Hirsch, Barry T.; Schumacher, Edward J.

Working Paper

Match Bias in Wage Gap Estimates Due to Earnings Imputation

IZA Discussion Papers, No. 783

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Hirsch, Barry T.; Schumacher, Edward J. (2003) : Match Bias in Wage Gap Estimates Due to Earnings Imputation, IZA Discussion Papers, No. 783, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/21456

DISCUSSION PAPER SERIES

IZA DP No. 783

Match Bias in Wage Gap Estimates Due to Earnings Imputation

Barry T. Hirsch
Edward J. Schumacher

May 2003

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Match Bias in Wage Gap Estimates Due to Earnings Imputation

Barry T. Hirsch
Trinity University and IZA Bonn

Edward J. Schumacher
Trinity University

Discussion Paper No. 783
May 2003

IZA

P.O. Box 7240
D-53072 Bonn
Germany

Tel.: +49-228-3894-0
Fax: +49-228-3894-210
Email: iza@iza.org

IZA Discussion Paper No. 783
May 2003

ABSTRACT

Match Bias in Wage Gap Estimates Due to Earnings Imputation*

About 30% of workers in the CPS have earnings imputed. Wage gap estimates are biased toward zero when the attribute being studied (e.g., union status) is *not* a criterion used to match donors to nonrespondents. An expression for "match bias" is derived in which attenuation equals the sum of match error rates. In practice, attenuation can be approximated by the proportion with imputed earnings. Union wage gap estimates with match bias removed are presented for 1973-2001. Estimates in recent years are biased downward 5 percentage points. Bias in gap estimates accompanying other non-match criteria (public sector, industry, etc.) is examined.

JEL Classification: J3, J5, C81

Keywords: wage differentials, hot deck imputation, match bias, CPS, union wage premiums

Corresponding author:

Barry T. Hirsch
Department of Economics
Trinity University
San Antonio, TX 78212
USA
Tel.: +1 850 644 3586
Fax: +1 850 644 4535
Email: bhirsch@trinity.edu

* The authors particularly appreciate the assistance of Anne Polivka at the Bureau of Labor Statistics, as well as comments from David Card, Christopher Carpenter, George Deltas, Simon Woodcock, and session participants at the Econometric Society and Canadian Economic Association meetings. The CPS data set was developed with David Macpherson at Florida State University.

I. Introduction

The Current Population Survey (CPS) provides the principal data source for estimates of union-nonunion wage premiums and sectoral wage differentials.[1] As widely recognized, many individuals surveyed in the CPS (and other household surveys) either refuse to report their earnings or "proxy" respondents in their household are unable to report earnings (Lillard, Smith, and Welch, 1986; Rubin, 1983).[2] Rather than compile official statistics based on large numbers of incomplete records, the Census allocates or imputes earnings for those with missing values. During the 1980s, fewer than 15% of workers in the CPS had earnings imputed. This figure rose with the 1994 change in CPS earnings questions, and has continued to increase in recent years. In 2001, 31% of all private and public sector wage and salary employees in the CPS earnings files had weekly earnings imputed by the Census.

Despite its prevalence, earnings imputation has been given relatively little attention in the large empirical literature on wage differentials, much of it based on the CPS.[3] The principal reason is that it is believed that earnings are imputed accurately on average, so that non-systematic error in the dependent variable does not bias explanatory variable coefficients. The prevailing view is stated succinctly by Angrist and Krueger (1999) in their comprehensive survey article on empirical methods. After comparing regression estimates with and without inclusion of allocated earners (and with and without weighting), the authors state: "The results in Table 12 suggest that estimates of a human capital earnings function using CPS and Census data are largely insensitive to whether or not the sample is weighted …, and whether or not observations with allocated values are included in the sample." (Angrist and Krueger, 1999, p. 1354). Interestingly, the wage equations estimated by Angrist and Krueger, using the March CPS, contained neither sectoral (industry and public sector) nor union status variables. Had these

[1] For an analysis of union wage gap studies through the early 1980s, see Lewis (1986). Recent analyses of wage gaps over time include Blanchflower (1999) and Bratsberg and Ragan (2002). Hirsch and Macpherson (2002) provide annual CPS union wage gap estimates for 1973-2001 for alternative worker and sectoral groups (industry and public/private). Frequently cited studies on interindustry wage differentials include Krueger and Summers (1988), Dickens and Katz (1987), and Gibbons and Katz (1992). Literature on wage differentials in the public sector is summarized in Gregory and Borland (1999).

[2] Groves and Couper (1998) provide an analysis of factors determining household response rates in six national surveys, each having households linked to records in the 1990 decennial census.

[3] There has been considerable attention given to mismeasurement in *reported* earnings in the CPS (Mellow and Sider, 1983; Bound and Krueger, 1991; Bollinger, 1998). These authors have been careful to delete allocated earners from their analysis. Hirsch and Schumacher (1998) omit allocated earners, noting the possible mismatch between the union status of workers and donors. Hirsch and Macpherson (2000, Appendix), find that wage differential estimates among air transport workers, particularly pilots, are sensitive to treatment of allocated earners.

variables been included, they are likely to have arrived at a different conclusion.

The Census allocates earnings using a "hot deck" imputation method that matches each nonrespondent to an individual or "donor" whose characteristics are identical. The donor's reported earnings are then assigned to the nonrespondent. Among the more important characteristics used in matching a donor to a nonrespondent are gender, age, education, and hours worked, four strong correlates of earnings. Two characteristics *not* used are union status and sector (e.g. industry) of employment.

The principal argument of this paper is straightforward. The research literature in labor economics abounds with estimates of wage differentials with respect to worker and job attributes. If the attribute under study is *not* used as a Census match criterion in selecting a donor, wage differential estimates (with or without controls) are biased toward zero. This bias is large and exists independent of any bias from the nonrandom determination of missing earnings (i.e., response bias). This paper analyzes the systematic "match bias" attaching to estimated wage differentials for attributes that are not imputation match criteria. We focus in particular on estimates of union wage premiums, with limited attention given to the estimation of industry and public sector wage differentials.

In what follows, we first discuss the imputation methods used by the Census to allocate earnings for nonrespondents. A general expression is derived that provides a measure of match bias (or attenuation) in wage gap estimates, absent covariates. This is subsequently expanded to a regression framework with covariates. Correlation between union status (in the case of union gap estimates) and the explicit match criteria improves match quality and mitigates bias in wage gaps without covariates. That same correlation can exacerbate bias in regression gap estimates accounting for covariates.

Failure to account for earnings imputation causes a substantial understatement in the union wage gap. This bias is particularly severe since 1994. Changes over time in how allocated earners are designated in the CPS have led researchers to report misleading *changes* in union wage gaps. In particular, what appears as a large and puzzling drop in the CPS union gap between 1978 and 1979 (Freeman, 1986; Lewis, 1986) is accounted for in large part by changes in the treatment of workers with allocated earners. A set of time-consistent union wage gap estimates for the 1973-2001 period indicates a pattern that differs in several respects from existing evidence. Although our emphasis is on union wage

gaps, similar match bias is found in estimates of industry, public sector, city size and region, and other differentials studied extensively in the labor economics literature.

II. Census Imputation Methods for Allocating Earnings

The Census allocates missing earnings using "hot deck" imputation methods. Most familiar to researchers is the hot deck method used to impute earnings in the March CPS Annual Demographic Files (for details, see Lillard, Smith, and Welch, 1986). Using this method, matching of a nonrespondent with a donor is done in steps, with each step involving a less detailed match requirement. For example, suppose there were just four matching variables – sex, age, education, and occupation. The matching program would first attempt to find an exact match on the combination of variables, where each is segmented at a relatively detailed level. When there is not a successful match at a given level, matching proceeds to the next step where a less detailed breakdown is used, say, broader occupations and age categories. As emphasized by Lillard, Smith, and Welch, the probability of a close match declines the less common an individual's characteristics.

Much of our current knowledge about the labor market in general, and union and nonunion wages in particular, is based on research using the CPS Outgoing Rotation Group (ORG) Earnings Files. The CPS-ORG files are made up of the quarter sample of individuals in the monthly survey asked, among other things, usual weekly earnings, hours worked, and union status.

The CPS-ORG files use an imputation procedure called the "cell hot deck" method, which differs from the method used in the March CPS. The Census creates cells based on the following seven categories: gender (2 cells), age (6), race (2), education (3), occupation (13), hours worked (8), and receipt of tips, commissions or overtime (2), a matrix of 14,976 possible combinations.[4] The Census keeps all cells "stocked" with a donor, insuring that an exact match is always found. The donor in each cell is the most recent person surveyed by the Census with reported earnings and all the characteristics. When a new person with those characteristics is surveyed and reports earnings, the Census replaces the previous occupant of the cell. To insure an occupant of each cell, the Census reaches back as far as necessary within a given survey month and then to previous months and years. When surveyed

[4] Details on the coding of variables used to form the Census cells can be provided by the authors.

individuals do not report earnings, their earnings are imputed by assigning the value of (nominal) earnings reported by the current donor occupying the cell with an exact match of characteristics.[5]

Location is not an explicit match criterion using the cell hot dock, but files are sorted by location and nonrespondents are matched to the most recent donor match (i.e., the geographically closest person moving backward in the file).[6] If matched to someone in a similarly-priced neighborhood, the donor is more likely to have earnings similar to the nonrespondent than if the match is based exclusively on the mix of attributes defining each cell. Downward bias in wage gap estimates is mitigated as the difference between reported and imputed wages shrink. Mitigation of bias from this "location effect" is likely to be very small, except for nonrespondents in highly populated cells.

Although not the focus of this paper, attention has been given in the literature to alternative imputation methods that address shortcomings in standard hot deck methods. An imputation procedure is regarded as "proper" if it restores fully the sampling variability. Single imputation procedures are "not proper" because they do not incorporate information about the uncertainty associated with the choice of the value to impute. "Multiple imputation" methods select multiple donors for each missing observation (or, stated alternatively, create multiple data sets) and permit the researcher to account for the variability associated with the assignment of an imputed value.[7]

The Census hot deck procedures assume either no response bias, or "ignorable response bias" whereby the match criteria capture differences in earnings. For example, the likelihood of nonresponse might vary with schooling, occupation, and other match attributes, but as long as the earnings of respondents and nonrespondents *within* cells are equivalent, there is no response bias resulting from the imputation procedure. "Nonignorable response bias" occurs if the earnings of donors with the same

[5] A brief discussion of Census/CPS hot deck methods is contained in the U.S. Department of Labor, 2002, p. 9.3). A more detailed description was provided by economists at the BLS and Census Bureau. Although the "cell hot deck" procedure has been used for the CPS-ORG files since their beginning in 1979, the selection categories have not been identical over time. Prior to 1994, there were 6 usual hours worked categories and thus 11,232 cells. Beginning in 1994 usual work hours could be reported as "variable." Two additional hours cells were added for workers reporting variable hours, one for those who are usually full-time and one for those usually part-time.

[6] In the March CPS, region serves as an explicit match criterion for selecting donors.

