Match Bias in Wage Gap Estimates Due to Earnings Imputation

Barry T. Hirsch
Department of Economics
Trinity University
715 Stadium Drive
San Antonio, TX 78212-7200
Voice: (210)999-8112
Fax: (210)999-7255
e-mail: bhirsch@trinity.edu
homepage: http://www.trinity.edu/bhirsch

Edward J. Schumacher
Department of Economics
East Carolina University
Greenville, NC 27858
Voice: (252)328-1083
Fax: (252)328-6743
e-mail: schumachere@mail.ecu.edu
homepage: http://www.ecu.edu/econ/faculty/schumachere

September 2001

The authors appreciate very helpful discussions with Anne Polivka of the Bureau of Labor Statistics and suggestions from Christopher Carpenter, George Deltas, Simon Woodcock, and session participants at the Econometric Society and Canadian Economic Association meetings. The CPS data sets were developed with the assistance of David Macpherson at Florida State University.

Abstract

About a quarter of the workers surveyed in the monthly Current Population Survey (CPS) have weekly earnings imputed by the Census using a "cell hot deck" method that matches nonrespondents with similar donors. Even if imputed earnings provide an unbiased measure of average earnings, wage differential estimates are systematically understated when the attribute being studied is *not* a criterion used by the Census to match donors to nonrespondents. A general expression for "match bias" is derived. This bias can be approximated by $\Omega\Gamma$, where Ω is the proportion of workers with imputed earnings and Γ is the unbiased log wage gap estimate. Union status is not a match criterion. We estimate private sector union-nonunion wage gaps using the CPS for 1973-2000. Union gap estimates understate actual wage gaps by about 5 percentage points owing to a large downward bias in imputed earnings for union nonrespondents, matched primarily to nonunion donors, and a small upward bias for nonunion nonrespondents. The sharp decline in CPS union wage gap estimates between 1978 and 1979, long a puzzle in the literature, is largely the result of inclusion of imputed earnings in CPS earnings records beginning in 1979 following their exclusion in 1973-78. Using an alternative hot deck imputation procedure in which union status is an explicit match criterion, the match bias in union wage gap estimates is removed. Other examples of wage gap bias are examined using CPS data for 1996-2000. Public sector and industry wage differentials are biased downward in a fashion similar to that seen for union wage gaps.

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. In the 2000 CPS earnings files, 30% of all private and public sector wage and salary employees 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 coefficient estimates. 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. As will be evident

[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 (2001). Hirsch and Macpherson (2001) provide annual CPS union wage gap estimates for 1973-2000 for alternative worker and sectoral groups (industry and public/private). Frequently cited studies on interindustry wage differentials include Krueger and Summers (1988) and Dickens and Katz (1987). Later studies include Gibbons and Katz (1992), Helwege (1992) and Kim (1998). Literature on wage differentials in the public sector is summarized in Gregory and Borland (1999).

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

[3] An exception is a paper by Hirsch and Macpherson (2000, Appendix), who estimate relative wages in the air transport industry before and after deregulation. Airline wage differential estimates over time are found to be sensitive to the treatment of those with imputed earnings. There has been considerable attention given to

below, had these 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.[4] This bias is large and exists independently 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 more 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 Bureau of the Census in allocating earnings for nonrespondents. A general expression is then derived that provides a measure of match bias in wage gap estimates. Subject to simplifying assumptions, match bias with respect to attributes *not* included as a Census match criterion is equal to $\Omega\Gamma$, where Ω is the proportion of workers with allocated earnings and Γ is the unbiased wage gap estimate. Absent the simplifying assumptions, the more general bias expression applies.

Empirical evidence indicates that match bias is large, with $\Omega\Gamma$ providing a close approximation. The failure to account for earnings imputation causes a substantial understatement in the magnitude of the union wage gap, and this bias is particularly severe since 1994. Moreover, changes over time in how allocated earners have been designated in the CPS have led researchers to report misleading *changes* in

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.

[4] As seen below, the bias is mitigated if selection attributes are correlated with the attribute being studied.

the union premium. In particular, what appears to be a large and puzzling drop in the CPS union premium 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 1973-2000 period indicates a pattern that differs in several respects from extant evidence. Using data from the 1996-2000 CPS, further insight into the nature of match bias is provided through estimation of union gaps based on alternative hot deck matching schemes that do and do not use union status as a match criterion. Finally, match bias similar to that seen in union wage gap estimates is found in estimates of industry, public sector, city size and region, and other earnings 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, assume 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, the 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 hot deck 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. 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 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

[5] A brief discussion of Census/CPS hot deck methods is contained in the U.S. Department of Labor, 2000, pp. 9.1-9.4). 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 (and for the May 1973-78 earnings files), 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.

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 match characteristics as nonrespondents provide a biased estimate of earnings (Rubin, 1983, 1987).[8]

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), unconditional or conditional on controls. 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 toward zero in estimating

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

[7] For a discussion of imputation issues, see Rubin (1987). Rubin (1983, 1987) has proposed multiple imputation procedures that are both proper (i.e., preserve variance to take account of the uncertainty of imputed values) and that explicitly 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. For example, one can compare the earnings of those participating in a job training program with the earnings of matched individuals not participating, but with a similar likelihood (i.e., propensity score) of having participated. For a description of matching methods for estimating treatment effects, see Angrist and Krueger (1999) and Heckman, LaLonde, and Smith (1999). Heckman, Ichimura, and Todd (1998) compare the use of propensity scores with matching methods. Propensity score methods can be advantageous where dimensionality is high (making exact matching infeasible) and one has a reliable model to predict "treatment" (e.g., earnings nonresponse). The match bias described in this paper would not be corrected through propensity score matching since nonunion donors will continue to be matched to union nonrespondents (and vice-versa).

