E(\epsilon_i) + (q_1 - q_0)E(\epsilon_i u_i^*) = 0$.

where v_i' is orthogonal to x_i.[7] Combining these two expressions with the formula for γ_1 yields

$$\gamma_1 = \frac{\operatorname{cov}[u_i^* - \pi - c(x_i - \bar{x}), u_i - p - (q_1 - q_0)\cdot c(x_i - \bar{x})]}{\operatorname{var}[u_i - p - (q_1 - q_0)\cdot c(x_i - \bar{x})]},$$

$$= \frac{\operatorname{cov}[u_i^*, u_i] - (q_1 - q_0)c'V_{xx}c}{\operatorname{var}[u_i] - (q_1 - q_0)^2 c'V_{xx}c},$$

where V_{xx} is the population variance-covariance matrix of x_i. Let $R^2 = (q_1 - q_0)^2 c'V_{xx}c/\operatorname{var}[u_i]$ denote the (theoretical) R-squared coefficient from a linear probability model for observed union status. Then the previous expression can be written as

$$\gamma_1 = \frac{\gamma_1^o - R^2/(q_1 - q_0)}{1 - R^2}, \tag{4}$$

where γ_1^o is the attenuation coefficient in a model with no other covariates. Note that the addition of x's that are correlated with true union status (i.e., that lead to an $R^2 > 0$) exacerbates the attenuation effect of measurement error.[8]

Suppose that consistent estimates of q_0 and q_1 are available. Then a consistent estimate of the true union wage effect δ can be obtained in two steps by first estimating an unrestricted regression of wages on observed union status and the x's (providing a consistent estimate of $\delta\gamma_1$), and then using the estimates of q_0 and q_1, together with estimates of the observed fraction of union workers and the R^2 coefficient from a linear probability model of observed union status, to form a consistent estimate of the attenuation coefficient γ_1 from equation (4).

Allowing for a Correlation Between Union Status and Unobserved Wage Determinants

The preceding analysis relies on the maintained assumption that union status is orthogonal to the unobserved components of wages. The availability of multiple observations on wages and union status for the same individual over time provides an opportunity to relax this assumption. Specifically, suppose that the observed wage of individual i in period t is determined by

$$w_{it} = a_t + \beta_t x_i + \delta u_{it}^* + \epsilon_{it}, \tag{5a}$$

where u_{it}^* denotes the true union status of individual i in period t, x_i represents a vector of observed control variables,[9] and ϵ_{it}, the unobserved component of

[7] This follows from the observation that $E[u_i(x_i - \bar{x})] = E[D_{i0}(x_i - \bar{x})] + E[(D_{i1} - D_{i0})u_i^*(x_i - \bar{x})] = (q_1 - q_0)\operatorname{cov}[u_i^*, x_i]$, where the notation is the same as in footnote 6.

[8] Nevertheless, if the fraction of explained variance in a linear model for observed union status is relatively low, and if the misclassification rates are small, then the multivariate attenuation coefficient γ_1 is not too different from the univariate coefficient γ_1^o.

[9] The vector x_i is assumed to include time-invariant characteristics, as well as the complete history of any time-varying characteristics.

wages, is decomposable into the sum of a permanent and a transitory component:

$$\epsilon_{it} = \alpha_i + \epsilon'_{it}, \tag{5b}$$

with $\Sigma_t \epsilon'_{it} = 0$. In general, both the permanent and transitory components of ϵ_{it} may be correlated with union status in any given period. In this paper, however, I make the simplifying assumption that union status is correlated only with the permanent component of ϵ_{it}. Specifically, let U^*_{ih} represent an indicator variable for the hth possible "union history" in the longitudinal sample ($h = 1, 2 \ldots H$). Then I assume that $E(U^*_{ih}\epsilon'_{it}) = 0$ for all h and all t. This assumption imposes testable restrictions in even a two-period setting: in particular, it implies that the wage changes of union joiners and union leavers are equal and opposite in sign (see below).

Following Chamberlain (1982), I further assume that the permanent component of wages can be decomposed into a linear function of the observed covariates and indicators for all but one of the possible union histories:

$$\alpha_i = \phi_1 + \sum_{h=2}^{H} U^*_{ih}\phi_h + \lambda x_i + \xi_i, \tag{6}$$

where ξ_i is an error component with $E(\xi_i U^*_{ih}) = E(\xi_i x_i) = 0$. Equations (5) and (6) together generate a multivariate regression model for observed wages in each period that depends on x_i and on the indicators for the union histories. If only two periods of data are available—as is the case for the sample analyzed below—then the possible union histories correspond to the set {00, 01, 10, 11}, where '01', for example, refers to the union status of an individual who is nonunion in period 1 and unionized in period 2. In the two-period case, the complete model for wages is

$$\begin{aligned} w_{i1} &= a_1 + \phi_1 + (\beta_1 + \lambda)x_i + (\delta + \phi_{10})U^*_{i10} + \phi_{01}U^*_{i01} \\ &\quad + (\delta + \phi_{11})U^*_{i11} + \xi_i + \epsilon'_{i1}, \\ w_{i2} &= a_2 + \phi_1 + (\beta_2 + \lambda)x_i + \phi_{10}U^*_{i10} + (\delta + \phi_{01})U^*_{i01} \\ &\quad + (\delta + \phi_{11})U'_{i11} + \xi_i + \epsilon'_{i2}, \end{aligned}$$

where the union history '00' is treated as the omitted category.

If true union status is unobservable then these equations are not directly estimable. As in a one-period model, however, it is possible to express wages in terms of the observed union status indicators using a series of auxiliary regressions. Let $U_i = (U_{i10}, U_{i01}, U_{i11})$ represent a vector of observed union history dummies (treating the indicator for a '00' history as the omitted category), and consider the set of auxiliary regressions:

$$U^*_{ih} = \gamma_{0h} + \gamma_h U_i + \gamma_{xh} x_i + \eta_{ih}, \qquad h = \{10, 01, 11\}. \tag{7}$$

Then wages are related to x_i and the vector of observed union histories by

$$(8a)\qquad w_{i1} = a_1' + \{\beta_1 + \lambda + (\delta + \phi_{10})\gamma_{x10} + \phi_{01}\gamma_{x01} + (\delta + \phi_{11})\gamma_{x11}\}x_i + \{(\delta + \phi_{10})\gamma_{10} + \phi_{01}\gamma_{01} + (\delta + \phi_{11})\gamma_{11}\}U_i + e_{i1},$$

$$(8b)\qquad w_{i2} = a_2' + \{\beta_2 + \lambda + \phi_{10}\gamma_{x10} + (\delta + \phi_{01})\gamma_{x01} + (\delta + \phi_{11})\gamma_{x11}\}x_i + \{\phi_{10}\gamma_{10} + (\delta + \phi_{01})\gamma_{01} + (\delta + \phi_{11})\gamma_{11}\}U_i + e_{i2},$$

where

$$e_{i1} = (\delta + \phi_{10})\eta_{10} + \phi_{01}\eta_{01} + (\delta + \phi_{11})\eta_{11} + \xi_i + \epsilon_{i1}'$$

and

$$e_{i2} = \phi_{10}\eta_{10} + (\delta + \phi_{01})\eta_{01} + (\delta + \phi_{11})\eta_{11} + \xi_i + \epsilon_{i2}'$$

are orthogonal to the vector (U_i, x_i).