[7] Rubin (1983, 1987) has proposed multiple imputation procedures that are proper and that model the likelihood of having a missing value. Imputed values are obtained from multiple donors who have similar probabilities of being in the nonresponse group (e.g., similar "propensity scores" constructed from logit estimation). Treatment effects can be estimated using identical methods (Angrist and Krueger 1999; Heckman, LaLonde, and Smith 1999; Heckman, Ichimura, and Todd 1998).

match characteristics as nonrespondents provide a biased estimate of earnings (Rubin, 1983, 1987).

The bias examined in this paper occurs independently of whether or not there is nonignorable response bias. Even if nonrespondents are selected randomly, there will be a "match bias" toward zero in wage gap estimates associated with non-match criteria (union status, industry, public sector, etc.).

III. Imputed Earnings and Match Bias in Wage Differential Estimates

Let Γ represent the unbiased estimate of W_u-W_n, the difference in *mean* log wages between two groups *u* and *n* (union and nonunion in our example), absent covariates. The subsequent section extends the analysis to a regression framework with covariates. The analysis applies to the case where the wage differential attribute being studied is *not* a match criterion used to identify donors. Below we show conditions under which the match bias is equal to $\Omega\Gamma$, where Ω is the proportion of workers with imputed earnings. Although these conditions may not be satisfied exactly, $\Omega\Gamma$ may provide a good approximation of bias in many applications (i.e., proportionate attenuation in the wage gap is approximated by Ω).

Below we first derive the general formula for match bias absent covariates, and then show under what circumstances the bias simplifies to $\Omega\Gamma$. For the purpose of exposition, assume that there exist two groups, union and nonunion, with W_u and W_n representing unbiased measures of their mean log wages and Γ is the log wage differential. Union and nonunion nonresponse rates are designated Ω_u and Ω_n, with rates of response being $(1\text{-}\Omega_u)$ and $(1\text{-}\Omega_n)$. Let ρ_u be the proportion of union donors and $(1\text{-}\rho_u)$ the proportion of nonunion donors assigned to *union* nonrespondents. Likewise, ρ_n is the proportion of union donors and $(1\text{-}\rho_n)$ the proportion of nonunion donors assigned to *nonunion* nonrespondents.

The *measured* earnings W_u' and W_n' for union and nonunion workers (i.e., "edited" earnings) will be the weighted average of those reporting earnings and nonrespondents with imputed earnings. That is,

(1) $$W_u' = (1\text{-}\Omega_u)W_u + \Omega_u [\rho_u W_u + (1\text{-}\rho_u)W_n]$$

(2) $$W_n' = (1\text{-}\Omega_n)W_n + \Omega_n [\rho_n W_u + (1\text{-}\rho_n)W_n]$$

where the bracketed expressions are mean wages for union and nonunion workers with imputed earnings.

The measured or observed union wage gap in most empirical studies is W_u'-W_n', with match bias, *B*, being the difference between an unbiased and biased wage gap estimates, or

(3) $$B = (W_u\text{-}W_n) - (W_u'\text{-}W_n')$$

$$= W_u - W_n - [(1-\Omega_u)W_u + \Omega_u[\rho_u W_u + (1-\rho_u)W_n]] + [(1-\Omega_n)W_n + \Omega_n[\rho_n W_u + (1-\rho_n)W_n]].$$

Simplification of equation (3) yields the following general expression for the extent of match bias:

(4) $B = [(1-\rho_u)\Omega_u + \rho_n\Omega_n]\Gamma,$

where $\Gamma = W_u - W_n$. The term in brackets represents attenuation in Γ. An "attenuation coefficient" γ, with 1.0 representing no attenuation and zero complete attenuation, can be defined as

(4′) $\gamma = 1-[(1-\rho_u)\Omega_u + \rho_n\Omega_n].$

Interpretation of (4) and (4′) is straightforward. The term in brackets is the sum of mismatch rates for both groups of workers. The term $(1-\rho_u)\Omega_u$ represents the number of false negatives or, probabilistically, $Prob(u^d = 0 \mid u = 1)$, the probability of a match with a nonunion donor given that the nonrespondent is union. The term $\rho_n\Omega_n$ measures the false positive rate or $Prob(u^d = 1 \mid u = 0)$, the probability of a union earnings donor given a nonunion nonrespondent. There would be no match bias if either there were no allocated earners ($\Omega_u = \Omega_n = 0$) or no donor mismatch ($(1-\rho_u) = \rho_n = 0$).

Equation (4) can be simplified further. If the union-nonunion donor mix is identical for union and nonunion respondents, so that $\rho_u = \rho_n = \rho$, match bias is:

(5) $B = [(1-\rho)\Omega_u + \rho\Omega_n]\Gamma.$

Finally, assuming an equivalent donor mix and equal rates of nonresponse, so that $\Omega_u = \Omega_n = \Omega$, the match bias formula reduces to the simple expression:

(6) $B = \Omega\Gamma$

with the degree of attenuation equal to Ω and an attenuation coefficient

(6′) $\gamma = 1-\Omega.$

Evident from equation (5) is that bias is likely to exceed $\Omega\Gamma$ (where Ω is the full-sample nonresponse rate) if we assume the union density of donors is less than .50 (i.e., $1-\rho > \rho$) and if union workers have nonresponse rates exceeding nonunion workers. In the event that the nonresponse rate for union workers is less than for nonunion workers, bias is less than $\Omega\Gamma$.

As seen subsequently, $\Omega\Gamma$ provides a reasonable approximation of the match bias in union-nonunion wage gaps. The reasons are twofold. First, nonresponse rates are similar for union and nonunion workers. Second, although correlation between union status and the explicit match criteria acts

to mitigate bias, this is offset by increased bias within a regression framework due to inclusion of wage covariates correlated with union status as controls (see the next section).

The match bias in log wage gaps has been shown above. The upward adjustment to roughly correct for match bias is:

(7) $W_u - W_n = \Gamma = (W_u' - W_n')/(1-[(1-\rho_u)\Omega_u + \rho_n\Omega_n]) = (W_u' - W_n')/\gamma$

where the denominator γ is the attenuation coefficient (i.e., one minus the bias). For example, if 25% of individuals have their earnings imputed by the Census ($\Omega_u = \Omega_n = .25$) and the donor mixes are equal, then union gap estimates should be adjusted upward by a third (1/(1-.25) = 1.333) from, say, .15 to .20. In practice, researchers have information on Ω_u and Ω_n, but not the donor mix ρ_u and ρ_n. The latter can be self-generated, as seen subsequently, by implementing one's own imputation procedure.

To understand more fully the nature of match bias, we offer a simple example. Once again, assume equivalent rates of nonresponse and donor mix for union and nonunion respondents so that the bias formula $\Omega\Gamma$ applies. Assume that 10% of private sector workers are union members, that there is a .20 log wage gap between union and nonunion workers, and that 25% of workers in the CPS have earnings allocated, with union status not a match criterion. In selecting donors for those with missing earnings, let 10% of union nonrespondents be matched to union donors and 90% to nonunion donors. Likewise, among nonunion workers with missing earnings, let 90% be matched to nonunion donors and 10% to union donors. Union workers with imputed earnings have their earnings understated by .18 (.90 times the .20 union wage differential) so that the average of union earnings for those with and without imputed earnings is understated by .045 (.25 imputed earners times .18). Turning to nonunion workers with imputed earnings, their earnings are overstated by .02 (.10 times the .20 union differential), so the average of nonunion earnings is overstated by .005 (.25 imputed earners times .02). Taken together the measured union-nonunion wage differential is .15 rather than .20, biased downward by .05 due to the understatement of union earnings (.045) and overstatement of nonunion earnings (.005). Stated alternatively, with bias $B = \Omega\Gamma = .05$, attenuation of Γ is equal to $\Omega = .25$ and the attenuation coefficient

is $\gamma = (1-\Omega) = .75$. For the 25% of the sample with earnings imputed, there exists no union wage gap.[8]

In Table 1, we examine how sensitive is match bias, shown in equation (4), to changes in imputation rates and donor mix. The illustrative example discussed above is seen in line 1. Imputation rates of .25 for union and nonunion workers, an equal donor mix of 10% union, and an unbiased wage gap of .20 leads to downward bias of .05 log points and an attenuation coefficient of .75. Evident from lines 2-4 is that an increase (decrease) in the union relative to nonunion imputation rate increases (decreases) bias, given a union proportion in the donor mix of less than .50. In lines 5-6, the mitigating effect of a differential donor mix is seen. If union workers are matched to donors of whom 18% are union and nonunion workers are matched to 9% union workers, bias falls from .05 in line 1 to .046 in line 5. Were union workers matched to 50% union donors and nonunion workers to 3% union donors (with imputation rates of .26 and .24), bias would decline to .027 (line 7). Line 8 demonstrates that if union status is an explicit match criterion ($\rho_u = 1$, $\rho_n = 0$), there is no match bias.

Included in line 9 are the actual imputation rates in our 1996-2001 CPS sample ($\Omega_u = .265$, $\Omega_n = .257$) and the donor mix subsequently obtained using our own hot deck procedure ($\rho_u = .180$, $\rho_n = .091$). Predicted match bias is .048 and the attenuation coefficient is .759, close to the values from the simple approximation in line 1. Recall that the bias shown to this point applies to mean differences in union and nonunion log wages; that is; wage gaps absent covariates.[9]

IV. Match Bias as a Form of Measurement Error: Attenuation with Regression Covariates

In the previous section, we identified match bias absent covariates. In a wage regression including

[8] Although Card does not identify or discuss the issue of match bias, he points out in a footnote: "For simplicity, I have deleted all observations with imputed earnings data … The union wage gap for men with allocated earnings is roughly 0." (Card, 1996, p. 968, fn. 22).

[9] Although the focus here is on cross-sectional studies, similar match bias exists for longitudinal studies examining the correlation between wage change and the change in union status (or other non-match attributes). If earnings are imputed in *both* years 1 and 2, the bias can be approximated by $\Omega\Gamma$, just as in the cross sectional analysis, assuming Ω is the sample proportion with earnings imputed in both years. Among workers whose earnings are imputed in both years, there will be zero correlation between earnings change and union status change. For those whose earnings are imputed in year 1 only, the extent of bias depends on whether one is a union joiner (U_{01}) or leaver (U_{10}). Estimated wage gaps for joiners would show little bias since roughly 90% of imputed earners are correctly matched to nonunion donors in year 1. Bias would be substantial for leavers since only about 10% of imputed earners are correctly matched to union donors in year 1. If earnings are imputed in year 2 only, the opposite scenario occurs, with a substantial bias for union joiners and a minor bias for leavers. Imputation can either mitigate or exacerbate measurement error bias toward zero resulting from misclassified union status, depending on whether or not imputed earners with misclassified union status are matched to an earnings donor with the same *measured* union status. Longitudinal studies that examine misclassification bias in union status include Freeman (1984), Card (1996), and Hirsch and Schumacher (1998).

union status plus correlated covariates, identifying match bias is not straightforward. Recasting match bias as a form of misclassification or measurement error in the right-hand-side (RHS) union variable, rather than error in the left-hand-side (LHS) wage variable correlated with union status, permits us to address this issue based on existing literature.[10] Indeed, the match bias measure we provided in the previous section is equivalent to the measure given for attenuation bias due to misclassification of a binary variable (union status), absent covariates or assuming zero correlation between union status and covariates (Aigner 1973, p. 53, eq. 11; Freeman 1984, p. 8, eq. 9).