[8] The recent Health and Retirement Study (HRS) uses a hot deck imputation procedure intended to reduce response bias. Individuals not willing to report specific income or asset values are then asked to specify the "range" of values in which their income or assets should be placed. A hot deck procedure is first used to impute dollar values to those reporting a range of values. In turn, those not reporting either dollar values or range values have their values imputed based on the "range sample" of respondents. This procedure appears to account, at least in part, for nonignorable response bias, since those refusing to report specific values but willing to report a value range had higher incomes and asset values than the average for all respondents reporting specific values. For discussion of HRS procedures, see Moon and Juster (1995, pp. S141-42) and Smith (1995, pp. S162-66).

Γ is equal to $\Omega\Gamma$, where Ω is the proportion of workers with imputed earnings. Although these conditions may not be satisfied exactly, $\Omega\Gamma$ provides a good approximation of the bias in many applications. In some cases, $\Omega\Gamma$ provides an *upper-bound* approximation of the match bias, with correlation between union status and the explicit match criteria causing bias to be mitigated. In other cases, differences in union and nonunion imputation rates can exacerbate the bias beyond $\Omega\Gamma$.

Below we first derive the general formula for match bias, 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 $(1-\Omega_u)$ and $(1-\Omega_n)$. Let ρ_u be the proportion of union donors and $(1-\rho_u)$ the proportion of nonunion donors assigned to *union* nonrespondents. Likewise, ρ_n is the proportion of union donors and $(1-\rho_n)$ 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 including respondents and nonrespondents) will be the weighted average of those reporting earnings and those with imputed earnings. That is,

(1) $W_u' = (1-\Omega_u)W_u + \Omega_u\, [\rho_u W_u + (1-\rho_u)W_n]$

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

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

The measured or observed union wage gap in most empirical studies is:

(3) $W_u' - W_n'$

with the extent of match bias, B, being the difference between an unbiased and biased wage gap estimates, or

(4) $B = (W_u - W_n) - (W_u' - W_n')$

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

Simplification of equation (4) yields the following general expression for the extent of match bias:

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

where $\Gamma = W_u - W_n$.

Interpretation of the bias expression (5) is straightforward. Match bias is equal to the sum of each group's rate of nonresponse times its rate of donor mismatch, all multiplied by the union wage gap. Absent nonresponse or donor mismatch in the event of imputation, there would be no match bias.

Equation (5) can be simplified further with certain assumptions. If the union-nonunion donor mix is identical for union and nonunion respondents, so that $\rho_u = \rho_n = \rho$, the match bias is:

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

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:

(7) $B = \Omega\Gamma.$

Evident from equation (6) 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 $\Omega\Gamma$.

The validity of the $\Omega\Gamma$ bias approximation shown in (7) depends on the bracketed terms in (1) and (2) being equivalent, and thus canceling out in (4). These terms represent the average imputed wage for those not reporting earnings. For $\Omega\Gamma$ to be correct, nonresponse rates for union and nonunion workers must be equivalent ($\Omega_u = \Omega_n$) and each group of nonrespondents must be matched to the same mix of union and nonunion donors. For the case where the characteristic (union status) is an explicit match criterion, the donor mixes are no longer equivalent, union nonrespondents being matched to union donors and vice-versa, eliminating any match bias. As seen subsequently, $\Omega\Gamma$ provides a rather close approximation of the degree of match bias in union-nonunion wage gaps. The reasons are twofold. First, nonresponse rates are similar for union and nonunion workers. Second, bias is mitigated little by the correlation between union status and explicit match criteria.

The match bias in measured estimates of log wage gaps has been shown above. This bias measure can be used to adjust upward estimated gaps to correct for the bias. In the case where match bias is approximated by $\Omega\Gamma$, uncorrected estimates can be adjusted by multiplying by $1/(1\text{-}\Omega)$. That is:

(8) $W_u - W_n = \Gamma = (W_u' - W_n')/(1\text{-}\Omega)$

For example, if 25% of individuals have their earnings imputed by the Census (Ω = .25), then estimates of the union gap that include workers with imputed earnings should be adjusted upward by a third (1/(1-.25)=1.333) from, say, .15 to .20. In the more general case, the correction for match bias is:

(9) $W_u - W_n = \Gamma = (W_u' - W_n')/[1 - (1\text{-}\rho_u)\Omega_u + \rho_n\Omega_n]$.

In practice, researchers have information on Ω_u and Ω_n, but no information about the donor mix ρ_u and ρ_n.

To provide a flavor for the nature of the bias, simple examples are helpful. 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 differential between union and nonunion workers, and that 25% of workers in the CPS have their 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% be matched with 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). Absent "mitigating factors" that would cause the donor mix to differ between union and nonunion workers, match bias is exactly $\Omega\Gamma$ or .05, given that Ω=.25 and Γ=.20. Stated alternatively, for the 25% of the sample with earnings imputed, there exists no union wage gap.