Expressions for the γ_h coefficients from the auxiliary regressions (7) are easily derived under the assumption that the union status misclassification rates are constant across individuals, and independent over time for the same individual. Specifically, assume

$$(9)\qquad P(u_{i1}, u_{i2}|u_{i1}^*, u_{i2}^*, x_i) = P(u_{i1}|u_{i1}^*)P(u_{i2}|u_{i2}^*) \quad \text{with}$$
$$P(u_{it} = 1|u_{it}^*) = q_0, \qquad u_{it}^* = 0,$$
$$= q_1, \qquad u_{it}^* = 1,$$

where u_{it}^* and u_{it} are indicators for the true and observed union status of individual i in period t, respectively. For any particular observed union history j and any true history k, let $\tau_{jk} = P(U_{ij} = 1|U_{ik}^* = 1)$. Under the assumptions specified in equation (9), τ_{jk} is a simple function of q_0 and q_1.[10] Let T represent the 4-by-4 matrix whose jth row and kth column is τ_{jk}, let $\pi = (\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11})$ represent the vector of probabilities of the true union histories, and let $p = (p_{00}, p_{10}, p_{01}, p_{11})$ represent the vector of probabilities of the observed union histories. Then the observed and true union status probabilities are related by $p = T\pi$.

The auxiliary coefficients γ_h from equation (7) can be calculated by first projecting U_{ih}^* and U_i on x_i, and then projecting the residual component of U_{ih}^* on the residual component of U_i. Denote the linear projection of an indicator for *true* union history h on the observed x's by

$$(10)\qquad U_{ih}^* = \pi_h + c_h(x_2 - \bar{x}') + v_{ih}, \qquad h = \{10, 01, 11\}.$$

Similarly, denote the projection of the jth *observed* union status indicator on x_i by

$$(11)\qquad U_{ij} = p_j + \zeta_j(x_i - \bar{x}) + v_{ij}', \qquad j = \{10, 01, 11\}.$$

[10] For example, $P(U_{10} = 1|U_{11}^* = 1) = q_1(1 - q_1)$.

If union status misclassification rates are constant across individuals and constant over time (as specified by equation (9)), then the coefficients ζ_j in equation (11) are related to the coefficients c_h in equation (10) by $\zeta = c\Omega'$, where $\zeta = [\zeta_{10}, \zeta_{01}, \zeta_{11}]$, $c = [c_{10}, c_{01}, c_{11}]$, and Ω is a 3-by-3 matrix whose (j,k) element is $\tau_{jk} - \tau_{j00}$. Using equations (10) and (11), the auxiliary regression coefficients in equation (7) can be written as

$$\gamma_h = \{\mathrm{var}[U_i] - \Omega c' V_{xx} c \Omega'\}^{-1} \cdot \{\mathrm{cov}[U_i, U_{ih}^*] - \Omega c' V_{xx} c\}, \tag{12}$$

where V_{xx} is the variance-covariance matrix x_i. For given values of V_{xx} and c, the coefficients γ_h are functions of the misclassification rates and the vector of true union status probabilities.

Assuming that estimates of V_{xx}, c, and the misclassification rates are available, the coefficients of the observed union status indicators in the wage equations (8a) and (8b) are functions of 7 parameters: the union wage effect δ, the coefficients $\{\phi_{10}, \phi_{01}, \phi_{11}\}$, and the probabilities $\{\pi_{10}, \pi_{01}, \pi_{11}\}$. In this paper I use a two-step estimation procedure for deriving estimates of the union wage effect δ. First, I estimate unrestricted reduced-form regressions for wages in each period that include the observed covariates as well as indicators for the observed union histories (i.e., unrestricted versions of equations (8)). I also estimate linear probability models for the observed union status indicators as functions of the observed x-variables (providing estimates of c and V_{xx}). I then combine the 6 reduced-form union status coefficients from equations (8) with estimates of the sample fractions of each observed union history (3 estimated probabilities) and use a second-stage minimum-distance estimator to fit these 9 sample moments as functions of the 7 structural parameters ($\delta, \phi_{10}, \phi_{01}, \phi_{11}, \pi_{10}, \pi_{01}, \pi_{11}$), treating q_0, q_1, V_{xx}, and c as fixed constants.[11] The second stage models are over-identified with 2 degrees of freedom, providing a test of the assumptions underlying the model.

2. ESTIMATING THE MISCLASSIFICATION RATE OF UNION COVERAGE IN THE CPS

The estimation procedure outlined above relies on the availability of external information on union status misclassification rates. In the empirical work reported below I apply the procedure to a panel data set formed from the 1987 and 1988 Current Population Surveys. A distinctive feature of the Current Population Survey (CPS) is the availability of information from a 1977 validation survey that was designed to measure the reliability of employee-provided job data. This survey collected wage and union status information for a sample of workers, and then gathered the same data from each respondent's employer.[12]

[11] The second-stage estimator minimizes a quadratic form in the deviations between the actual and predicted reduced-form parameters, using the inverse covariance matrix of the estimated reduced-form parameters as a weighting matrix.

[12] This survey has been previously analyzed by Mellow and Sider (1983) and Freeman (1984).

TABLE I

CROSS-TABULATIONS OF EMPLOYER AND EMPLOYEE REPORTS OF UNION COVERAGE: ADULT MEN IN JANUARY 1977 CPS

1. *All Industries*

Employee Report	Employer Report: Union	Employer Report: Nonunion
Union	523 (30.4%)	43 (2.5%)
Nonunion	46 (2.7%)	1106 (64.4%)

2. *Manufacturing Industries*

Employee Report	Employer Report: Union	Employer Report: Nonunion
Union	246 (46.2%)	14 (2.6%)
Nonunion	12 (2.3%)	261 (49.0%)

3. *Trade and Service Industries*

Employee Report	Employer Report: Union	Employer Report: Nonunion
Union	97 (15.9%)	16 (2.6%)
Nonunion	14 (2.3%)	483 (79.2%)

Notes: The entries in each panel are the number of cases and the percent of responses (in parentheses). Sample consists of 1718 men age 24–66 who report a valid wage, with nonmissing reports of union coverage from both the employer and employee. Union status refers to coverage of job by a union contract.