The logic is as follows. The Census hot deck procedure matches an earnings nonrespondent with an earnings donor identical with respect to match characteristics, but not non-match characteristics such as union status. Donor earnings on the LHS can be treated as a valid observation whose RHS characteristics, apart from union status (or other non-match attributes) are measured without error. The regression then includes valid donor observations whose misclassified union status produces a biased estimate of the union wage gap. Identification of measurement error bias in the union coefficient thus provides an estimate of the match bias from earnings imputation within a regression framework.[11]

The classical errors-in-variables approach assumes measurement error in an explanatory variable x_1 that is uncorrelated with its true value. Although typically a reasonable assumption for a continuous variable, it is necessarily false for binary variables. If true union status $u^* = 1$, then misclassification error in observed union status u must be negative; if $u^* = 0$ error must be positive. As shown by Aigner (1973) and others, the binary errors-in-variables case results in least squares coefficient estimates biased toward zero and which, absent additional assumptions, is difficult to identify. Bollinger (1996) has established bounds for the true coefficient, with the least squares estimate providing the lower bound and

[10] We thank Thomas Lemieux and an anonymous referee for providing this insight, along with guidance on the appropriate literature.

[11] We ignore several complications. Although most explanatory variables are included in the Census match list, there will exist measurement error among donor observations for covariates that are not match criteria (e.g., industry status). Second, donors are obtained from the CPS and may be included twice in the regression sample, first as a donor and second as a regular observation. Replicate observations will lead to downward bias in standard errors. Pairs of replicates have union misclassification error in the donor observations, but not in the paired regular observations. Third, a separate issue is the degree of measurement error in the reported union status variable, unrelated to earnings imputation. In order to focus on imputation match bias, we ignore the standard form of measurement error, which is not so large in wage level analysis. Measurement error has been a focus in longitudinal studies (Freeman 1984; Bollinger 1996; Card 1996). Farber and Western (2002) note that as union *density* falls below .50, random reporting error biases density upward, substantially so as density becomes very small (when true density is zero, all bias is positive). Bias would be in the opposite direction if union density exceeded .50.

a variant of reverse regression providing an upper bound.

Card (1996, pp. 958-60) has identified a measure of attenuation bias resulting from misclassification error in union status, a measure requiring external information on the rate of misclassification error. Card's formulation also accounts for differences in the true but unobserved union density and the observed union density (these are assumed identical in Aigner and Freeman). In the context of our match bias problem, true density can diverge from observed density if one treats the earnings sample *including donors* as the true sample. Card's approach can be readily applied here. Misclassification in union status is a function of Census earnings allocation rates, which we directly measure, and the rates at which union workers are assigned nonunion donors and nonunion workers assigned union donors, which we can approximate based on our own hot deck procedure.

Card first derives the attenuation coefficient measure γ^0 in the case where there are no covariates. Following his notation, u^* = true but unobserved union status, indicator variable u = observed union status, $q_1 = Prob(u = 1 \mid u^* = 1)$, $q_0 = Prob(u = 1 \mid u^* = 0)$, true union density $\pi = \bar{u}^*$, observed density $P = \bar{u}$ or $P = q_1\pi + q_0(1-\pi)$.[12] Card (1996, p. 959) shows

(8) $$\gamma^0 = \pi/P \cdot [(q_1-P)/(1-P)].$$

In practice, γ^0 yields values very close to our attenuation coefficient, $\gamma = 1-[(1-\rho_u)\Omega_u + \rho_n\Omega_n]$, which treats the original RHS sample as the true sample (recall that our measure γ is identical to misclassification bias shown in Aigner and Freeman).[13] Card (1996, p. 960) next shows that γ^1, the attenuation coefficient from misclassification error *with covariates*, can be approximated by

(9) $$\gamma^1 = [\gamma^0 - R^2/(q_1-q_0)] \,/\, 1-R^2$$

where γ^0 is the attenuation bias absent covariates, (q_1-q_0) is one minus the misclassification errors, and R^2 is the explained variance from a regression of observed union status on all other covariates. Using our notation, equation (9) translates to

(9′) $$\gamma^1 = [\gamma^0 - R^2/(1 - (1-\rho_u)\Omega_u - \rho_n\Omega_n)] \,/\, 1-R^2.$$

If union status is correlated with the explanatory variables (i.e., $R^2 > 0$), bias from misclassification error

[12] q_0 is the "false positive" and $(1-q_1)$ the "false negative" rate Card assumes $q_0<q_1$. Translating into our earlier notation, $(1-q_1)$ is equivalent to $(1-\rho_u)\Omega_u$ and q_0 equivalent to $\rho_n\Omega_n$.

[13] Using values in line 9 of Table 1 we obtain $\gamma = .759$. Allowing $\pi \neq P$, Card's measure produces $\gamma^0 = .756$.

is exacerbated by addition of covariates.[14] The intuition is that bias is a function of the error variance divided by the variance of observed union status, *conditional on covariates*. Earnings covariates correlated with union status reduce variance in the denominator.

In subsequent work using a pooled 1996-2001 CPS sample, we regress union status on other earnings covariates and obtain an R^2 of .1136 (much of the explanatory power results from industry, occupation, and region dummies). Utilizing the values shown in Table 1, line 9, we obtain an estimate of $\gamma^1 = .684$, as compared to $\gamma = .759$ absent covariates or $\gamma^0 = .756$ accounting for changes in the sample composition but not covariates. We later compare predicted rates of attenuation with those obtained in our empirical work. Results indicate that these match bias approximations work well in practice and that treatment of LHS imputation error as a form of RHS measurement error is a fruitful strategy.

V. Data Description and CPS Allocation Flags for Imputed Earnings

In our analysis of union wage gaps, the data sources are the May 1973 through May 1981 CPS and the CPS-ORG earnings files for 1983 through 2001. Subsequent analysis of industry and other sectoral wage differentials is based on the combined 1996-2001 CPS-ORG sample.

The CPS-ORG earnings files made available to researchers are prepared by the Census for use by the Bureau of Labor Statistics (BLS), which then makes these files available to the research community. The information provided on the BLS's CPS earnings files regarding allocated earnings has varied in important ways over time. Table 2 describes these differences and provides the percentage of earnings records identified as being allocated. Allocation rates are provided for two samples. First, figures are compiled for *all* employed wage and salary workers ages 16 and over with reported positive usual weekly earnings (weekly earnings designated as missing are retained for 1973-78). Second, allocation rates are compiled for our estimation subsample of private nonagricultural workers.

The May 1973-78 CPS earnings files formed the basis for much early research on labor unions and industry differentials, among other topics.[15] On these files, individuals who do not report earnings are included, but weekly earnings are listed as missing. Hence, researchers using the May 1973-78 CPS to

[14] Recall that correlation between union status and covariates included as Census match criteria leads to lower match error and thus decreases attenuation in the mean wage gap (i.e., an increase in γ^0).

[15] Perhaps most importantly, many of the empirical studies on unionization by Richard Freeman, James Medoff, and their students at Harvard used the May 1973-78 CPS (for a summary, see Freeman and Medoff, 1984).

estimate wage equations, knowingly or unknowingly, *exclude* allocated earners. As seen in Table 2, during 1973-78, the percentage of wage and salary workers whose weekly earnings are *missing* ranges between 18% and 22%. These are primarily workers who did not report earnings.

Beginning in 1979, imputed earnings were included in the edited earnings field, along with allocation flags designating which individuals have reported earnings and which imputed earnings. This was true for the monthly CPS-ORG files, which began in January 1979 but did not yet include union status information, and the May 1979, 1980, and 1981 CPS earnings files, which included union status.[16] The percentages designated as allocated in our 1979-81 estimation samples are 19% in the May 1979 half sample and 16% in the May 1980 and 1981 quarter samples. Allocation rates found for the full-year 1979-82 ORG files (which do not include union status) are shown in the column for all wage and salary workers (the table note contains corresponding rates from the May files). Turning to the CPS-ORG monthly earnings files for 1983-88, allocation rates were 14%-15% in most years (the exception is 1986).

Beginning in January 1989, earnings allocation flags included in the CPS-ORG are unreliable. They designate about 4% of workers as having imputed earnings, roughly a quarter of those who in fact had their earnings allocated. An alternative method exists to identify allocated earners. The ORG files during these years contain an "unedited" weekly earnings variable. Those with *missing* unedited weekly earnings (and valid edited weekly earnings) are designated as having earnings allocated. Those with non-missing unedited earnings are assumed to have been earnings respondents. This method appears to provide a reliable measure of nonresponse for most workers.[17] About 15% of workers in our 1989-92 estimation samples are designated as nonrespondents based on this method. A slightly higher rate (16.7%) is found for 1993.

Following CPS revisions in 1994, there were no usable earnings allocation flags in the ORG files for January 1994 through August 1995 (the unedited weekly earnings variable is not provided). During

[16] The May 1979 and 1980 CPS include union status information for all rotation groups, while the May 1981 CPS includes it for only the quarter sample. *Earnings* are reported for only a half sample in May 1979 and quarter samples in 1980 and 1981. There were no union questions in 1982. Union status questions were asked every month to a quarter sample (the outgoing rotation groups) beginning with the January 1983 CPS-ORG.

[17] Using unedited earnings to identify allocated earners in 1989-93 was recommended by Thomas Lemieux and David Card. We ignore the Census earnings allocation flag for 1989-93, since some workers designated as allocated have non-missing unedited and edited weekly earnings whose values are equivalent. Analysis using the 1988 ORG allows one to evaluate the method used for 1989-93, since one can compare workers identified by missing unedited weekly earnings with those designated as imputed based on the Census earnings allocation flag.

September 1995, an accurate allocation flag for the usual weekly earnings variable was included. For the period September 1995 through 1998, 22%-24% of individuals had imputed earnings. The substantial increase in earnings allocation from about 17% in 1993 to 22%-24% in 1995-98 is likely to be the result of changes in the CPS. The series of questions used by the Census to form the "edited" usual weekly earnings field became more complex following the 1994 CPS redesign (Polivka and Rothgeb, 1993). If a response is missing or replaced on any part of the sequence of questions, the Census utilizes its imputation procedure. Although procedures have been consistent since 1994, there has been a clear and worrisome increase in nonresponse since 1998, with the nonresponse rate in 2001 being 31%.

Table 3 uses the 1996-2001 CPS sample to compare characteristics of private sector wage and salary workers with and without allocated (imputed) earnings. For the most part, nonrespondents tend to be similar to respondents among measurable attributes. Allocated earners tend to be a little older, more likely in the largest cities, and more likely to be black and full time. As expected, nonresponse to the earnings question is higher when another household member (a proxy) is interviewed, providing a possible instrument for nonreporting in attempts to account for response bias (Bishop et al. 1999).