In order to provide some notion of how sensitive is match bias to alternative rates of imputation and differences in the donor mix, equation (5) is used to calculate match bias given different parameter values. Results are shown in Table 1. The illustrative example just discussed is represented in line 1. Imputation rates of .25 for union and nonunion workers, an equal donor mix of 10% union, and an unbiased log wage gap of .20 leads to downward bias of .05 log points. Evident from lines 2 and 3 is that holding all else constant, an increase the union relative to nonunion imputation rate increases bias, given that the union proportion in the donor mix is less than .50. Bias is mitigated if the nonunion imputation rate is higher (line 4). In lines 5-7, the mitigating effect of a differential donor mix is seen. If union workers are matched to donors of whom 17% are union and nonunion workers are matched to 9% union workers, the 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 for union and nonunion), the bias would decline sharply to .027 (line 8). Line 9 demonstrates that if union status is an explicit match criterion (ρ_u=1 and ρ_n=0), there is no match bias. In line 10, included are the actual imputation rates in our 1996-2000 CPS sample ($\mathbf{\Omega}_u$=.262, $\mathbf{\Omega}_n$=.256) and the donor mix subsequently obtained using our own hot deck procedure (ρ_u=.171, ρ_n=.088). Predicted match bias is .048, close to the .05 obtained from the simple approximation in line 1.

Although the focus in this paper is on cross-sectional wage measurement, a similar type of 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 $\mathbf{\Omega}\Gamma$, just as in the cross sectional analysis, assuming $\mathbf{\Omega}$ is the sample proportion with earnings imputed in both years and 1-$\mathbf{\Omega}$ is the proportion with earnings imputed in neither year. Focusing exclusively on the group whose earnings are imputed, there will be zero correlation between earnings change and union status change, since nonrespondents would be matched to, say, roughly 90% nonunion donors and 10% union donors in both years. For those whose earnings are imputed in just year 1, the degree of bias depends on whether one is a union joiner (U_1=0, U_2=1) or leaver (U_1=1, U_2=0). 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. Note that 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.[9]

IV. 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 2000. Subsequent analysis of industry and other sectoral wage differentials is based on the combined 1996-2000 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. The May 1973-78 CPS earnings files formed the basis for much early research on labor unions and industry differentials, among other topics.[10] On these files, individuals who do not report earnings are included, but weekly earnings are listed as missing. Hence, research articles using the May 1973-78 CPS, knowingly or unknowingly, *exclude* allocated earners in estimating union and sectoral (among other) wage gaps.

Table 2 presents figures on the percentage of individual earnings records in the CPS with missing earnings in 1973-78 and earnings explicitly *designated* as allocated by the Census for 1979 forward. Figures are compiled for *all* employed wage and salary workers ages 16 and over and for the private nonagricultural sector sample used in subsequent analysis. During 1973-78, the percentage of wage and salary workers whose weekly earnings are *missing* ranges between 18% and 21%. These are primarily if not exclusively workers who did not report earnings. Beginning in 1979, imputed earnings were included in the earnings variable field, along with allocation flags designating which individuals have reported

[9] Longitudinal studies that focus on bias from the misclassification of union status include Freeman (1984), Card (1996), and Hirsch and Schumacher (1998).

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.[11] The percentages allocated were 18.5% in the May 1979 half sample and 16% in the May 1980 and 1981 quarter samples.

Turning to the CPS-ORG monthly earnings files for 1983-88, earnings allocation rates were 14%-15% in most years.[12] Beginning in January 1989, earnings allocation flags included with the CPS-ORG are incomplete. They designate only about 4% of workers as having imputed earnings, roughly a quarter of those who in fact had their earnings allocated.[13] Hence, for the years 1989-93, it is impossible to identify all allocated earners. In what follows, we use information for years in which all allocated earners can be identified in order to infer the effects of imputed earnings on union wage gaps in 1989-93.

Following changes in the CPS that began in 1994, there were no usable earnings allocation flags in the ORG files for January 1994 through August 1995. Beginning September 1995, an accurate allocation flag for usual weekly earnings was included. For the period from September 1995 through 1998, 22%-24% of individuals had imputed earnings. The increase in the allocated earners share from about 14% in 1983-88 to about 25% or more in recent years is not due to a surge in nonresponse, although there has been an increase in recent years. 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.

Table 3 uses the 1996-2000 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

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

[11] 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.

[12] The exception is 1986. We have not computed the allocation rates in the full-year 1979-82 CPS-ORG files because these do not include the union status variables.

be similar to respondents, at least among measurable attributes. Allocated earners tend to be a little older, more likely in the largest cities, and more likely full time. As expected, nonresponse to the earnings question is higher when another household member (a proxy) is interviewed. The attribute most of concern to us is union status. Union density is 9.4% among respondents versus 9.7 among nonrespondents. Among union members, 26.2% have their earnings imputed, compared to 25.6% of nonunion workers. Although these differences are small, the higher nonresponse rate among union than nonunion workers will slightly increase the match bias. In general, the similarity in measured characteristics among respondents and nonrespondents suggests that there may be little difference in earnings function parameters attaching to those attributes explicitly included as match criteria for samples with and without allocated earners included (the result reported by Angrist and Krueger, 1999).

V. Union Wage Gaps

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 adjustment for the match bias associated with imputed earnings (also see Figure 1). The union wage gap without controls is the difference between the mean log wage for union members minus the mean for nonunion workers. The union wage gap with controls is the coefficient on a union membership dummy variable from a log wage equation with inclusion of standard control variables.[14] 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 1998-2000) are assigned the mean above the cap based on the assumption that the upper tail of the earnings distribution follows a Pareto distribution.[15] Controls included are years of schooling, potential experience and experience squared (interacted with gender), dummy variables for gender, race and ethnicity (3), marital status (2),

[13] An analyst at the CPS stated that those designated as allocated do in fact have an imputed earnings value, but that about three-quarters of those with imputed earnings are not designated as allocated.