The 1977 validation survey provides a unique source of information on the misclassification rates in CPS union-status questions.[13]

Table I presents a series of cross-tabulations of union coverage responses from employees and employers in the 1977 validation survey. The sample consists of 1,718 men age 24–66 who reported union status and earnings data, and whose employers also reported valid union coverage information. The upper panel of the table reports the cross-tabulation of union responses for workers in all industries, while the middle and lower panels give sector-specific cross-tabu-

[13] Unfortunately, the union status questions in the regular Current Population Survey are not exactly the same as the question in the January 1977 study. The CPS asks individuals if they are members of a labor union or employee association, and if not, whether they are covered by a labor union on their job. The January supplement asked both the employer and employee whether the employee's pay rate was set by a union contract. I assume that misclassification rates measured by the 2-part questions in the regular CPS are the same as the error rates in the January 1977 question.

lations for employees in manufacturing and trade and services sectors, respectively.

These simple cross-tabulations display two striking features: (i) in each table, the two off-diagonal probabilities are approximately equal; and (ii) the off-diagonal probabilities are similar across sectors, even though the overall unionization rate is much higher in manufacturing than in trade and services. Most analysts of the CPS validation survey have assumed that the employer responses to the union status question are "true" and that the employee responses are measured with error.[14] Under this assumption, however, a symmetric cross-tabulation will *only* arise if the relative error rates of union and nonunion workers vary with the odds of union coverage. In particular, if the true unionization rate is π, and the employers' responses are correct, then the probability that the employer reports coverage and the employee reports noncoverage is $\pi(1-q_1)$, whereas the probability that the employer reports noncoverage and the employee reports coverage is $(1-\pi)q_0$. Symmetry of the cross-tabulation therefore requires $q_0/(1-q_1) = \pi/(1-\pi)$. In the manufacturing sector, $\pi \approx .5$, implying that the false positive rate and false negative rate are about equal for manufacturing workers. In trade and services, on the other hand, $\pi \approx .2$, implying that the false negative rate is 4 times greater than the false positive rate in that sector.

An alternative to the hypothesis that relative error rates vary systematically by industry is that both employer and employee responses are measured with error, and that the misclassification rates are about equal. To pursue this idea, suppose that union and nonunion employers and employees all have the same probability q of reporting the incorrect union status. Then the cross-tabulations in Table I are functions of only two parameters: the true fraction of union coverage (π) and the misclassification rate ($q = q_0 = 1 - q_1$). It is easy to see that in this "symmetric misclassification model" the off-diagonal probabilities of the cross-tabulation will be equal and independent of the true level of union coverage.[15] Both features are displayed in Table I.

A more formal way to test the symmetric misclassification model is by a goodness-of-fit test—the model has 2 parameters and can be fit to the 3 independent elements of the cross-tabulation by minimum chi-square methods. The best fit to the overall table (in Panel 1) yields $q = 0.027$ and $\pi = 0.321$: the associated test statistic is 0.10 (with 1 degree of freedom). The misclassification rate is estimated relatively precisely, with a standard error of 0.0014. Assuming a 2.7% misclassification rate but treating the true union density as a free parameter gives chi-squared statistics of 0.24 for manufacturing (with $\pi = 0.485$) and 0.21 for trade and services (with $\pi = 0.167$). This simple model therefore provides an acceptable fit to the overall and sector-specific cross-tabulations.

[14] See, e.g., Mellow and Sider (1983). This was apparently the assumption that motivated the design of the study.

[15] The probability of observing either of the conflicting classifications is $\pi(1-q)q + (1-\pi)q(1-q) = q(1-q)$, independent of π.

TABLE II

ESTIMATED CROSS-SECTIONAL WAGE EQUATIONS USING ALTERNATIVE MEASURES OF UNION STATUS

Measure of Union Status	No Covariates			Covariates Included[a]		
	(1)	(2)	(3)	(4)	(5)	(6)
1. Employee reported coverage	0.125 (0.023)	—	—	0.205 (0.022)	—	—
2. Employer reported coverage	—	0.124 (0.023)	—	—	0.196 (0.022)	—
3. Product of employee and employer reports	—	—	0.137 (0.024)	—	—	0.225 (0.023)
R-squared	0.016	0.016	0.019	0.308	0.305	0.313

Notes: Standard errors in parentheses. The sample is described in Table I, and includes 1718 observations. The dependent variable in all models is the logarithm of hourly wages. The mean and standard deviation of the dependent variable are 1.746 and 0.460, respectively.

[a] Models also include education, potential experience and its square, indicators for nonwhite race, residence in the South, public-sector employment, and one-digit industry and occupation dummies.

Further evidence of symmetric measurement errors in the employers' and employees' union coverage responses is presented in Table II. This table shows the estimated union status coefficients from cross-sectional wage regressions fit to the January 1977 CPS sample using three alternative union measures: the worker-reported measure (row 1); the firm-reported measure (row 2), and their product (row 3). Columns 1–3 present estimated union coefficients from models with no other covariates, whereas columns 4–6 present coefficients from models that include a standard set of control variables (education, potential experience and its square, race, region dummies, and industry and occupation dummies).

If employers' union responses are treated as correct, then the attenuation formulas developed in Section 1 imply that the coefficients in columns 1, 2, and 3, should be related in the ratios of 1.00:0.89:0.96.[16] On the other hand, if employers and employees are assumed to have the same misclassification rates, the predicted ratio of the coefficients is 1.00:1.00:1.04.[17] Assuming that union status error rates are constant across individuals (and employers), the predicted

[16] In the absence of other control variables, the expected attenuation of the estimated union coefficient is $(\pi q_1 - p\pi)/p(1-p)$, where π is the true union rate, p is the mean of the observed union indicator, and q_1 is the probability of observed union status, given true status. Assuming that the employer response is correct the data in the upper panel of Table I imply $\pi = 0.331$ and $q_1 = 0.92$ for the employee response. An indicator formed from the product of the employer and employee responses has the same probability of a correct classification given true union coverage (i.e., $q_1 = 0.92$) and has mean πq_1.

[17] If employer and employee responses have the same misclassification rates, then the probability that both responses are 1 given true union coverage is $(q_1)^2$, where q_1 is the probability that either the worker or the firm reports union coverage when it is true.

ratios of the coefficients across the models with other control variables are approximately equal to the ratios in models without covariates, since observed union coverage has a relatively low coefficient of multiple correlation with the covariates included in Table II (see equation (4)).