The attribute of most concern to us is union status. Union density is 9.5% among respondents versus 9.8% among nonrespondents. Among union members, 26.5% have their earnings imputed, compared to 25.7% of nonunion workers. Although these differences are small, the higher nonresponse rate among union than nonunion workers increases match bias slightly. In general, the similarity in measured characteristics among respondents and nonrespondents suggests that there may be little bias in earnings function parameters attaching to those attributes *included* as imputation match criteria. This is effectively the result reported by Angrist and Krueger (1999) based on wage regressions from the March CPS with and without inclusion of allocated earners.

VI. Union Wage Gap Estimates With and Without Match Bias Correction, 1973-2001

This section compares standard estimates of union wage gaps, with and without "correction" for match bias. This comparison provides insight into what has been a puzzle regarding changes in the union premium in the late 1970s and early 1980s and into the magnitude of recent declines in the premium. Table 4 provides estimates of union-nonunion log wage gaps for all private sector nonagricultural wage

and salary workers, with and without control for standard CPS worker and job characteristics, and with and without inclusion of workers whose earnings are imputed. The union wage gap without covariates is the difference in mean log wages for union and nonunion workers. The union gap with covariates is the coefficient on a union membership dummy variable from a log wage equation with inclusion of standard control variables. The regression estimates in Table 4 are also shown in Figure 1, along with the proportions of workers whose earnings are missing (1973-78) or allocated (beginning in 1979).

Hourly earnings are defined as usual weekly earnings divided by usual hours worked per week. Top coded earnings (at $999 in 1973-88, $1,923 during 1989-97, and $2,885 in 1998-2001) are assigned the mean above the cap based on the assumption that the upper tail of the earnings distribution follows a Pareto distribution.[18] Omitted for quality control are a small number of workers with implicit wages less than $3.00 and greater than $150 (in 2001 dollars). Controls included are years of schooling, potential experience and its square (interacted with gender), dummy variables for gender, race and ethnicity (3), marital status (2), part-time status, region (8), large metropolitan area, industry (8), and occupation (12).[19]

In what follows, we characterize estimates from the full sample, including allocated earners, as "not corrected" for match bias. Estimates from samples in which allocated earners are excluded (in those years possible) are characterized as "corrected". For the years 1994-95, where allocated earners cannot be identified, we provide estimates of what the union gap would be were it estimated for a sample with allocated earners excluded. This is done by adjusting the 1994-95 gaps upward by .043 log points, the average difference for 1996-98 in estimates with and without allocated earners.[20] In a later section, wage gap estimates are presented in which the full sample is included, but with nonrespondent earnings assigned by us using hot-deck imputation in which union status is an explicit match criterion. Following

[18] Estimates of gender-specific means above the cap for 1973-2001 are shown in Hirsch and Macpherson (2002, p. 6). These values are approximately 1.5 times the cap, with somewhat smaller female than male means and modest growth over time. For observations with non-positive usual hours worked per week or "variable hours" after 1994, we use hours worked the previous week. Absent information on hours worked, observations are dropped.

[19] Ignored are issues such as specification, the endogeneity of union status, differences between nonmember covered and not covered, unmeasured worker and job attributes, and employer-employee selection on skills and tastes (e.g., Card, 1996; Hirsch and Schumacher, 1998).

[20] In an earlier version of the paper, prior to our identifying nonrespondents in 1989-93 (see footnote 16), "corrected" regression estimates for 1989-93 were obtained by adjusting upward the "not corrected" gap estimates by .031 log points, the 1983-88 average difference between estimates including and excluding allocated earners. For the years 1973-78, it is possible to approximate what estimated union gaps would be had the sample included those with imputed earnings. This was done in the earlier version by subtracting .033 log points from the "corrected" 1973-78 regression estimates, .033 being the average difference in the two series during 1979-81.

that analysis, we conclude that estimating union gaps based on the respondent sample only (i.e., excluding allocated earners) is a simple and reasonable approach, but that it fails to *fully* account for the understatement in relative union wages owing to match bias. In addition, both the "corrected" and "not corrected" sets of estimates may contain some unknown degree of noningnorable response bias.[21]

Our analysis helps resolve what has long been a puzzle in the literature – the large decline in estimated union wage gaps between 1978 and 1979 (Freeman, 1986; Lewis, 1986). Researchers including all valid earnings records have unknowingly excluded nonrespondents during May 1973-78, but included them in years since 1979. For example, using the standard approach, our estimates indicate a 6 percentage point decline in the private sector union gap between those years, from .205 in 1978 to .148 in 1979 (the dotted line in Figure 1). Exclusion of allocated earners in 1979 eliminates match bias and produces time consistent estimates between 1973-78 and later years. We obtain an estimate for May 1979 of .180, a more modest .025 decline from the .205 estimate for 1978. Although these results are not entirely consistent with changes seen in contract data (Freeman 1986), any remaining discrepancy can be readily reconciled by the relatively small sizes or possible non-representativeness of the May samples (for further attempts at explanation, see Freeman 1986). The corrected wage gap pattern seen in Figure 1 does make economic sense, since 1979 was a period with much unanticipated inflation and contractual union wages may not have adjusted upward so quickly as did nonunion wages.[22]

Although there is a general consensus that union wage effects rose in the mid- and late-1970s (at least through 1978), there is disagreement over whether union wage premiums were maintained in the early 1980s and whether or not premiums have declined in recent years. As seen in Figure 1, the "corrected" union wage gap series indicates clearly higher premiums in 1983-85 than in 1977-78. This pattern is consistent with contract data (Freeman 1986) and evidence from the BLS Employment Cost Index (Hirsch, Macpherson, and Schumacher 2003). Such a conclusion would not follow using uncorrected CPS estimates, since one obtains wage gaps for 1973-78 similar to the "squares" in Figure 1,

[21] If the earnings of nonrespondents differ in ways not accounted for by measurable variables, *neither* sample contains accurate information on the earnings of nonrespondents, the one sample omitting nonrespondents and the other matching them to donors that differ from them in an unknown manner. As stated previously, the match bias considered in the paper exists even if nonresponse is random.

[22] Inflation during 1979 (December 1978 to December 1979) was 13.3%, as measured by the CPI-U. COLAs, when used, did not provide full adjustment for inflation.

followed by a drop down to the "diamonds" for 1979 forward.

More recently, the *uncorrected* CPS data indicate a sizable decline in the gap since the early 1990s. For example, the premium between 1993 and 2001 drops by .052 log points, from .180 to .128. Part of this decline, however, reflects recent increases in the proportion of allocated earners. During the same 1993-2001 period, the corrected series declines .035 log points, from .217 to .182. Hence, use of the uncorrected CPS will cause researchers to overstate the decline in relative union-nonunion wages. As seen in Table 4, CPS wage gaps *absent controls* indicate an even larger closing of union wage differentials than do the regression results.[23] Such closing is similar to that seen in the BLS Employment Cost Index (ECI), although the ECI has fixed industry-occupation weights. The BLS Employer Costs for Employee Compensation (ECEC), which uses current weights, surprisingly reveals no pattern in relative union wages. Reconciliation of CPS, ECI and ECEC differences is beyond the scope of this paper.[24]

Absent correction for imputation match bias, one would not only overstate the decline in the union premium since 1994, but also find it difficult to distinguish between real year-to-year changes in the union gap and variation due to changes in the number and composition of allocated earners. For example, the full sample union gap drops sharply between 1998-99 and then shows little change in the next year. But the initial drop was the result of the large increase in allocated earners between 1998-99. The corrected wage gap series shows little change in the wage gap in 1998-99 and a modest decline the following year. Similarly, an increase in allocated earners in 2001 causes the uncorrected CPS series to overstate decline in the union wage gap.

VII. Predicted versus Observed Attenuation in Wage Gap Estimates

Match bias resulting from imputed earnings is readily evident in both the unadjusted and regression-based union wage gaps. Prior to 1994, the regression gaps appear biased downward by about .03 log points – an average .031 during 1983-88 and .033 during 1989-93. Since 1994, inclusion of allocated earners causes a more substantial understatement in union gaps, an average .043 log points in

[23] In January issues of *Employment and Earnings* BLS publishes median weekly earnings for union and nonunion full time workers, based on the CPS-ORG files. For the reasons outlined in this paper, these figures should understate union-nonunion earnings differences.

[24] Hirsch, Macpherson, and Schumacher (2003) attempt to reconcile union-nonunion wage growth patterns in the CPS, ECI, and ECEC. They uncover numerous puzzles, but find few solutions. They have the most confidence in CPS estimates.

1996-98 and .054 in 1999-2001. The large bias in recent years is in line with expectations, given the increase in the proportion of allocated earners (see the triangles in Figure 1).

We earlier presented match bias and attenuation coefficient measures for cases absent covariates (equation 4′) and with covariates (equation 9′). Both measures require information not generally available to researchers (i.e., rates of donor mismatch). We also suggested (see equations 6 and 6′) that $\Omega\Gamma$ might provide a rough approximation of the match bias associated with wage gap estimates or, equivalently, that the attenuation coefficient can be approximated by $\gamma \approx (1-\Omega)$. Does $(1-\Omega)$ provide a good approximation of match bias attenuation? This cannot be answered definitively since Γ, the unbiased union gap, is not known precisely if there is nonignorable response bias or if union gaps differ for the included and excluded samples. We can assess, however, whether the ratio of the union gaps with and without allocated earners is roughly equal to $(1-\Omega)$. For example, in 1983 the proportion imputed is .138, so $1-\Omega = .862$. The ratio of the 1983 regression wage gaps is .194/.227 = .853, very close to $1-\Omega$. In 2001, with allocation rate $\Omega = .305$ and $1-\Omega = .695$, we obtain a wage ratio .128/.182 = .703, again nearly identical to $1-\Omega$. Figure 2 compares the predicted and observed rates of attenuation by year over the entire 1979-2001 period (ignoring 1994 and using September-December 1995 results). As seen in Figure 2, the simple attenuation estimate $1-\Omega$ tracks observed attenuation remarkably well. It has a mean absolute deviation from the annual wage gap ratio of only .013.[25]

We also calculated predicted attenuation rates from equation 4′, based on union-specific imputation rates and estimated rates of donor match based on our own hot-deck procedure.[26] The mean absolute prediction error using the equation 4′ measure was .019, a little larger than that seen from equation 6′. Equation 4′ tended to slightly under-predict match bias (average deviation of the attenuation coefficient was .017), as should be expected from *both* equations 4′ and 6′.

What makes these results intriguing is that regression match bias, *given covariates*, ought to have been larger than Ω. The puzzle is why $1-\Omega$ tracks closely the observed attenuation, rather than being

[25] Note that the deviation in predicted and observed attenuation is *not* the difference in the log wage gap. If the wage gap absent attenuation were .20, then the prediction error would be .013 times .20, or .0026 log points. In contrast to the mean absolute deviation, the mean deviation is effectively zero (-.0004).