[14] Ignored are issues such as specification, the endogeneity of union status, unmeasured worker and job attributes, and employer-employee selection on skills and tastes (e.g., Card, 1996; Hirsch and Schumacher, 1998).

[15] Estimates of gender-specific means above the cap for 1973-2000 are shown in Hirsch and Macpherson (2001, p. 6). These values are approximately 1.5 times the cap, with somewhat smaller female than male means and modest growth over time.

part-time status, region (8), large metropolitan area, industry (8), and occupation (12).

In what follows, we characterize the full sample, including allocated earners, as "not corrected" for match bias. We also provide a set of estimates in which all allocated earners are excluded (in those years possible) or coefficients are adjusted to approximate the effect of excluding allocated earners. These estimates are characterized as being "corrected" for match bias. Either set of estimates may contain some unknown degree of noningnorable response bias. In a later section, union wage gap estimates are presented in which the full sample is included, but earnings are assigned by us using hot-deck imputation in which union status is an explicit match criterion. Although our principal interest is focused on the regression-based union wage gaps, the "raw" or unadjusted log wage gaps show the effect of imputation prior to control for other wage and union status correlates.[16]

The downward bias expected from inclusion of imputed earnings is readily evident in both the unadjusted and regression-based union wage gaps. In 2000, the mean log wage difference is .200 with allocated earners included, but rises sharply to .255 with exclusion of allocated earners. More relevant for economists is the match bias evident in regression-based estimates. In 1983 (the first full-year ORG including union status), the regression-based private sector wage gap rises from .202 to .236 following exclusion of those with imputed earnings. In 2000, the union gap rises substantially, from .137 using standard methods to .188 following exclusion of imputed earners. Regression-based union wage gaps are biased downward by about .03 log points in years prior to 1994. Since 1994, inclusion of allocated earners causes a more substantial understatement in union wage gaps, by an average .046 log points for 1996-98 and by more than .05 in 1999 and 2000. The large bias since 1994 is in line with expectations, given the increase in the proportion of allocated earners after 1994, in particular 1999 and 2000.

For years with missing or incomplete allocation flags, we provide estimates of what the union gap would be were it estimated using only workers who report earnings. For the years 1989-93, where allocation flags are incomplete, we obtain the corrected wage gap results by adjusting upward the biased

[16] In the January issue of *Employment and Earnings* (and reprinted in the *Statistical Abstract of the United States*), the BLS publishes median usual weekly earnings for union and nonunion full time workers. Match bias in these figures is likely to be similar to that seen in our unadjusted log wage gaps.

or unadjusted gap estimates by .031 log points, the 1983-88 average difference between estimates including and excluding allocated earners. For 1994-95 we adjust the "All Worker" gap upward by .046 log points, the average difference for 1996-98. For the years 1973-78 we have the "reverse" problem, needing an approximation of what estimated union gaps would be had the sample included those with imputed earnings. To complete the "not corrected" series, we subtract .035 log points from the "corrected" 1973-78 estimates, .035 being the average difference in the two series during 1979-81.

It was shown above that the downward bias in wage gap estimates can be approximated by **Ω***Γ*. How good is **Ω***Γ* at approximating match bias? This cannot be answered exactly since *Γ*, the unbiased union gap, is not known precisely if there is nonignorable response bias. But we can get an idea given that we know the value of **Ω**, the proportion of allocated earners, in many of the years, and even if there exists nonignorable response bias, it need not affect *relative* union-nonunion wages across the two samples.[17] Returning to 1983, **Ω** = .139 (Table 2), so upward adjustment of the estimated gap including allocated earners of .202 by 1/(1-**Ω**), or 1.161, leads to a predicted "true" gap of .235, nearly identical to the .236 value obtained by exclusion of those with imputed earnings.

In most years, the predicted gap based on the simple match bias formula is slightly above the gap obtained from the sample excluding allocated earners. For example, in 2000, 30% of our CPS estimation sample has imputed earnings (**Ω** = .304). The 2000 private sector gap including all earners is .137. Upward adjustment of this figure by 1/(1-**Ω**) or 1.437, leads to a predicted gap of .197, slightly higher than the .188 obtained when those with imputed earnings are excluded. In short, in most years our approximate bias formula **Ω***Γ* provides a close upper-bound estimate of the bias in wage gap estimates when the attribute under investigation (e.g., union status) is *not* a match criterion used by the Census to impute earnings. Knowledge about the average proportion of allocated earners generally provides sufficient information to make a reasonable adjustment to union wage gap estimates, and, most likely,

[17] Nonignorable response bias will affect both the samples with and without imputed earners. 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. We have no priors as to the direction of bias

other wage gap estimates for attributes not used by the Census in their hot deck procedure.

An exception to this generalization is 1999, when the spread between our estimates with and without union status as a match criterion clearly exceeds $\Omega\Gamma$, the maximum bias predicted for a non-match attribute. Adjusting upward the all-worker estimate of .137 by 1.395 based on the .283 value of **Ω**, we predict an upward bound estimate of .191, as compared to the actual value of .200. Note that a higher union than nonunion nonresponse rate will tend to cause a bias greater than $\Omega\Gamma$, while correlation between union status and the match criteria mitigates bias. Nonresponse rates for union and nonunion workers are highly similar in most years, but the year with the largest gap is 1999, with a union nonresponse rate of 29.5% and nonunion rate of 28.1%. In 1996, there was also a noticeable gap, the union rate being 23.8% and nonunion rate 22.6%. In both 1996 and 1999, the full sample union gaps drop sharply from the previous year and then misleadingly show little change between 1996-97 and 1999-2000. In contrast, the sample that corrects for match bias shows a gradual and steady decline throughout the late 1990s (as seen in Figure 1).