Inspection of the coefficient estimates in Table II reveals that the union wage effects are approximately equal when union status is measured by either the employer's response or the employee's response, but rise when union status is measured by the product of their responses. This pattern is inconsistent with the assumption that the employers' responses are error-free, but is fully consistent with the hypothesis of equal misclassification rates in the employer and employee responses. Based on this evidence, and the cross-tabulations in Table I, I draw two main conclusions. First, both employee and employer union responses seem to contain measurement errors. Second, the misclassification rate in employee-reported union status in the CPS survey is on the order of 2.5–3.0 percent.[18]

3. LONGITUDINAL DATA FROM THE CPS

A panel data set with at least two observations per individual is required to implement the estimator developed in Section 1. While a number of potentially suitable data sets are available, I have elected to construct a two-period panel data set from the 1987 and 1988 Current Population Surveys. The main advantages of this data set are the large sample size, which permits a detailed investigation of union wage effects for different "skill groups," and the availability of information on union status misclassification rates. Offsetting these advantages is the relatively high attrition rate induced by the CPS sample design.[19] This section summarizes the construction of the CPS data set and presents some descriptive information on the resulting panel.

Every month one quarter of respondents in the CPS are administered supplemental questions on wage rates and union status for their main job. Twelve months later, one-half of these individuals are asked the same questions again.[20] I have used a statistical matching algorithm (described in the Appendix) to link information for adult men from corresponding months of the 1987 and 1988

[18] Freeman (1984) presents data from the May 1979 CPS, in which individuals were asked about their union status in two separate parts of the questionnaire. Conflicting union status reports were given by 3.2 percent of individuals. I fit the symmetric misclassification model to these data and obtained an estimate of the misclassification rate of 1.66 percent (with a chi-squared test statistic of 4.04). I regard this as a lower bound on the misclassification rate in the CPS, and perhaps indicative of the rate of miscoding by interviewers and transcribers.

[19] The CPS interviews residents of a rotating sample of housing units. Individuals who move out of a given housing unit are replaced by the individuals who move in. This fact, and potential confusion that arises if two individuals of similar age and sex live in the same housing unit, lead to a high nonmatching rate across interviews.

[20] The CPS design includes 8 rotation groups. Each group is surveyed for 4 months, then taken out of the sample for 8 months, and then surveyed for 4 months. Groups completing their 4th and 8th months in the survey answer the earnings and union status questions.

surveys. The algorithm compares the men in a particular household in 1987 to the men in the same household in 1988, and computes a match probability for each potential pair. The match probabilities depend on age, race, education, and marital status. Each person in the 1987 sample is then assigned his "best match," and deleted from the sample if the match probability falls below a critical value.

A relatively conservative critical value for the match probability yields an overall match rate of 69 percent.[21] A key correlate of the matching rate is age—the match rate rises from 50 percent for 25 year olds to around 80 percent for individuals over age 55. Match rates are also higher for whites than nonwhites (69.6% versus 62.7%), and for union than nonunion workers (73.2% versus 67.2%), but are fairly similar across occupation and education categories.

Table III illustrates some of the differences between the overall sample of adult male workers in the 1987 CPS and the subset of observations that are successfully matched to a 1988 record. The first column in the upper panel of the table shows the mean characteristics of individuals with valid earnings data for 1987 who could potentially match to a 1988 record.[22] The lower panel presents regression coefficients from a standard cross-sectional wage model fit to this sample. Column 2 presents similar information for the subset of men who are successfully matched to a 1988 observation. The matched sample is older, has a lower fraction of nonwhites, and a higher fraction of unionized workers. Some of the regression coefficients are also slightly different in the matched sample.

The empirical analysis in the next section is based on a subset of observations in the matched sample with valid (nonimputed) wages for both 1987 and 1988. This restriction eliminates men who were working in 1987 but were unemployed or out of the labor force in the same month in 1988, as well as individuals with imputed 1988 wage data. The characteristics of this "balanced" subsample are presented in the third column of the table. Relative to a representative cross-section of adult male workers (column 1), the balanced subsample has similar average age and education, but a lower fraction of nonwhites and Hispanics. The coefficients of a standard wage regression are also similar between the balanced subsample and the overall cross-section, although the returns to experience and the union-nonunion wage gap are slightly lower in the balanced subsample.

[21] At this critical value an individual record will *only* match if the respondent's age grows by 1 year between the 1987 and 1988 surveys, if the respondent's race and veteran status are the same in the two surveys, and if reported education is either fixed or increases by 1 year.

[22] For simplicity, I have deleted all observations with imputed earnings data in this table (and all subsequent analyses). Approximately 15 percent of individuals in the CPS have allocated earnings—this rate is not much different between matchers and nonmatchers. However, the inclusion of observations with allocated earnings affects some of the characteristics of the data, including the estimated union wage premium. The union wage gap for men with allocated earnings is roughly 0.

TABLE III

COMPARISONS OF VARIOUS SAMPLES OF ADULT MEN IN THE 1987 CURRENT POPULATION SURVEY

	All with Nonallocated Wage (1)	Subset Matched to 1988 (2)	Subset Matched with 1988 Wage (3)
Sample Characteristics:			
Sample Size	32,803	22,810	19,044
Average Age	39.2	40.6	40.1
Average Education	13.1	13.1	13.2
Percent Nonwhite	11.6	10.7	10.2
Percent Hispanic	6.0	4.8	4.6
Percent Union	26.5	28.1	28.8
Mean Log Wage	2.32	2.35	2.37
Standard Deviation of Log Wage	0.55	0.54	0.52
Estimated Regression Coefficients[a]			
Education	0.083	0.082	0.082
	(0.001)	(0.001)	(0.001)
Experience	0.038	0.034	0.033
	(0.001)	(0.001)	(0.001)
Experience^2	−0.063	−0.052	−0.051
(coefficient ×100)	(0.002)	(0.002)	(0.002)
Nonwhite	−0.187	−0.182	−0.176
	(0.008)	(0.010)	(0.010)
Hispanic	−0.134	−0.135	−0.137
	(0.011)	(0.014)	(0.016)
Union	0.183	0.166	0.154
	(0.006)	(0.007)	(0.007)
R-squared	0.290	0.279	0.282

Notes: See text for description of samples. Samples include men age 24–66 in rotation group 4 of the 1987 CPS monthly files.

[a] Regression models for log hourly wage. All models include 8 region dummies and indicators for central city and suburban residence.

4. UNION EFFECTS BY POSITION IN THE WAGE DISTRIBUTION

This section applies the estimation method outlined in Section 1 to the matched 1987–1988 CPS sample. Recognizing that the union wage effect may vary with a worker's skill level, and that the selection process into unionized jobs may lead to differing selection biases at different skill levels, the models are estimated separately for five different "skill groups." The groups are defined by quintiles of predicted wages in the nonunion sector, using an equation fit to an independent sample of workers in the 1987 and 1988 CPS.