[26] Union imputation rates are similar or slightly higher than nonunion rates. Based on our hot deck procedure, match of a union donor to union nonrespondents (ρ_u) is about twice as likely as a union match to nonunion nonrespondents (ρ_n). Both rates fell steadily between 1979-81 and 2001, ρ_u from .344 to .160 and ρ_n from .173 to .088.

systematically too high (i.e., understating bias). Figure 2 shows the predicted attenuation given regression covariates, based on Card's (1996) measurement error attenuation coefficient γ^1 (equation 9′). These are calculated based on union-specific imputation rates, estimates of donor match, and the R^2s from regressions of union status on other covariates (all components of 4′, 6′, and 9′ are available on request). As seen in Figure 2, annual measures of γ^1, which should best predict attenuation, instead overstate observed differences in union gaps estimated from samples with and without allocated earners. The mean absolute difference between observed attenuation and that predicted from equation 9′ is an average .046 over the 1979-2001 period, substantially larger than the average error of .013 based on 1-Ω.

Additional information is needed to sort out this puzzle. Is the attenuation coefficient γ^1 from Card a poor predictor of true match bias? Are our estimates of γ^1 too high owing to differences in the actual but unreported union status of donors in the CPS and our estimates of donor mismatch rates? Or are the measures of predicted attenuation based on γ^1 approximately correct, implying that actual match bias is even greater than the observed difference between the estimates from samples with and without allocated earners? In the next section, we conduct our own imputation procedure, which permits us to draw inferences regarding the reliability of the Card measure. Although we cannot answer the above questions with certainty, we conclude that the third explanation is the most likely one. True attenuation from match bias appears to be larger than the observed attenuation based on samples with and without allocated earners. The sample of allocated earners, omitted from the corrected sample, may have an unobserved union wage advantage larger than that seen for the sample that reports earnings.

VIII. Results Using Alternative Imputation Matching Criteria

This paper has argued that Census earnings imputation causes estimates of wage differentials to be biased downward when the attribute being studied is not used as an imputation match criterion. The bias is sizable for the measurement of relative union-nonunion earnings, causing recent union wage gaps to be understated by at least 5 percentage points. In this section, an alternative hot deck imputation procedure is implemented. Instead of using Census imputation values, wage values are obtained using simple hot deck matching methods with and without union status as a match criterion. The purpose of this exercise is threefold. First, it demonstrates whether the large discrepancy between estimated wage gaps with and

without the inclusion of allocated earners is in fact the result of union status being excluded as a match criterion. Second, union gap estimates obtained when allocated earners are excluded can be compared with results obtained for the full sample using an imputation method with union as a match criterion. Third, information is gathered on the mix of donors matched to union and nonunion nonrespondents, which allows us to calculate and compare expected and observed attenuation.

A cell hot deck imputation procedure is used. It includes 240 cells classified by gender (2 groups), age (4), education (3), occupation (5), and full or part time status (2). These 240 classifications are less detailed than the Census hot deck procedure using 14,976 cells. The limited number of cells insures that a match for all nonrespondents is found, it eases the computational burden, and permits the use of multiple imputation since there are generally many possible donors in a cell. Our program assigns a log real wage to nonrespondents (i.e., those whose earnings have been imputed by the Census) by randomly selecting a donor from among all those with exactly the same combination of characteristics. In order to account for variability in match values, 50 imputation rounds for each individual, with replacement, are performed. A second hot deck imputation procedure is then performed, identical to that described above except that it adds union membership as a match criterion, thus resulting in 480 combinations or cells. Union wage gaps using the 50 alternative data sets are then estimated. Reported in Table 5 are the union coefficient estimates based on the first round of hot decking (and its standard error), as well as the mean and standard deviation of the 50 coefficient estimates.

Results are summarized in Table 5. We use data from a sample of private sector workers in the 1996-2001 CPS-ORG files (n = 719,632; see the table note for a list of control variables).[27] Based on the Census imputation procedure, we obtain a "full sample" union log wage gap of .145. When we exclude the 25.8% of the sample with allocated earnings, the estimated wage gap rises .05 log points to .193. When we use our own multiple hot deck procedure, without union status as a match criterion, we obtain a log gap of .141, very close to the value obtained based on the more detailed Census procedure. When we add union status as a match criterion (i.e., move from 240 to 480 match cells), a (nearly) full sample

[27] Results in this section are presented based on multiple hot deck imputation using the *pooled* 1996-2001 sample. In the previous section, predicted attenuation rates (Figure 2) are based on donor match rates (ρ_u and ρ_n) obtained from a single hot deck imputation *by year* for 1983-93, September-December 1995, and 1996-2001. Because of small samples, donor rates for May 1979-81 are based on imputation using a combined sample.

union log wage gap of .208 results, somewhat higher than the .193 gap obtained by simply omitting allocated earners from the sample. For estimates with and without union status as a match criterion, the estimated gap from round 1 is similar to the mean union gap across the 50 rounds, reflecting little variation across the 50 sets of earnings data.[28]

Using our own imputation scheme provides us with information on the union status of donors and the extent to which match bias is mitigated. For the first round of the analysis (corresponding to the point estimate shown in Table 5), nonunion nonrespondents are matched to donors who are 9.1% union, while union nonrespondents are matched to 18.0% union (the sample mean among respondents or potential donors is 9.5% union). As seen in Table 1, line 9, the predicted attenuation coefficient, absent covariates, is $\gamma = .759$ (equation 4′); with covariates it is $\gamma^1 = .684$ (equation 9′).

Observed attenuation is found to be nearly identical to that predicted. Focusing on *unadjusted* wage gaps using our own imputation procedures (results not shown), we find a union-nonunion gap of .203 without union status as a match criterion and a gap of .265 with union as a match criterion. The implied attenuation coefficient is .2025/.2648 = .765, highly similar to the predicted .759 attenuation. This compares to an attenuation coefficient of .2105/.2657 = .792 based on unadjusted union gaps from the Census samples with and without allocated earners.

The regression attenuation seen in the bottom half of Table 5 is .1410/.2081 = .678, again nearly identical to the $\gamma^1 = .684$ predicted based on observed donor match rates and the correlation between union density and the regression covariates. In short, the Card measure of attenuation (equation 9′) predicts extremely well, when evaluated in an appropriate manner (i.e., based on alternative estimates from the full sample with and without "measurement error"). We observe less attenuation, .1452/.1928 = .753, using the CPS samples with and without allocated earners.

The results shown in Table 5 confirm that it is the exclusion of union status as a match criterion that accounts for the large difference in union gap estimates between the samples with and without the inclusion of allocated earners. Similar results using the Census procedure and our hot deck procedure

[28] R^2 values in Table 5 are as expected. The full sample with Census imputation (top left panel) yields a slightly higher R^2 than our hot deck method without union as a match criterion, but less than our method with union as a criterion. All of the full-sample R^2s are less than the R^2 with allocated earners excluded from the sample.

(the left column of Table 5) suggests that our simple imputation method provides matches roughly similar to those in the CPS and yields meaningful information on match donor rates ρ_u and ρ_n, information not available from the Census.

The question remains why the γ^1 measure of attenuation, which accounts for presence of covariates, predicts observed attenuation in samples with and without allocated earners far less well than does the simple approximation 1-Ω (see Figure 2). Based on results in Table 5, we concluded that γ^1 is a reliable measure of attenuation when tested correctly. Thus, two possible explanations remain. Part of the explanation could reflect more precise donor match rates using Census matching than with our simpler method (i.e., a higher likelihood of matching a union donor to a union nonrespondent and vice-versa). Reasonable changes in the values of ρ_u and ρ_n, however, have too small an impact on γ^1 to account for the difference. The alternative explanation is that nonrespondents in the CPS are not fully representative, with there being a higher than average (but unobserved) union wage gap among the omitted sample of allocated earners than among the observable sample of respondents.[29]

In short, we have found that the simple method of omitting allocated earners from the estimation sample provides a "reasonable" although not ideal approximation of a wage gap estimate purged of match bias. Observed attenuation based on samples with and without allocated earners is closely approximated by the aggregate rate of imputation (in the case of the 1996-2001 sample, Ω = .258 or γ = 1-Ω = .742). True attenuation, at least in the case of union wage gaps, is somewhat larger. Whether or not the same conclusion would apply to other wage determinants is not examined here. The analysis in this section simply reinforces the central point of the paper. Census earnings imputation causes a substantial attenuation in wage gap estimates for non-match criteria. Match bias warrants attention from researchers conducting empirical analysis.

[29] A separate issue is the possibility of response bias. We estimate a Heckman selection model in which earnings response (i.e., no imputation) is a function of all wage determinants, plus the *Proxy* variable designating if another household member provided survey information. *Proxy* is a highly important determinant of nonresponse, but a relatively unimportant determinant of the wage (*Proxy* has a -.02 coefficient in the wage equation). In principle, the selection model should account for response bias. The union gap from the selectivity-adjusted MLE wage equation is .1916 (.0020), as compared to .1928 (.0020) obtained by OLS for the sample excluding allocated earners and to .2081 (.0018) obtained with the full sample using our own hot deck imputation with union as a match criterion. Bishop et al. (1999) have used proxy status as an instrument for earnings imputation in order to measure the response bias in earnings using the March CPS. They do not consider the match bias that is the focus of our paper.

IX. Sectoral Wage Differentials

Neither industry of employment nor class of worker (i.e., private, federal, state, and local) is used as a match criterion in the Census cell hot deck earnings imputation procedure. Thus, estimates of wage differences across employment sectors should be biased downward, in a manner similar to that seen for union status. Below, we briefly examine industry wage dispersion, public sector wage differentials, and wage differentials associated with other non-match attributes.

We do not attempt here to explore the source of industry differences in earnings. Our own reading of the literature suggests that much of the dispersion in industry wages reflects the matching of highly skilled workers to high productivity and high wage workplaces. That being said, large wage differences across industries show up in cross-sectional wage regressions with standard and augmented sets of control variables, and longitudinal analysis finds individual wage changes associated with changes in industry.[30] Regardless of one's interpretation of the evidence, the size of measured industry wage differentials is understated by the inclusion of workers with imputed earnings.

We estimate log wage equations using the CPS-ORG files for 1996-2001, with and without the inclusion of allocated earners. We include a similar set of controls as in the union analysis, except that 27 industry dummies are now included, with mining the reference group (full results are available on request). We measure dispersion in log wages across the 28 industries (with a zero base for mining) by, alternatively, the standard deviation and the mean absolute deviation.

As expected, measured dispersion across industries is substantially higher when allocated earners are excluded than when included. As seen in Table 6, the standard deviation is .131 with allocated earners excluded, compared to .103 for the full sample, an understatement in dispersion of 21.4%. A similar result is found using the mean absolute deviation, .096 with allocated earners excluded versus .076 with them included, a 20.8% understatement. In results not shown, a similar pattern is found during earlier years, although bias is less severe owing to a lower proportion of allocated earnings records.