Accounting for workers with imputed earnings 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 allocated earners 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 .209 in 1978 to .150 in 1979. 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 .184, a more modest .025 decline from the estimate of .209 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 pattern shown in the far right column of Table 4 does make economic sense, since 1979 was a period with much unanticipated inflation

in union gap owing to nonresponse bias; we would be surprised if any such bias were large. As stated previously, and contractual union wages may not have adjusted upward so quickly as did nonunion wages.[18]

Figure 1 provides a graphical representation of our two *time-consistent* union wage gap estimates for 1973-2000, one biased downward owing to imputation match bias and the other "corrected" (approximately) for match bias. Time-consistent estimates are crucial for understanding changes over time in the union wage premium. 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. For example, two recent papers (Blanchflower, 1999; Bratsberg and Ragan, 2001) conclude that CPS union premiums have shown little trend since 1983. CPS evidence in Hirsch and Macpherson (2001) indicates a modest decline in union premiums, while data from the BLS Employment Cost Index (ECI) indicates a substantial closing of the private sector union-nonunion wage gap.

Reconciliation of ECI and CPS figures is beyond the scope of this paper. Our corrected time-consistent union gap estimates, however, indicate that union premiums obtained in the late 1970s were maintained or increased during the early 1980s. Such a conclusion would not necessarily follow using uncorrected CPS estimates, since one would obtain wage gaps similar to the "squares" in Figure 1 for 1973-78, but then jump down to the "diamonds" for 1979 forward. Using either the corrected or uncorrected series one would conclude that there has been a modest narrowing of the private sector union wage gap since the mid-1980s. Absent correction for imputation match bias, however, one would overstate the decline since 1994 owing to the increase in allocated earners and find it difficult to distinguish real trends in the union wage gaps from year-to-year variation (e.g., during 1996-2000) due to changes in the number and composition of allocated earners.

VI. 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

the match bias considered in the paper exists even if nonresponse is random.

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

is sizable for the measurement of relative union-nonunion earnings, causing recent union wage gaps to be understated by about 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 Census allocated earners is in fact the result of union status being excluded as a match criterion. Second, union wage gap estimates obtained when Census allocated earners are excluded can be compared with results obtained for the full sample using an imputation method with union status as a match criterion. Third, information is gathered on the mix of donors matched to union and nonunion nonrespondents, providing some direct evidence on the extent to which match bias is mitigated.

A cell hot deck multiple 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 far 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 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 combined sample of private sector workers in the 1996-2000 CPS-ORG files (n=640,231; see the table note for a list of control variables). Based on the Census imputation procedure, we obtain a "full sample" union log wage gap of .151. When we exclude the 25.7% of the sample with allocated earnings, the estimated wage gap rises .05 log points to .201. When we use our own multiple hot deck procedure, without union status as a match criterion, we obtain a log gap of .147, 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 union log wage gap of .216 results, a bit higher than the .201 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 highly similar to the mean union gap across the 50 rounds, reflecting little variation across the 50 sets of earnings data.[19]

Using our own imputation scheme also provides us with information on the union status of earnings 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 .088% union, while union nonrespondents are matched to 17.1% union members (the sample mean among respondents or potential donors is 9.4%). The predicted bias based on the 25.7% nonresponse rate in the sample, an equal donor mix, and a .200 unbiased gap estimate is $\Omega\Gamma$ =.0514. With the unequal union donor mix of .171 and .088 to union and nonunion nonrespondents, respectively, and the differential rate of nonresponse for union and nonunion workers (26.2% and 25.6%, respectively), the predicted bias (see Table 1, line 10) is .0479 (i.e., $[\rho_n\Omega_{\mathrm{u}} + \rho_u\Omega_{\mathrm{n}}]\Gamma$ =[.829*.262 + .088*.256]*.20 = .0479. In short, match bias is mitigated only slightly by the correlation of union status with explicit match criteria.

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 our simple procedure and the Census procedure (the left column of Table 5) and when excluding imputed earners and using union status as a match criterion (the right column) indicate that imputed earners are not so unique a sample that their exclusion substantially biases estimates, given the presence of standard control variables. The higher union wage

[19] 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.

gap estimate using the full sample with union status as a match criterion than with the sample omitting allocated earners (.216 versus .201), however, suggests that to the extent that nonrespondents are not fully representative, they have characteristics associated with higher than average union wage premiums.

In short, we have found that omitting allocated earners from the estimation sample provides a reasonable approximation of a wage gap estimate purged of match bias. But a check on such results is warranted. The strategy employed here has been to retain the full sample, but to use an alternative imputation procedure in which the gap attribute under study is an explicit match criterion.

VII. 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 in turn 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.[20] 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 1996-2000, 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 .132 with allocated earners excluded, compared to .104 for the full sample, an understatement in dispersion of 21.2%. A similar result is found using the mean absolute deviation, .097 with allocated earners excluded versus .077 with them included, a 20.6% 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 private sector.[21] In each case, estimated gaps including allocated earners are biased toward zero. Bias in the postal-private gap estimate is particularly large, with a log differential of .241 obtained among those reporting earnings, versus a biased measure of .184 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.[22]

VIII. Conclusion and Implications

Researchers have not given sufficient attention to what can be substantial bias in wage differential 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

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

[21] In the regressions estimating public sector differentials, we omit controls for union status and industry, effecting a comparison of public sector workers with union and nonunion private sector workers across all industries. Such an approach corresponds to comparability laws that mandate compensation for public sector workers equivalent to that for similar levels of work in the private sector. Discussion of the issues involved is contained in Linneman and Wachter (1990).