A. Defining the Predicted Wage Quintiles

To develop a simple index of skill I fit a flexible wage equation to the pooled sample of nonunion workers in the "unmatchable" subset of the 1987 and 1988 CPS file (i.e., individuals in the 1987 CPS who would not be interviewed in 1988

TABLE IV

CHARACTERISTICS OF MEN IN 1987 CPS, BY PREDICTED WAGE QUINTILE

	Predicted Wage Quintile				
	1	2	3	4	5
All Workers in Quintile:					
Average Age	34.4	37.0	41.6	39.5	43.8
Average Education	10.3	12.0	12.8	14.7	16.8
Percent Nonwhite	26.6	12.0	4.9	8.5	3.4
Percent Union	23.5	30.3	33.1	24.7	19.7
Mean Log Wage	1.98	2.20	2.34	2.48	2.73
Standard Deviation of Log Wage	0.45	0.46	0.45	0.47	0.51
Nonunion Workers in Quintile:					
Average Age	33.5	36.1	40.6	38.8	43.8
Average Education	10.3	12.1	13.0	14.8	16.7
Percent Nonwhite	25.0	10.0	4.9	8.4	2.9
Mean Log Wage	1.89	2.10	2.27	2.47	2.75
Standard Deviation of Log Wage	0.44	0.46	0.48	0.50	0.54
Union Workers in Quintile:					
Average Age	37.1	39.1	43.4	41.7	43.9
Average Education	10.3	11.8	12.4	14.1	16.9
Percent Nonwhite	31.9	16.7	5.0	9.0	5.5
Mean Log Wage	2.26	2.43	2.49	2.52	2.66
Standard Deviation of Log Wage	0.38	0.34	0.34	0.35	0.37
Difference: Union-Nonunion:					
Mean Log Wage	0.37	0.33	0.21	0.05	−0.09
Standard Deviation of Log Wage	−0.06	−0.12	−0.14	−0.16	−0.16

Notes: Sample consists of men age 24–66 in rotation group 8 of monthly 1987 CPS files. Only observations with a nonallocated wage measure are included. Sample size is 33,385. Observations are stratified into quintiles on the basis of a predicted wage in the nonunion sector. See text for description of prediction equation.

and individuals in the 1988 CPS who had not been interviewed in 1987).[23] Using the estimated coefficients from this equation I then constructed a predicted wage for union and nonunion workers in the 1987 CPS sample (including those in the matched 1987–88 sample and those in the unmatchable sample). This predicted wage provides an index of observed skill for each individual that is unaffected by any distortionary effect of unions on the pay structure of unionized jobs.

Table IV shows the characteristics of workers in the unmatchable subset of the 1987 CPS, stratified into quintiles on the basis of their predicted nonunion wage. The table gives overall means for each quintile as well as means for the

[23] The equation includes region and central city dummies, 11 education dummy variables, linear and quadratic experience terms, indicators for veteran status, nonwhite race and Hispanic origin, and interactions between the race and experience terms and 3 broad education classes.

TABLE V

UNION FREQUENCIES AND ESTIMATED UNION WAGE EFFECTS: ADULT MEN IN MATCHED 1987–88 CPS FILE

	Percent Union		Cross-Sectional Union Wage Gap		Probabilities of Union Histories:			Estimated Reduced Form Coefficients					
								1987 Log Wages			1988 Log Wages		
Quintile[a]	1987	1988	1987	1988	'10'	'01'	'11'	'10'	'01'	'11'	'10'	'01'	'11'
1	25.4	25.1	0.37	0.36	5.14	4.87	20.22	0.230	0.109	0.413	0.066	0.208	0.403
			(0.02)	(0.02)				(0.030)	(0.031)	(0.017)	(0.031)	(0.032)	(0.018)
2	33.1	33.8	0.29	0.28	4.14	4.83	28.92	0.171	0.098	0.314	0.093	0.158	0.306
			(0.01)	(0.01)				(0.034)	(0.032)	(0.016)	(0.033)	(0.030)	(0.015)
3	36.2	36.5	0.20	0.18	3.82	4.16	32.33	0.131	0.043	0.210	0.095	0.119	0.191
			(0.01)	(0.01)				(0.033)	(0.031)	(0.014)	(0.033)	(0.031)	(0.014)
4	26.4	26.0	0.06	0.05	3.97	3.64	22.38	−0.012	−0.060	0.071	−0.012	−0.053	0.063
			(0.02)	(0.02)				(0.039)	(0.041)	(0.019)	(0.039)	(0.041)	(0.019)
5	22.3	22.4	−0.13	−0.13	3.74	3.79	18.57	−0.081	−0.156	−0.155	−0.095	−0.110	−0.138
			(0.02)	(0.02)				(0.040)	(0.040)	(0.020)	(0.039)	(0.039)	(0.019)
						All Quintiles Pooled							
	28.8	28.9	0.16	0.15	4.15	4.25	24.69	0.087	0.006	0.173	0.024	0.069	0.167
			(0.01)	(0.01)				(0.016)	(0.016)	(0.008)	(0.016)	(0.016)	(0.008)

Notes: Standard errors in parentheses. Estimated on sample of matched observations of men age 24–66 in the 1987 and 1988 CPS with valid (nonallocated) wages for 1987 and 1988. See text for list of covariates included in estimation.
[a] Observations are sorted into quintiles on the basis of a predicted nonunion wage. See text.

union and nonunion workers within each quintile. Not surprisingly, individuals in the lower quintiles are younger and less-educated, and are also more likely to be nonwhite. Within quintiles the characteristics of union and nonunion workers are similar, although union workers are more likely to be nonwhite. Comparisons of the wages of union and nonunion workers in each quintile (in the bottom two rows of the table) reveal two interesting patterns. First, the gap in average (log) wages between union and nonunion workers declines with the general level of skill: from a wage differential of 37% in quintile 1 to a differential of −9% in quintile 5.[24] Second, unionized workers in each quintile have lower wage dispersion than their nonunion counterparts (see Freeman (1980) and Freeman and Medoff (1984)).

B. Estimation Results

Table V reports information on the union status probabilities and union wage differentials for men in the matched 1987–1988 CPS data set. The sample is stratified into 5 quintiles using the same cutoffs for the predicted wage quintiles as in Table IV.[25] The first 2 columns of the table report the unionization rate by quintile and year. The extent of union coverage in the matched data set is

[24] Johnson and Youmans (1971) present an early analysis of the variation in union wage effects by skill (in their case, by age and education).

[25] Consequently, the 5 groups are not of exactly equal size in the matched panel. The sample sizes by quintile are 3695, 3600, 4395, 3347, and 4007.

slightly higher than in the 1987 cross-section, but shows a very similar pattern across the predicted wage quintiles. The third and fourth columns report estimated cross-sectional union wage differentials for 1987 and 1988 from models that include the full set of covariates used to form the predicted wage quintiles. Across quintiles the regression-adjusted union wage gaps show the same pattern as the unadjusted gaps in Table IV. The cross-sectional union wage gap is large and positive for the lowest wage quintile and negative for the fifth quintile.