Table 6 also provides estimates of public sector differentials, comparing non-postal federal, postal, state, and local worker wages to those for workers with similar measured characteristics across the entire

[30] Among the articles in this literature, see Dickens and Katz (1987), Krueger and Summers (1988), Helwege (1992), Gibbons and Katz (1992), and Kim (1998).

private sector.[31] In each case, estimated gaps including allocated earners are biased toward zero. The column labeled "Observed Attenuation" provides the ratio of the full-sample result to the result with allocated earners omitted. Bias in the postal-private gap estimate is particularly large, with a log differential of .255 obtained among those reporting earnings, versus a biased measure of .188 from a sample of workers including allocated earners.

In addition to industry and public sector differentials, Table 6 provides selected wage differentials estimated with and without inclusion of allocated earners. In every case, differentials with respect to attributes not used as a Census match criterion are biased toward zero when allocated earners are included in the estimation sample. As seen in Table 6, this applies to such wage correlates as Hispanic, marital status, veteran status, foreign born, and city size.

X. Conclusion and Implications

Researchers have not given sufficient attention to what can be substantial bias in wage gap estimates owing to earnings imputation by the Census. The "match bias" identified in this paper is *not* the result of response bias, nor is it related to improper accounting for the uncertainty of imputed values. It does not reflect a deficiency in hot deck methods vis-à-vis alternative imputation approaches (e.g., propensity score matching and multiple imputation). Rather, for an attribute not used as an imputation match criterion (or, more broadly, not used to predict a missing wage), wage gap estimates with respect to that attribute are systematically biased toward zero. Absent covariates, the match bias is equal to the sum of match error rates (false negatives plus false positives). Bias is exacerbated in a regression framework with covariates, unless the covariates are uncorrelated with the attribute under study.

In practice, attenuation from match bias can be *roughly* approximated by Ω, the proportion of allocated earners. In 2001, over 30% of wage and salary workers had weekly earnings imputed by the Census. Excluding allocated earners from estimation samples appears to provide a simple and reasonable approach to estimating wage gaps for non-match attributes. Where precision is important, further analysis is warranted. In the case of the union wage premium, for example, match bias exceeds that

[31] In the regressions estimating public sector differentials, we omit controls for union status and industry, effecting a comparison of public workers with union and nonunion private workers across all industries. This approach comports well with comparability laws mandating public sector compensation equivalent to that for similar levels of work in the private sector. Discussion of the issues involved is contained in Hirsch, Wachter, and Gillula (1999).

observed by comparing the results from estimation samples with and without allocated earners.

We have shown that bias from imputed earnings in the estimation of union and sectoral wage gaps is considerable. The analysis applies to other wage characteristics studied in the literature that are not Census imputation match criteria. Although not exhaustive, a list of CPS-ORG wage gap estimates affected by match bias includes ethnicity, immigrant status, marital status, presence and number of children, not-for-profit employment, veteran status, and city size. In each case, differentials are understated when allocated earners are included. A similar argument applies to supplements attached to the CPS, which permit study of wage gaps with respect to company tenure, employer size, job training, displacement, and shift work. As discussed briefly, match bias affects longitudinal estimates, albeit in a more complex manner. Earnings imputation should affect measurement of both the level and trend in earnings dispersion, as well as union effects on inequality. Although this paper has focused on the cell hot deck procedure used in the monthly Census earnings files, a similar (but more complex) match bias exists using the March CPS. And earnings imputation is not limited to the CPS, being used in the NLS surveys, PSID, SIPP, and other household surveys.

It is worth emphasizing that this paper is *not* intended as criticism of Census imputation procedures. We do not argue that the Census should necessarily use industry, union status, or other attributes as match criteria in their hot deck procedure. The Census match variables are firmly based on a supply-side explanation for earnings determination, including demographic, hours, and human capital (schooling, age, occupation) variables. There is a nontrivial cost to adding variables (cells) to the match procedure, with cell sizes becoming smaller and a declining probability of finding a recent donor, let alone a donor living nearby. This is particularly true for union status since a small proportion of private employees are members. Alternatives to hot deck matching that incorporate a larger number of attributes might well be preferable, but analysis of alternative procedures lies beyond the scope of our paper.[32]

A principal implication of this paper is that *researchers* need to pay close attention to how wage differential estimates are affected by the presence of records with imputed earnings. A few researchers

[32] Lillard et al. (1986) argue that missing earnings should be predicted with an explicit regression model, with addition of a random term to restore variance. Given adequate survey information, researchers can implement their preferred imputation procedure.

ignore allocated earners because they are unaware of their presence. Most researchers, however, are aware, but see little cause for concern. The prevailing view is that as long as the Census does a good job imputing earnings *on average*, random individual error on the dependent variable does not bias coefficient estimates. In practice, including or excluding allocated earners has not appeared to make much difference, resulting in highly similar coefficients on schooling, potential experience (i.e., age and schooling), and other variables that are explicit match criteria (Angrist and Krueger, 1999).

As this paper has shown, error in imputing individual earnings is not random, being correlated with earnings attributes not used as match criteria. The extent of match bias is proportional to the share of the sample with imputed earnings. Substantial match bias is found for union and sectoral wage gaps, not surprising given that about 30% of earnings records in the CPS currently contain imputed values. The Census might best improve the quality of research, first, by insuring that reliable allocation flags are provided with publicly available data sources and, second, by providing more information to the research community on the match criteria and methods by which earnings are imputed.

References

Aigner, Dennis J. "Regression with a Binary Independent Variable Subject to Errors of Observation." *Journal of Econometrics* 1 (1973): 49-59.

Angrist, Joshua D. and Alan B. Krueger. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics, Vol. 3A*, edited by Orley C. Ashenfelter and David Card. Amsterdam: Elsevier, 1999.

Bishop, John A., John P. Formby, and Paul D. Thistle. "Mitigating Earnings Imputation Bias: Evidence from the CPS." Unpublished manuscript, 1999.

Blanchflower, David. "Changes Over Time in Union Relative Wage Effects in Great Britain and the United States." In *The History and Practice of Economic: Essays in Honour of Bernard Corry and Maurice Peston, Vol. 2*, edited by Sami Daniel, Philip Arestis and John Grahl. Northampton, Mass.: Edward Elgar, 1999.

Bollinger, Christopher R. "Bounding Mean Regressions when a Binary Regressor is Mismeasured." *Journal of Econometrics* 73 (August 1996): 387-99.

———. "Measurement Error in the Current Population Survey." *Journal of Labor Economics* 16 (July 1998): 576-94.

Bound, John and Alan B. Krueger. "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics* 9 (January 1991): 1-24.

Bratsberg, Bernt and James F. Ragan Jr. "Changes in the Union Wage Premium by Industry–Data and Analysis." *Industrial and Labor Relations Review*, 56 (October 2002).

Card, David. "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis." *Econometrica* 64 (July 1996): 957-79.

Dickens, William T. and Lawrence F. Katz, "Inter-Industry Wage Differences and Industry Characteristics." In *Unemployment and the Structure of Labor Markets*, edited by Kevin Lang and Jonathan S. Leonard. New York: Basil Blackwell, 1987.

Farber, Henry S. and Bruce Western. "Accounting for the Decline of Unions in the Private Sector, 1973-98." In *The Future of Private Sector Unionism in the United States*, edited by James T.Bennett and Bruce E. Kaufman. Armonk, New York: M.E. Sharpe, 2002.

Freeman, Richard B. "Longitudinal Analyses of the Effects of Trade Unions." *Journal of Labor Economics* 2 (January 1984): 1-26.

Freeman, Richard B. "In Search of Union Wage Concessions in Standard Data Sets." *Industrial Relations* 25 (Spring 1986): 131-45.

Freeman, Richard B. and James L. Medoff. *What Do Unions Do*? New York: Basic Books, 1984.

Gibbons, Robert and Lawrence Katz. "Does Unmeasured Ability Explain Inter-Industry Wage Differentials?" *Review of Economic Studies* 59 (July 1992,): 515-35.

Gregory, Robert G. and Jeffrey Borland. "Recent Developments in Public Sector Markets." In *Handbook of Labor Economics, Vol. 3C*, edited by Orley C. Ashenfelter and David Card. Amsterdam: Elsevier, 1999.

Groves, Robert M. and Mick P. Couper. *Nonresponse in Household Interview Surveys*. New York: John Wiley, 1998.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65 (April 1998): 261-94.

Heckman, James, Robert LaLonde, and Jeffrey Smith. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics, Vol. 3A*, edited by Orley C. Ashenfelter and David Card. Amsterdam: Elsevier, 1999.

Helwege, Jean. "Sectoral Shifts and Interindustry Wage Differentials." *Journal of Labor Economics* 10 (January 1992): 55-84.

Hirsch, Barry T. and David A. Macpherson. "Earnings, Rents, and Competition in the Airline Labor Market." *Journal of Labor Economics* 18 (January 2000): 125-55.

———. *Union Membership and Earnings Data Book: Compilations from the Current Population Survey*. Washington D.C.: The Bureau of National Affairs, 2002.

Hirsch, Barry T., David A. Macpherson, and Edward J. Schumacher. "Measuring Union and Nonunion Wage Growth: Puzzles in Search of Solutions." In *The Changing Role of Unions: New Forms of Representation*, edited by Phani Wunnava. Armonk: NewYork: M.E. Sharpe, 2003.

Hirsch, Barry T. and Edward J. Schumacher. "Unions, Wages, and Skills." *Journal of Human Resources* 33 (Winter 1998): 201-19.

Hirsch, Barry T., Michael L. Wachter, and James W. Gillula. "Postal Service Compensation and the Comparability Standard." *Research in Labor Economics* 18 (1999): 243-79.

Kim, Dae Il. "Reinterpreting Industry Premiums: Match-Specific Productivity." *Journal of Labor Economics* 16 (July 1998): 479-504.

Krueger, Alan B. and Lawrence H. Summers. "Efficiency Wages and the Inter-industry Wage Structure." *Econometrica* 56 (March 1988): 259-93.

Lewis, H. Gregg. *Union Relative Wage Effects: A Survey*. Chicago: University of Chicago Press, 1986.

Lillard, Lee, James P. Smith, and Finis Welch. "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation." *Journal of Political Economy* 94 (June 1986): 489-506.

Mellow, Wesley and Hal Sider. "Accuracy of Response in Labor Market Surveys: Evidence and Implications." *Journal of Labor Economics* 1 (October 1983): 331-44.

Polivka, Anne E. and Jennifer M. Rothgeb, "Overhauling the Current Population Survey: Redesigning the Questionnaire," *Monthly Labor Review* 116 (September 1993): 10-28.

Rubin, Donald B. "Imputing Income in the CPS: Comments on 'Measures of Aggregate Labor Cost in the United States'." In *The Measurement of Labor Cost*, edited by Jack E. Triplett. Chicago: University of Chicago Press and National Bureau of Economic Research, 1983.

Rubin, Donald B. *Multiple Imputation for Nonresponce in Surveys*. New York: John Wiley, 1987.

U.S. Department of Labor, Bureau of Labor Statistics, *Current Population Survey: Design and Methodology, Technical Paper 63RV*, March 2002, available www.bls.census.gov/cps/tp/tp63.htm.