[22] Approximately half the CPS sample has answers provided by a "proxy" respondent, typically another family member. In work not shown, match bias is found to be similar in both the proxy and non-proxy samples. A paper by Bishop, Formby, and Thistle (1999) uses proxy status as an instrument for earnings imputation in order to estimate the response or reporting bias in earnings using the March CPS. They do not consider or mention the match bias that is the focus of this paper.

would not be remedied by the use of alternative hot deck methods, propensity score matching, or multiple imputation procedures. Rather, for an attribute not used as an imputation match criterion, wage gap estimates with respect to that attribute are systematically understated. A simple approximation of the match bias is $\Omega\Gamma$, where Ω is the proportion of allocated earners and Γ is the unbiased measure of the wage gap. Relaxing the assumptions that $\Omega_u = \Omega_n$ and that the mix of donors is equivalent for union and nonunion respondents allows one to formulate the more general measure of match bias, $(1-\rho_u)\Omega_u + \rho_n\Omega_n$. Match bias is mitigated by correlation between the attribute under study and the explicit match criteria. In our examples, however, there is little mitigation of bias.

We have shown that bias from imputed earnings in the estimation of union wage gaps and sectoral wage differentials (industry and public-private employment) 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 (a two-way race classification is used for matching but not Hispanic status), immigrant status, marital status, presence and number of children, 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. 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 use industry, union status, or other attributes as explicit match criteria in their hot deck procedure. The Census match variables are firmly based on a supply-side explanation for earnings determination, with demographic, hours, and human capital (schooling, age, occupation) variables selected. There is a nontrivial cost to adding variables (cells) to the match procedure. As match attributes are added, cell sizes become smaller and the probability of finding a recent donor, let alone a donor living nearby, declines. This is particularly true of union status since a small proportion of private employees are union members. Alternative matching procedures (e.g., propensity score methods) could incorporate a larger number of attributes, but match bias would remain since union (nonunion) workers will continue to be matched to nonunion (union) donors.

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 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. Moreover, 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 and causes coefficient bias on those attributes not used as match criteria. The extent of match bias is proportional to the share of the sample with imputed earnings. Not surprising, given that over a quarter of earnings records in the current CPS contain imputed values, substantial match bias is found for union and sectoral wage gaps. 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

Angrist, Joshua D. and Alan B. Krueger. “Empirical Strategies in Labor Economics.” In *Handbook of Labor Economics, Vol. 3A*, ed., Orley Ashenfelter and David Card. Amsterdam: Elsevier, 1999, 1277-366.

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*. ed. Sami Daniel, Philip Arestis and John Grahl. Northampton, Mass.: Edward Elgar, 1999, 3-32.

Bollinger, Christopher. “Measurement Error in the Current Population Survey.” *Journal of Labor Economics* 16 (3), 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 (1), January 1991, 1-24.

Bratsberg, Bernt and James F. Ragan Jr. “Changes in the Union Wage Premium by Industry – Data and Analysis.” Unpublished manuscript, Kansas State University, 2001.

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*, ed. Kevin Lang and Jonathan S. Leonard. New York: Basil Blackwell, 1987, 48-89.

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

Freeman, Richard B. “In Search of Union Wage Concessions in Standard Data Sets.” *Industrial Relations* 25 (2), 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 (3), July 1992, 515-35.

Gregory, Robert G. and Jeffrey Borland. “Recent Developments in Public Sector Markets.” In *Handbook of Labor Economics, Vol. 3C*, ed., Orley Ashenfelter and David Card. Amsterdam: Elsevier, 1999, 3573-630.

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 (223), 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*, ed., Orley Ashenfelter and David Card. Amsterdam: Elsevier, 1999, 1865-2097.

Helwege, Jean. "Sectoral Shifts and Interindustry Wage Differentials." *Journal of Labor Economics* 10 (1), 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 (1), January 2000, 125-55.

Hirsch, Barry T. and David A. Macpherson. *Union Membership and Earnings Data Book: Compilations from the Current Population Survey*. Washington D.C.: The Bureau of National Affairs, 2001.

Hirsch, Barry T. and Edward J. Schumacher. "Unions, Wages, and Skills." *Journal of Human Resources* 33 (1), Winter 1998, 201-19.

Kim, Dae Il. "Reinterpreting Industry Premiums: Match-Specific Productivity." *Journal of Labor Economics* 16 (3), July 1998, 479-504.

Krueger, Alan B. and Lawrence H. Summers. "Efficiency Wages and the Inter-industry Wage Structure." *Econometrica* 56 (2), 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 (3), June 1986, 489-506.

Linneman, Peter D. and Michael L. Wachter. "The Economics of Federal Compensation." *Industrial Relations* 29 (1), Winter 1990, 58-76.

Mellow, Wesley and Hal Sider. "Accuracy of Response in Labor Market Surveys: Evidence and Implications." *Journal of Labor Economics* 1 (4), October 1983, 331-44.

Moon, Marilyn and F. Thomas Juster. "Economic Status Measures in the Health and Retirement Study." *Journal of Human Resources* 30 (Supplement), 1995, S138-57.

Polivka, Anne E. and Jennifer M. Rothgeb, "Overhauling the Current Population Survey: Redesigning the Questionnaire," *Monthly Labor Review* 116 (9), 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*. ed. Jack E. Triplett. Chicago: University of Chicago Press and National Bureau of Economic Research, 1983, 333-43.