Columns 5–7 of Table V give the sample fractions of each of the four possible union histories in each skill group. The fractions of union joiners and union leavers range from 4 to 5 percent, with relatively higher rates of mobility in the lower quintiles. Presumably, not all of the observed union transitions reflect a true change in union status. Indeed, if the misclassification rate is 2.8 percent, then one would expect to see a 2.7 percent union joining rate and a 2.7 percent union leaving rate, even in the absence of *any* real mobility between sectors. Close to one-half of the observed union status transitions over a two year period therefore can be attributed to measurement error.

Columns 8–13 give the reduced form wage coefficients corresponding to equations (8a) and (8b) in Section 1. In addition to a set of indicators for observed union status (whose coefficients are reported) the models include the same set of education, race, potential experience, and region variables used to form the predicted wage quintiles. Inspection of the coefficients of the observed union status variables suggests that some of the differences of the cross-sectional union wage gap across skill groups are attributable to differences in the unobserved characteristics of union and nonunion workers in each group. For example, the coefficient of the '01' history for 1987 wages is large and positive for quintiles 1 and 2, and large and negative for quintiles 4 and 5. Since individuals with a '01' history are nonunion in 1987 (ignoring measurement errors) these coefficients suggest that union joiners with lower observed skills have unobserved characteristics that generate above-average wages in the nonunion sector, whereas union joiners with higher observed skills have unobserved characteristics that generate below-average wages in the nonunion sector.

In the absence of measurement error, a simple method for eliminating unobserved heterogeneity between union and nonunion workers is to examine the wage changes of union joiners and leavers. These can be computed directly from the coefficients in Table V. For example, the average wage change of union joiners between 1987 and 1988 is the difference in the '01' coefficients between 1988 and 1987.[26] For the first quintile, this change is $0.208 - 0.109 = 0.099$. The average wage changes of joiners and leavers in each quintile are presented in Table VI, along with their associated standard errors. Compared to the cross-sectional estimates, these "fixed effects" estimates show less variation

[26] Since the reduced form wage equations do not restrict the coefficients of the observed covariates across the two years, differences computed in this way are regression-adjusted for the x variables.

TABLE VI

AVERAGE WAGE GAIN BETWEEN 1987 AND 1988 FOR UNION JOINERS AND UNION LEAVERS

Predicted Wage Quintile	Change in Mean Log Wage, 1987 to 1988: Joiners	Leavers
1	0.099	−0.164
	(0.024)	(0.023)
2	0.060	−0.078
	(0.023)	(0.025)
3	0.076	−0.036
	(0.023)	(0.024)
4	0.007	−0.000
	(0.031)	(0.029)
5	0.047	−0.014
	(0.029)	(0.030)
All	0.063	−0.063
	(0.012)	(0.012)

Note: Standard errors in parentheses. Estimates based on reduced-form parameter estimates in Table V. See text.

across skill groups, and suggest a uniformly positive union wage effect. It should be noted, however, that any reporting errors in observed union status will attenuate the measured wage gains or losses of observed union joiners or leavers. Furthermore, if misclassification rates are constant across skill groups, the degree of attenuation will tend to be higher for groups with lower observed union transition rates.[27]

The two-step estimation strategy described in Section 1 identifies the union wage effect in the presence of both unobserved heterogeneity and misclassification errors in union status. Results from the second-stage estimation, applied separately for each quintile and for the sample as a whole, are presented in Table VII. The models are estimated using the reduced-form coefficients for the observed union indicators in Table V as well the estimated fractions of each union history. The estimation assumes a fixed 2.8% misclassification rate, and uses the estimated coefficients from linear probability models for the observed union histories in each quintile.[28] Parameter estimates are reported in the first 7 columns of the table, along with a goodness-of-fit statistic in the eighth column. The two right-hand columns give implied estimates of two of the key auxiliary

[27] To check if misclassification rates vary across skill groups I divided the men in the 1977 CPS into predicted wage quintiles and computed the cross-tabulations of employer and employee union responses by quintile. The assumption of a fixed misclassification rate is easily accepted in all the quintiles.

[28] As specified in equation (11), these models are estimated for the observed '01', '10', and '11' histories using the same set of covariates included in the reduced-form wage models in Table V. The R-squared coefficients of the models range from 1–3 percent (for the models of the probability of an observed union joiner or leaver) to 8–10 percent (for the models of the probability of an observed union stayer).

TABLE VII

SUMMARY OF STRUCTURAL ESTIMATION, BY QUINTILE

Predicted Wage Quintile	Estimated Structural Parameters: δ	ϕ_{10}	ϕ_{01}	ϕ_{11}	π_{10}	π_{01}	π_{11}	Goodness of fit[a]	Implied Auxilliary Regression Coefficients[b] $\gamma(U_{10}^*\|U_{10})$	$\gamma(U_{01}^*\|U_{01})$
1	0.282	0.073	0.117	0.135	2.86	2.21	21.20	2.80	0.49	0.43
	(0.036)	(0.061)	(0.070)	(0.040)	(0.37)	(0.38)	(0.70)			
2	0.164	0.106	0.083	0.149	1.68	2.29	30.43	1.14	0.37	0.44
	(0.040)	(0.085)	(0.068)	(0.043)	(0.34)	(0.39)	(0.80)			
3	0.184	0.135	0.010	0.018	1.11	1.68	34.11	5.14	0.27	0.37
	(0.052)	(0.115)	(0.082)	(0.053)	(0.32)	(0.31)	(0.75)			
4	0.008	−0.076	−0.275	0.059	1.39	1.05	23.55	0.44	0.32	0.25
	(0.066)	(0.120)	(0.171)	(0.068)	(0.37)	(0.35)	(0.77)			
5	0.108	−0.296	−0.430	−0.255	1.11	1.23	19.52	2.04	0.27	0.30
	(0.065)	(0.148)	(0.147)	(0.068)	(0.33)	(0.32)	(0.65)			
All	0.169	−0.011	−0.060	0.002	1.60	1.73	25.98	1.46	0.36	0.38
	(0.026)	(0.043)	(0.042)	(0.027)	(0.32)	(0.33)	(0.76)			

Notes: Standard errors in parentheses. Parameters are estimated by minimum distance, fitting the reduced-form coefficients and union history probabilities in Table V. All estimates assume a 2.8 percent misclassification rate for union status reporting.

[a] Distributed as chi-squared with 2 degrees of freedom under the null hypothesis of a correctly specified model.

[b] Implied coefficients of auxilliary regression of indicator for true union status on set of observed union status indicators. $\gamma(U_{10}^*|U_{10})$ denotes the regression coefficient of an indicator for observed status '10' in an auxilliary regression model for true status '10'. $\gamma(U_{01}^*|U_{01})$ denotes the regression coefficient of an indicator for observed status '01' in an auxilliary regression model for true status '01'.

regression coefficients from equation (7): the coefficient of an indicator for an observed union leaver in an auxiliary regression for true union-leaving status (denoted by $\gamma(U_{10}^*|U_{10})$); and the coefficient of an indicator for an observed union joiner in an auxiliary regression for true union-joining status (denoted by $\gamma(U_{01}^*|U_{01})$).