Table 1:
Sensitivity of Match Bias to Alternative Assumptions

Line	ρ_u	ρ_n	Ω_u	Ω_n	Γ	B	γ
1.	.10	.10	.25	.25	.20	.0500	.750
2.	.10	.10	.26	.24	.20	.0516	.742
3.	.10	.10	.30	.20	.20	.0580	.710
4.	.10	.10	.20	.30	.20	.0420	.790
5.	.18	.09	.25	.25	.20	.0455	.773
6.	.18	.09	.30	.20	.20	.0528	.736
7.	.50	.03	.26	.24	.20	.0274	.863
8.	1.00	.00	.26	.24	.20	.0000	1.00
9.	.180	.091	.265	.257	.20	.0481	.759

Note. – Match bias is calculated by $B = [(1-\rho_u)\Omega_u + \rho_n\Omega_n]\Gamma$,where ρ_u = proportion union donors assigned to union nonrespondents, ρ_n = proportion union donors assigned to nonunion nonrespondents, Ω_u = proportion of union workers with imputed earnings, Ω_n = proportion of nonunion workers with imputed earnings, and $\Gamma = W_u - W_n$, the unbiased union-nonunion log wage gap. The attenuation coefficient is calculated as $\gamma = 1-[(1-\rho_u)\Omega_u + \rho_n\Omega_n]$; The biased wage gap equals $\gamma\Gamma$. Line 9 utilizes the values of ρ and Ω obtained in subsequent analysis. B and γ are strictly valid only for mean wage gaps absent covariates. Bias is exacerbated in a wage regression framework if union status is correlated with other covariates (see text).

Table 2:
Proportion of CPS Wage and Salary Earners Designated as Allocated, by Year

Year	All W&S Employees	Private Sector Estimation Sample	Year	All W&S Employees	Private Sector Estimation Sample
1973-78: Nonrespondents included in files with missing earnings; no allocated earnings designation. Shown below is the Proportion with Missing Weekly Earnings					
1973	.181	.186	1976	.198	.203
1974	.206	.211	1977	.178	.186
1975	.179	.186	1978	.216	.223
1979-88: Nonrespondents have weekly earnings imputed. Files include valid allocation designation. Shown below is the Proportion Designated as Allocated.					
1979	.165*	.187	1984	.147	.149
1980	.158*	.163	1985	.143	.144
1981	.152*	.160	1986	.107	.108
1982	.137*	–	1987	.135	.137
1983	.138	.138	1988	.144	.147
1989-93a: Nonrespondents have weekly earnings imputed. Allocation flag identifies about ¼ of allocated earners. Shown below is the Proportion Designated as Allocated.					
1989	.037	.037	1992	.042	.042
1990	.039	.040	1993	.046	.047
1991	.044	.044			
1989-93b: Nonrespondents have weekly earnings imputed. Unedited earnings used to identify allocated earners. Shown below is the Proportion with Missing Values for Unedited Weekly Earnings.					
1989	.148	.150	1992	.153	.155
1990	.150	.154	1993	.165	.167
1991	.153	.155			
1994-1995 (Aug): Nonrespondents have weekly earnings imputed. No valid allocation designation. Shown below is the Proportion Designated as Allocated.					
1994	.000	.000	1995 (Jan-Aug)	.000	.000
1995 (Sep)-Current: Nonrespondents have weekly earnings imputed. Files include valid allocation designation. Shown below is the Proportion Designated as Allocated.					
1995 (Sep-Dec)	.233	.228	1999	.276	.273
1996	.221	.217	2000	.298	.295
1997	.222	.219	2001	.309	.305
1998	.236	.232			

Note. – Data for 1973-81 are from the May CPS Earnings Supplements. Data for 1983-2000 are from the monthly CPS-ORG earnings files. Samples of "All W&S Employees" include all employed wage and salary workers ages 16 and over with positive values for usual weekly earnings (the 1973-78 samples include those with "missing" weekly earnings). The "Private Sector Estimation Samples" correspond to analysis presented in Table 4 and Figures 1-2. Additional restrictions are that observations be private sector nonagricultural wage and salary workers, with no missing observations on control variables included in the estimated wage equation, and a real wage between $3.00 and $150. Sample sizes for "All W&S Employees" are an average 50,028 for the years 1973-78 (including those with missing earnings), 25,596 in 1979, 16,085 in 1980, 14,713 in 1981, and an average 168,311 for 1983-2001. Table 4 provides sample sizes for the estimation samples.

* The 1979-82 figures for all wage and salary workers are imputation rates in the full-year 1979-82 ORG files (these do not include union status). Rates from the May files used in the analysis are 1979=.184, 1980=.159, and 1981=.161.

Table 3:
Characteristics of CPS Respondents and Allocated Earners

Variable	Allocated Earners	Earnings Respondents
Wage (2001$)	15.86	15.36
Age	39.59	37.64
Education	13.23	13.19
Male	.535	.515
Black	.120	.079
Asian	.047	.038
Hispanic	.087	.099
Married w/ spouse	.541	.563
Separated, divorced, widowed	.163	.157
MSA, medium	.399	.422
MSA/CMSA, large	.403	.324
Foreign born	.138	.126
Part time	.142	.189
Proxy respondent	.613	.488
Union member	.098	.095
N	185,685	533,947
	Union	Nonunion
Proportion of allocated earners	.265	.257
N	68,937	650,695

Note. – Data are from the 1996-2001 monthly CPS-ORG earnings files. The sample includes 719,632 private nonagricultural employed wage and salary workers, ages 16 and over.

Table 4
Private Sector Union Log Wage Differentials, With and Without Controls and Adjustment for Imputation Match Bias, 1973-2001

	Unadjusted Wage Gaps without Controls			Regression Wage Gaps with Controls		
	Not Corrected for Match Bias	Corrected for Match Bias	Observed Attenuation	Not Corrected for Match Bias	Corrected for Match Bias	Observed Attenuation
1973	--	0.309	--	--	0.161	--
1974	--	0.310	--	--	0.161	--
1975	--	0.314	--	--	0.176	--
1976	--	0.315	--	--	0.186	--
1977	--	0.362	--	--	0.214	--
1978	--	0.357	--	--	0.205	--
1979	0.270	0.319	.848	0.148	0.180	.819
1980	0.304	0.347	.877	0.162	0.193	.838
1981	0.312	0.364	.855	0.150	0.186	.807
1982	--	--	--	--	--	--
1983	0.323	0.366	.882	0.194	0.227	.853
1984	0.321	0.365	.880	0.201	0.233	.862
1985	0.318	0.359	.887	0.197	0.231	.853
1986	0.311	0.343	.905	0.190	0.214	.889
1987	0.302	0.345	.878	0.182	0.215	.847
1988	0.296	0.336	.881	0.173	0.204	.847
1989	0.296	0.336	.882	0.188	0.218	.862
1990	0.268	0.310	.865	0.171	0.206	.833
1991	0.255	0.298	.854	0.168	0.202	.830
1992	0.254	0.293	.866	0.172	0.203	.850
1993	0.265	0.310	.854	0.180	0.217	.830
1994	0.253	0.303*	.835	0.179	0.222*	.806
1995	0.235	0.285*	.825	0.174	0.217*	.801
1995^{P}	0.227	0.280	.811	0.165	0.208	.791
1996	0.234	0.286	.817	0.166	0.211	.784
1997	0.240	0.288	.833	0.169	0.209	.808
1998	0.223	0.272	.820	0.158	0.202	.784
1999	0.201	0.265	.761	0.139	0.199	.697
2000	0.191	0.244	.782	0.137	0.186	.738
2001	0.183	0.243	.752	0.128	0.182	.703

Note. – Data for 1973-81 are from the May CPS Earnings Supplements and for 1983-2001 from the monthly CPS-ORG earnings files. There was no union status variable in 1982. The sample includes employed private sector nonagricultural wage and salary workers ages 16 and over with positive weekly earnings and non-missing data for control variables (few observations are lost). The raw wage gap is the difference in mean log wages for union and nonunion workers. The regression wage gap is the coefficient on a dummy variable for union membership in a regression where the log of hourly earnings is the dependent variable. Control variables included are years of schooling, experience and its square (allowed to vary by gender), and dummy variables for gender, race and ethnicity (3), marital status (2), part-time status, region (8), large metropolitan area, industry (8), and occupation (12). Columns labeled "Not Corrected for Match Bias" include the full sample (workers with and without earnings allocated) for the years 1979-2001. Columns labeled "Corrected for Match Bias" attempt to include only workers reporting earnings. Columns labeled "Observed Attenuation" present the ratio of the uncorrected union gap to the corrected union gap (calculated prior to rounding). All allocated earners are identified and excluded for the years 1973-88 and 1996-2001. During 1989-95 allocation flags are either unreliable (in 1989-93) or not available (1994 through August 1995). For 1989-93, allocated earners are identified by a missing unedited weekly earnings variable. For 1994-95 the corrected gap (designated by *) is adjusted upward by the bias during 1996-98 (.050 for the raw gap and .043 for the regression gap). 1995^{P} is for the *partial* year September-December. Sample sizes are an average 32,102 for 1973-78 for the samples without allocated earners; for the samples including allocated earners 20,446 in 1979, 12,804 in 1980, 11,833 in 1981, and an average 133,613 for 1983-2001. See Table 2 for the proportion of allocated earners. Standard errors for the regression estimates are an average .006 for the years 1973-78, .007 in 1979, .009 in 1980, .009 in 1981, and an average .004 for 1983-2001.

Table 5
Union Wage Gap Estimates Using Alternative Hot Deck Imputation Methods With and Without Union Status as a Match Criterion, 1996-2001

	All workers, Census hot deck method, excludes union as match criterion	Excludes workers with Census imputation
CPS Imputation:		
Coefficient	.1452	.1928
Standard error	(.0018)	(.0020)
R^2	.473	.517
N	719,632	533,947
	All workers, own hot deck method, union status excluded as match criterion	**All workers, own hot deck method, union status included as match criterion**
Multiple Imputation, 50 rounds:		
Round 1 earnings data	.1410	.2081
Round 1 standard error	(.0018)	(.0018)
R^2	.464	.471
Mean gap across 50 sets	.1396	.2073
Standard deviation of coefficients	(.0010)	(.0009)
N	719,632	719,629

Note. – Data are from the 1996-2001 monthly CPS-ORG files. All regression wage gap estimates are from an identical specification including schooling, experience and its square (allowed to vary by gender), and dummy variables for gender, race and ethnicity (3), marital status (2), part-time status, foreign born, veteran status, region (8), large metropolitan area (2), industry (8), occupation (12), and year (3). This is the same sample and specification used subsequently in Table 6. The top panel relies on the Census cell hot dock imputation method with 14,976 cells, but excluding union status as a match criterion. The sample in the left column includes allocated earners. The right column "corrects" for imputation bias by excluding allocated earners. The bottom panel relies on the authors' multiple imputation hot deck procedure described in the text. In the left column nonrespondent earnings are imputed using 240 cells, where union status is not a match criterion. In the right column, match bias is corrected by using 480 cells, with union status as a match criterion. No donors were found for three union nonrespondents. The coefficients and standard errors shown are those obtained based on earnings values from the first of 50 hot deck imputation rounds. Also presented are the means and standard deviation of the union coefficients across regressions using the 50 sets of earnings data.