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

Smith, James P. "Racial and Ethnic Differences in Wealth in the Health and Retirement Study." *Journal of Human Resources* 30 (Supplement), 1995, S158-83.

U.S. Department of Labor, Bureau of Labor Statistics, *Current Population Survey: Design and Methodology, Technical Paper 63*, March 2000, available at 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	Bias
1.	0.10	0.10	0.25	0.25	0.20	0.0500
2.	0.10	0.10	0.26	0.24	0.20	0.0516
3.	0.10	0.10	0.30	0.20	0.20	0.0580
4.	0.10	0.10	0.20	0.30	0.20	0.0420
5.	0.17	0.09	0.25	0.25	0.20	0.0460
6.	0.17	0.09	0.26	0.24	0.20	0.0475
7.	0.17	0.09	0.30	0.20	0.20	0.0534
8.	0.50	0.03	0.26	0.24	0.20	0.0274
9.	1.00	0.00	0.26	0.24	0.20	0.0000
10.	.171	.088	.262	.256	0.20	0.0479

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. Line 10 utilizes the values of ρ and Ω obtained in subsequent analysis.

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

Year	No Sample Restrictions	Private Sector Estimation Sample	Year	No Sample Restrictions	Private Sector Estimation Sample
1973	.184	.192	1988	.145	.148
1974	.207	.216	1989	.037	.038
1975	.181	.192	1990	.039	.039
1976	.200	.207	1991	.044	.044
1977	.180	.192	1992	.042	.042
1978	.212	.228	1993	.046	.046
1979	.185	.188	1994	.000	.000
1980	.159	.164	1995 (Jan-Aug)	.000	.000
1981	.161	.161	1995 (Sep-Dec)	.233	.233
1982	–	–	1996	.222	.227
1983	.138	.139	1997	.222	.227
1984	.147	.149	1998	.236	.241
1985	.143	.144	1999	.276	.283
1986	.107	.109	2000	.298	.304
1987	.136	.138			

Note. – Data for 1973-81 are from the May CPS Earnings Supplements. Data for 1983-2000 are from the monthly CPS-ORG earnings files. The figures without sample restrictions include the sample of all employed wage and salary workers ages 16 and over. The figures for the estimation sample, corresponding to analysis presented in Table 4 and Figure 1, are for a sample of private sector nonagricultural wage and salary workers, with no missing observations on control variables included in the estimated wage equation. Figures shown for May 1973-78 represent the proportion of earnings records with missing values. The designated allocation flags for 1989-93 identify only some (about a quarter) of records with imputed earnings. Valid allocation flags are not included in the CPS-ORG for January 1994 through August 1995. Sample sizes prior to restrictions are an average 40,681 for the years 1973-78, 25,568 in 1979, 16,081 in 1980, 14,709 in 1981, and an average 167,992 for 1983-2000. Table 4 provides average sample sizes for the estimation samples.

Table 3:
Characteristics of CPS Respondents and Allocated Earners

Variable	Allocated Earners	Respondents
Wage (2000$)	14.85	14.48
Age	39.50	37.53
Education	13.19	13.14
Male	0.536	0.515
Black	0.118	0.080
Asian	0.045	0.037
Hispanic	0.084	0.096
Married w/ spouse	0.534	0.560
Separated, divorced, widowed	0.164	0.157
MSA, medium	0.399	0.420
MSA/CMSA, large	0.399	0.321
Foreign born	0.134	0.123
Part time	0.160	0.200
Proxy respondent	0.616	0.494
Union member	0.097	0.094
N	164,427	475,804
	Union	Nonunion
Proportion of allocated earners	.262	.256
N	60,871	579,360

Note. – Data are from the 1996-2000 monthly CPS-ORG earnings files. The sample includes 640,231 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, 1973-2000

	Unadjusted Wage Gaps, without Controls		Regression Wage Gaps, with Controls	
Year	Not Corrected for Match Bias	Corrected for Match Bias	Not Corrected for Match Bias	Corrected for Match Bias
1973	0.279	0.326	0.127	0.162
1974	0.280	0.327	0.129	0.164
1975	0.283	0.330	0.145	0.180
1976	0.289	0.336	0.155	0.190
1977	0.331	0.378	0.179	0.214
1978	0.330	0.377	0.174	0.209
1979	0.285	0.334	0.150	0.184
1980	0.321	0.363	0.162	0.196
1981	0.329	0.380	0.152	0.189
1983	0.345	0.386	0.202	0.236
1984	0.344	0.386	0.212	0.245
1985	0.339	0.378	0.209	0.243
1986	0.334	0.364	0.206	0.230
1987	0.326	0.366	0.200	0.232
1988	0.323	0.361	0.192	0.224
1989	0.311	0.349	0.193	0.224
1990	0.282	0.320	0.178	0.209
1991	0.268	0.307	0.174	0.205
1992	0.268	0.307	0.180	0.211
1993	0.280	0.318	0.187	0.218
1994	0.265	0.317	0.181	0.227
1995	0.253	0.305	0.177	0.223
1996	0.254	0.306	0.172	0.220
1997	0.253	0.304	0.172	0.215
1998	0.239	0.291	0.162	0.208
1999	0.210	0.276	0.137	0.200
2000	0.200	0.255	0.136	0.188