As suggested by the pattern of wage changes for union joiners and leavers, the measurement-error corrected longitudinal estimators of the union wage effect are uniformly positive, and are much less variable across quintiles than the cross-sectional wage gap. Interestingly, for the sample as a whole the corrected estimator is almost identical to the cross-sectional wage gap (17 percent versus 15–16 percent). At the extremes of the skill distribution, however, the corrected longitudinal estimator is much different: smaller than the cross-sectional estimator for the lowest quintiles (indicating a positive correlation between union coverage and the unobserved determinants of wages) and larger than the cross-sectional estimator for the highest quintiles (indicating a negative correlation between unionization and the unobserved determinants of wages). These results suggest that union workers with low levels of observed skill are positively selected, whereas union workers with high levels of observed skill are negatively selected. For union workers as a whole the selection biases for low- and high-skilled workers approximately offset each other.

The implied auxiliary regression coefficients relating indicators for the observed union transitions to the corresponding true transitions range from 25–50

TABLE VIII

ESTIMATED UNION WAGE EFFECTS UNDER ALTERNATIVE ASSUMPTIONS ON THE MISCLASSIFICATION RATE AND THE COVARIATES IN THE REDUCED FORM WAGE EQUATIONS

	Based on Reduced Form in Table V with Alternative Misclassification Rates			Based on Alternative Reduced Forms ($q = 0.028$)	
Predicted Wage Quintile	Base Case ($q = 0.028$) (1)	Low Estimate ($q = 0.025$) (2)	High Estimate ($q = 0.031$) (3)	No Covariates (4)	Industry Effects (5)
1	0.28	0.25	0.32	0.27	0.28
2	0.16	0.14	0.19	0.16	0.16
3	0.18	0.15	0.24	0.18	0.18
4	0.01	0.01	0.01	0.01	0.01
5	0.11	0.08	0.15	0.10	0.10
All	0.17	0.14	0.21	0.17	0.16

Notes: In Columns 1–3 estimates are obtained from unrestricted reduced form reported in Table V, using alternative values for the misclassification rate (q). In column 4 the estimates are obtained from reduced form models that exclude any other control variables. In column 5 the estimates are obtained from reduced form models that include 16 industry effects (8 effects for industry in each of 1987 and 1988). See text.

percent, with slightly higher values for the lower wage quintiles. These estimates imply that union status misclassification errors lead to a 50–75 percent attenuation in the average wage changes of observed union joiners and leavers, relative to the true wage changes of actual joiners or leavers.

As noted above, the second-stage structural models are over-identified with 2 degrees of freedom. The goodness-of-fit test statistics in Table VII are all below the corresponding 5% critical value (5.99). This suggests that the maintained assumptions of the statistical model—in particular the assumption that the transitory wage shocks are uncorrelated with true union status—are consistent with the data.

The structural parameter estimates, and especially the union wage effect δ, are relatively sensitive to the value of the misclassification rate assumed in the estimation. Table VIII shows the estimated values of δ under 3 alternative assumptions: $q = 0.028$ (the base case); $p = 0.025$ (a low estimate of the misclassification rate, given the evidence in Table I); and $p = 0.031$ (a high estimate). Higher values of the misclassification rate lead to larger estimates of the union wage effect, although the pattern of the estimated wage effects across quintiles is preserved.

The fourth and fifth columns of Table VIII report the results of two other specification checks. The parameter estimates in column 4 are obtained from reduced-form models with no other control variables. This specification is particularly simple because without additional x's, the auxiliary regression coefficients γ_h depend only on the misclassification rates and true union status probabilities, and are independent of the parameters of the linear probability models for the observed union status indicators (see equation (12)). The estimates of the union wage effects are very similar to the basis-case estimates from reduced-form models that include an extensive list of covariates. The estimates in column 5 are obtained from reduced-form models that include all the control

variables used in Table V as well as one-digit industry effects for the reported industry in each year.[29] Again, the estimated union wage effects are very similar to the basis-case estimates.[30]

In summary, the results of the structural estimation suggest two substantive conclusions. First, although a simple cross-sectional estimator provides a roughly unbiased estimator of the "true" union wage effect for a pooled sample of all workers together, the biases at either tail of the skill distribution are significant. The biases in the upper and lower tails are in opposite direction, with evidence of positive selection among union workers with lower observed skills and negative selection among union workers with higher observed skills. Second, even correcting for these selection biases, the union wage effect is bigger for workers with lower levels of observed skill.

5. INTERPRETATION OF THE RESULTS

What do these findings imply about the effects of unionization on the overall wage structure and the nature of the selection process into unionized jobs? One immediate implication of the finding that the "true" union wage effect is larger for less-skilled workers is that wage differences between broad skill groups tend to be compressed in the union sector. This is consistent with a long literature which finds that wage differentials by age, education, and region are typically smaller for unionized workers (see Lewis (1986) for a critical review of this literature).

A second implication of the results in Tables VII and VIII is that structural models which assume that the probability of union coverage is determined by a "single index" of observed and unobserved characteristics may be too restrictive. Most structural analyses of the union wage effect posit a three-equation model, consisting of an equation for the union wage for a given individual, an equation for the nonunion wage of the same individual, and a third equation defining a latent index (I_i) that determines the relative likelihood of holding a union job (see Lee (1978) and Robinson (1989), for example). In this class of models, the conditional expectation of any unobserved wage determinants given observed union status is a function *only* of the index I_i. Thus the selectivity biases in the union-nonunion wage gap are the same for any two groups of workers with the same probability of holding a union job. As shown in Table IV, individuals in the top and bottom quintiles of the observed skill distribution have (roughly) the same unionization rate. In the standard union selection model one would therefore expect similar selection biases to affect the union-nonunion wage differential for workers at the top and bottom of the observed skill distribution.

[29] The wage equation for 1987 includes a full set of dummies for industry in both 1987 and 1988. Likewise the wage equation for 1988 includes dummies for industry in both 1987 and 1988.

[30] Although the estimates of δ are insensitive to the choice of covariates, the estimated selection terms (the ϕ parameters) depend on the particular set of x's included in the reduced forms. On the other hand, the goodness-of-fit statistics and the estimated standard errors of δ are largely unaffected by the selection of control variables in the reduced forms.

Contrary to this prediction, however, the results in Table VII suggest that the selection biases are of opposite sign for these two groups.