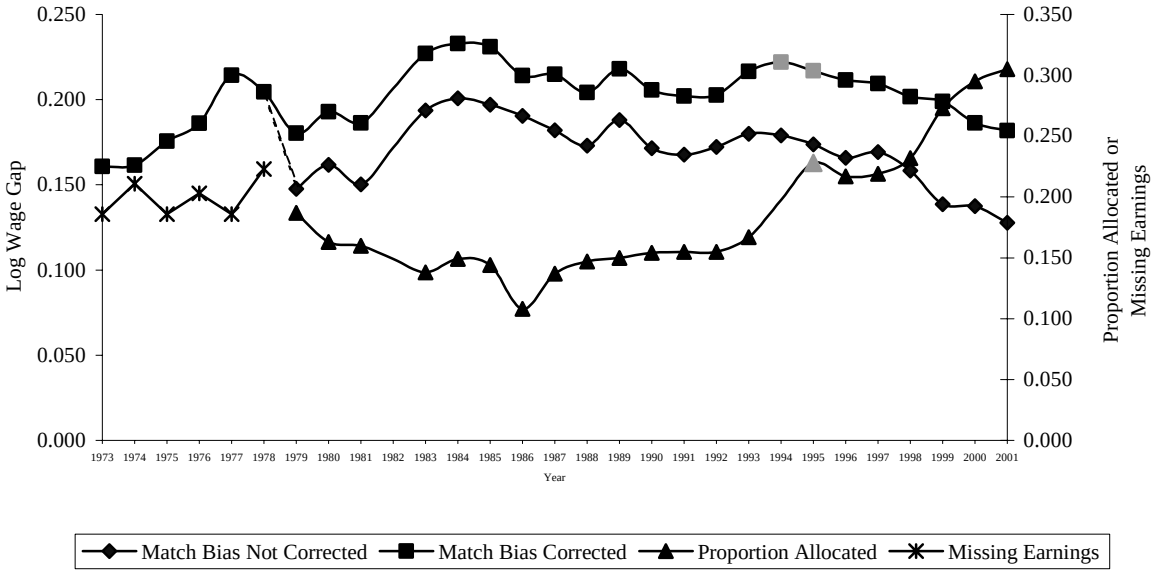
.

Table 6
The Effect of Earnings Imputation on Industry, Public Sector, and Other Selected Wage Differentials, 1996-2001

	Not Corrected for Match Bias	Corrected for Match Bias	Observed Attenuation
Industry:			
Standard Deviation	.103	.131	.784
Mean Absolute Deviation	.076	.096	.786
N	719,632	533,947	
Federal (non-postal)	.108	.143	.755
	(.003)	(.003)	
Postal	.188	.255	.739
	(.005)	(.006)	
State Government	-.036	-.043	.840
	(.002)	(.002)	
Local Government	-.032	-.040	.798
	(.002)	(.002)	
N	869,303	649,357	
Union	.145	.193	.753
	(.002)	(.002)	
Hispanic	-.089	-.107	.833
	(.002)	(.002)	
Married, Spouse Present	.083	.101	.828
	(.001)	(.002)	
Separated, Divorced, or Widowed	.034	.041	.829
	(.002)	(.002)	
Veteran	-.021	-.025	.816
	(.002)	(.002)	
Foreign Born	-.063	-.081	.776
	(.002)	(.002)	
MSA 100,000-2.5 Million	.089	.109	.823
	(.001)	(.001)	
MSA > 2.5 Million	.190	.239	.792
	(.002)	(.002)	
N	719,632	533,947	

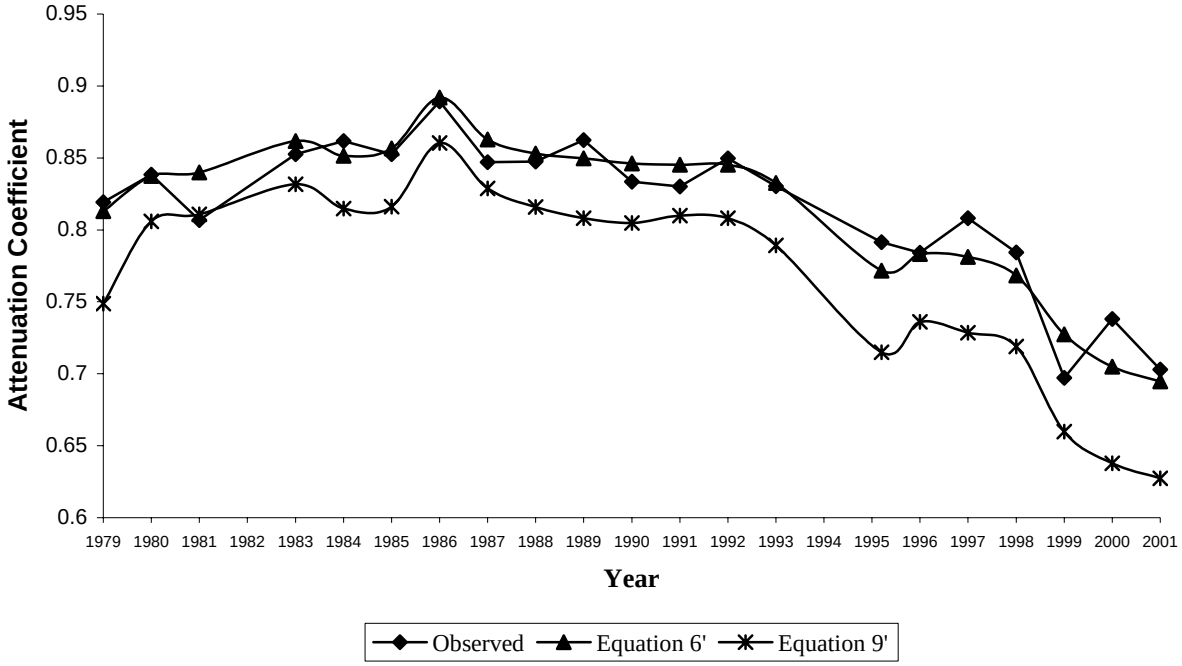
Note. – Data are from the 1996-2001 monthly CPS-ORG earnings files. Column labeled "Corrected" excludes allocated earners from the sample. "Observed attenuation" is the ratio of the "Not Corrected" to "Corrected" columns (prior to rounding). The top panel presents the dispersion in industry wages across the private nonagricultural sector. We report the unweighted standard deviations and absolute mean deviations for 28 industry classifications (i.e., the log differentials from 27 industry dummies and a zero reference group). Included variables are the same as in Table 4, except for inclusion of a more detailed industry breakdown, foreign born, veteran status, two rather than one city size dummies, and year dummies. The middle panel is based on a sample of private and public sector nonagricultural workers. The specification is the same as in the top panel, except for the inclusion of the public sector dummies and the exclusion of union status and industry dummies (see text for discussion). The bottom panel reports coefficients from the same regression used previously in the top portion of Table 5. It is identical to that in the top panel of this table, except that 8 rather than 27 industry dummies are included. Standard errors are in parentheses.

Figure 1:
Private Sector Union-Nonunion Wage Gaps and Earnings Allocation Rates



Note. – For details on estimation, see Table 4 and discussion in the text. Each wage gap series is time consistent, the "squared" line correcting approximately for match bias by omitting allocated earners, and the line with "diamonds" including allocated earners and match bias. Researchers who use all valid earnings records in CPS files would obtain wage gap estimates similar to the "squares" for 1973-78, when CPS files do not include imputed earnings, and the "diamonds" beginning in 1979, when CPS files include imputed earnings values. The 1978-79 "dotted line" connects the two series. In 1994-95, allocated earners cannot be excluded and the corrected gaps (the "light squares") are based on an approximation of the bias (see Table 4 and the text). The 1973-78 "*" series designates the proportion of the private sector estimation sample with missing earnings. The proportion of the estimation sample with earnings allocated in each year is designated by the "triangles." The allocation rate for 1989 to 1993 is defined as the proportion with a missing value for unedited weekly earnings and a valid value of edited earnings. The allocation rate for 1995 is shown for September-December only, the months with valid allocation flags. Allocation rates for 1994 and full-year 1995 are not available.

Figure 2:
Predicted and Observed Attenuation



Note. – Shown are the predicted and observed attenuation rates by year. An attenuation coefficient of 0 implies complete attenuation (bias) and a value of 1.0 implies no attenuation (bias). The observed attenuation coefficient is the ratio of the union coefficients with and without allocated earners included in the sample. Equation 6′ measures attenuation by $\gamma = 1\text{-}\Omega$, where Ω is the full sample proportion of allocated earners. Equation 9′ is adapted from Card (1996) and measures attenuation in the presence of covariates by $\gamma^1 = [\gamma^0 - R^2/(1 - (1\text{-}\rho_u)\Omega_u - \rho_n\Omega_n)] / 1\text{-}R^2$. See the text for explanation. There is no way to identify allocated earners in 1994. The observation for 1995 uses the September-December samples, which contain reliable allocation flags.

IZA Discussion Papers

No.	Author(s)	Title	Area	Date
768	J. J. Heckman S. Navarro-Lozano	Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models	6	04/03
769	L. Flood J. Hansen R. Wahlberg	Household Labor Supply and Welfare Participation in Sweden	3	04/03
770	A. Heitmueller	Coordination Failures in Network Migration	1	04/03
771	A. Calvó-Armengol Y. Zenou	Job Matching, Social Network and Word-of-Mouth Communication	5	05/03
772	E. Patacchini Y. Zenou	Search Intensity, Cost of Living and Local Labor Markets in Britain	3	05/03
773	A. Heitmueller	Job Mobility in Britain: Are the Scots Different? Evidence from the BHPS	1	05/03
774	A. Constant D. S. Massey	Labor Market Segmentation and the Earnings of German Guestworkers	1	05/03
775	J. J. Heckman L. J. Lochner P. E. Todd	Fifty Years of Mincer Earnings Regressions	5	05/03
776	L. Arranz-Aperte A. Heshmati	Determinants of Profit Sharing in the Finnish Corporate Sector	2	05/03
777	A. Falk M. Kosfeld	It's all about Connections: Evidence on Network Formation	6	05/03
778	F. Galindo-Rueda	Employer Learning and Schooling-Related Statistical Discrimination in Britain	5	05/03
779	M. Biewen	Who Are the Chronic Poor? Evidence on the Extent and the Composition of Chronic Poverty in Germany	1	05/03
780	A. Engellandt R. T. Riphahn	Temporary Contracts and Employee Effort	1	05/03
781	J. H. Abbring J. R. Campbell	A Structural Empirical Model of Firm Growth, Learning, and Survival	5	05/03
782	M. Güell B. Petrongolo	How Binding Are Legal Limits? Transitions from Temporary to Permanent Work in Spain	1	05/03
783	B. T. Hirsch E. J. Schumacher	Match Bias in Wage Gap Estimates Due to Earnings Imputation	5	05/03

An updated list of IZA Discussion Papers is available on the center‘s homepage www.iza.org.