Note. – Data for 1973-81 are from the May CPS Earnings Supplements and for 1983-2000 from the monthly CPS-ORG earnings files. The sample includes employed private sector nonagricultural wage and salary workers ages 16 and over with 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 1979-2000. Because the sample does not include allocated earnings in 1973-78, the "Not Corrected" series are adjusted downward by the average bias found during 1979-81, .047 for the raw gap and .035 for the regression gap. Columns labeled "Corrected for Match Bias" attempt to include only workers reporting earnings. All allocated earners are identified and excluded for the years 1973-88 and 1996-2000. For 1989-95 allocation flags are either unreliable (in 1989-93) or not available (1994 through August 1995). For 1989-93, the gaps are adjusted upward by the average imputation bias during 1983-88 (.038 for the raw gap and .031 for the regression gap). For 1994-95 the gap is adjusted upward by the bias during 1996-98 (.052 for the raw gap and .046 for the regression gap). Sample sizes are an average 32,765 for 1973-78 for the samples without allocated earners, and for the samples including allocated earners 20,619 in 1979, 12,937 in 1980, 11,940 in 1981, and an average 137,222 for 1983-2000. 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, .010 in 1981, and an average .004 for 1983-2000.

Table 5
Union Wage Gap Estimates Using Alternative Hot Deck Imputation Methods With and Without Union Status as a Match Criterion, 1996-2000

	All workers, Census hot deck method, excludes union as match criterion	Excludes workers with Census imputation
CPS Imputation:		
Coefficient	.1515	.2010
Standard error	(.0020)	(.0023)
R^2	0.442	0.489
N	640,231	475,804
	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	.1469	.2157
Round 1 standard error	(.0020)	(.0020)
R^2	0.439	0.447
Mean gap across 50 sets	.1467	.2160
Standard deviation of coefficients	(.0009)	(.0010)
N	640,231	640,229

Note. – Data are from the 1996-2000 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 two 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.

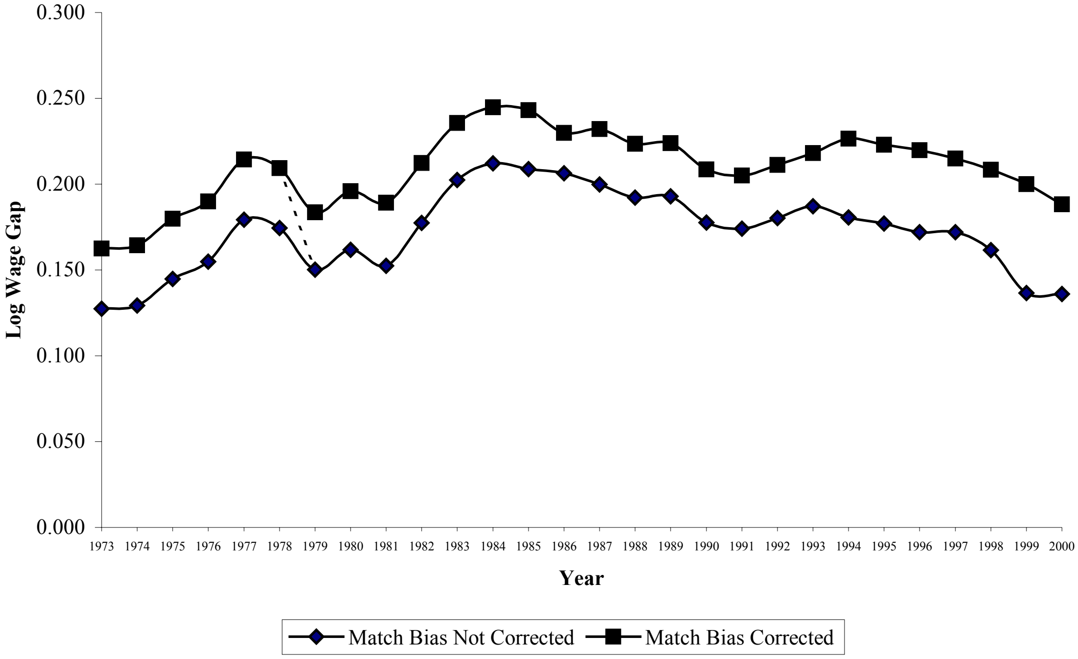
.

Table 6
The Effect of Earnings Imputation on Industry, Public Sector, and Other Selected Wage Differentials, 1996-2000

	Not Corrected for Match Bias	Corrected for Match Bias
Industry:		
Standard Deviation	.104	.132
Mean Absolute Deviation	.077	.097
N	640,231	475,804
Federal (non-postal)	.093	.121
	(.004)	(.004)
Postal	.184	.241
	(.006)	(.007)
State Government	-.034	-.041
	(.003)	(.003)
Local Government	-.028	-.035
	(.002)	(.002)
N	769,938	576,419
Hispanic	-.085	-.102
	(.002)	(.003)
Married, Spouse Present	.090	.106
	(.002)	(.002)
Separated, Divorced, or Widowed	.042	.050
	(.002)	(.003)
Veteran	-.022	-.028
	(.002)	(.002)
Foreign Born	-.066	-.083
	(.002)	(.002)
MSA 100,000-2.5 Million	.089	.108
	(.002)	(.002)
MSA > 2.5 Million	.183	.232
	(.002)	(.002)
N	640,231	475,804

Note. – Data are from the 1996-2000 monthly CPS-ORG earnings files. 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 a regression identical to that in the top panel, except that 8 rather than 27 industry dummies are included. Standard errors are in parentheses.

Figure 1: Union-Nonunion Private Sector Wage Gaps: With and Without Imputation Match Bias



Note. – For details on estimation, see Table 4 and discussion in text. Each of the series is time consistent, the "squared" line corrects for imputation match bias and the line with "diamonds" includes the bias. Researchers who use all valid earnings records in publicly available CPS files would obtain union 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.