In fact, the patterns of the selection biases by skill group and the tendency for unionized workers to be drawn from the middle of the skill distribution are more consistent with a *two-sided* selection model that incorporates both employer and employee behavior in the union selection process (see Abowd and Farber (1982)). To illustrate this point, suppose that the general productivity of a given individual (g_i) consists of two components:

$$g_i = z_i + a_i,$$

where z_i is an observable factor and a_i represents a productivity component that is observed by labor market participants but is unobserved in a conventional data set. Suppose that the wage in a nonunion job for a worker with general productivity g_i is

$$w_i^n = g_i + \epsilon_i^n, \tag{13a}$$

where ϵ_i^n represents the effects of randomness or other factors. Suppose further that the structure of wages is "flattened" in the union sector, so that the union wage of worker with productivity g_i is

$$w_i^u = \theta_0 + \theta_1 g_i + \epsilon_i^u, \tag{13b}$$

where $\theta_0 > 0$ and $0 < \theta_1 < 1$.

To complete the model, suppose that a worker is observed to hold a union job if two criteria are satisfied: (i) the worker's expected union wage exceeds his expected nonunion wage by more than the person-specific disutility that the individual attaches to working the union sector; and (ii) the worker's expected union wage is less than the sum of his general productivity plus a firm-specific match component. If ρ_i denotes the individual's disutility of working in a unionized job, the first of these conditions requires

$$g_i < \theta_0/(1-\theta_1) - \rho_i/(1-\theta_1). \tag{14a}$$

Similarly, if ω_i denotes an individual-specific match component at a unionized employer, the second condition requires

$$g_i > \theta_0/(1-\theta_1) - \omega_i/(1-\theta_1). \tag{14b}$$

This simple two-sided selection model has three implications that are broadly consistent with the findings in the previous section. First, by assumption, the "true" union-nonunion wage gap is lower for more highly skilled workers. Second, since highly productive workers are less likely to want to work in the union sector, whereas unionized employers are less likely to want to hire a low-productivity worker, the union sector is predicted to include more workers from the "middle" of the skill distribution, and relatively few workers from either tail. Finally, conditional on a high level of *observed* skill, the worker's selection criterion (14a) is more likely to be binding than the firm's selection

criterion (14b). Thus, for workers of higher levels of observed skill, those in the union sector are more likely to have negative values of the unobserved skill component a_i (i.e., a negative selection bias). On the other hand, conditional on a low level of observed skill, the firm's selection criterion is more likely to be binding than the worker's selection criterion. Unionized workers with lower levels of observed skill are therefore more likely to have higher values of a_i (i.e., a positive selection bias).

While a model with a two-sided selection process is broadly consistent with the findings in this paper, more research is clearly required to fully understand the effects of unions on the structure of wages, and to model the union selection process. In particular, the development and testing of a fully-specified dynamic model for wages and union status remain for future work.

Dept. of Economics, Princeton University, Princeton, N.J. 08544, U.S.A.

Manuscript received September, 1991; final revision received July, 1995.

APPENDIX

Construction of Matched CPS Sample

The data set is based on the merged monthly files of the outgoing rotation groups in the 1987 and 1988 CPS. The procedure for matching observations in the 1987 and 1988 files followed five steps:

1. Create a file containing one record for each household in the 4th rotation group of the 1987 CPS with one or more men age 24–67. Record for each male age 24–66 in the household (up to 7 men per household) the individual's age, race, education (highest grade attended), marital status, veteran status, and the number of people in the household. The 1987 file has 44,265 households.

2. Create a file containing one record for each household in the 8th rotation group of the 1988 CPS with one or more men age 24–67. Record the information listed above for each male age 24–67 (up to 7 men per household). The 1988 file has 42,318 households.

3. Merge the 1987 and 1988 households by CPS household identifier. The merged data set has 36,501 households.

4. For each individual in the 1987 household compute a "match probability" for matching with every observation in the 1988 household. Compute a "match probability" for matching each male in the 1988 household with every observation in the 1987 household.

5. Delete potentially matched observations with a "match probability" of 0.3 or less. Then retain only one matched observation per original observation in either the 1987 or 1988 data set. The final data set has 39,363 observations.

The "match probabilities" are assigned by comparing information in 1987 and 1988, following an algorithm developed by Joshua Gahm at the Bureau of Labor Statistics (document dated December 15, 1983). The algorithm penalizes matches with a change in age between 1987 and 1988 different than 1 year, with a change in race, with an unlikely change in marital status (e.g. married/separated/widowed in 1987 to never married in 1988), with a change in veteran status, or with a change in highest grade of schooling greater than 1 year. Consider a white married man age 30 in 1987 who reports nonveteran status and 12 years of schooling and who lives in a household with 4 people in 1987. A match to a married white man age 31 in 1988 with the same education and veteran status is assigned a probability of 0.49 (and is retained). A match to a man age 31 with a different race or veteran status, or an absolute change in education of 2 years, is assigned a probability of 0.16 (and is dropped).

Match rates for various groups are tabulated below:

Characteristic	Match Rate (%)
All	68.8
Age: 24–30	55.0
31–35	66.0
36–40	71.7
41–45	74.7
46–50	75.7
51–55	79.1
56–60	80.2
61–66	80.2
Race: white	69.6
nonwhite	62.7
Education: 0–11 years	67.1
12 years	69.5
13 + years	68.7
Veteran Status: veteran	74.0
nonveteran	66.0
Wage Allocation: no	69.5
yes	64.7

REFERENCES

ABOWD, JOHN, AND HENRY S. FARBER (1982): "Job Queues and the Union Status of Workers," *Industrial and Labor Relations Review*, 35, 354–367.

CHAMBERLAIN, GARY (1982): "Panel Data," in *The Handbook of Econometrics*, Volume 2, ed. by Z. Griliches and M. Intrilligator. New York: North Holland.

FREEMAN, RICHARD B. (1980): "Unionism and the Dispersion of Wages," *Industrial and Labor Relations Review*, 34, 3–23.

——— (1984): "Longitudinal Analysis of the Effects of Trade Unions," *Journal of Labor Economics*, 2, 1–26.

FREEMAN, RICHARD B., AND JAMES L. MEDOFF (1984): *What Do Unions Do*? New York: Basic Books.

JAKUBSON, GEORGE (1990): "Distinguishing Unobserved Heterogeneity and Measurement Error in Panel Estimates of the Union Wage Effect," Unpublished Manuscript, New York State School of Industrial and Labor Relations, Cornell University.

JOHNSON, GEORGE E., AND KENWOOD C. YOUMANS (1971): "Union Relative Wage Effects By Age and Education," *Industrial and Labor Relations Review*, 24, 171–179.

LEE, LUNG-FEI (1978): "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables," *International Economic Review*, 19, 415–433.

LEWIS, H. GREGG (1986): *Union Relative Wage Effects: A Survey*. Chicago: University of Chicago Press.

MELLOW, WESLEY, AND HAL SIDER (1983): "Accuracy of Response in Labor Market Surveys: Evidence and Implications," *Journal of Labor Economics*, 1, 331–344.

ROBINSON, CHRIS (1989): "The Joint Determination of Union Status and Union Wage Effects: Some Tests of Alternative Models," *Journal of Political Economy*, 97, 639–667.