Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching

Christopher R. Bollinger, *University of Kentucky*

Barry T. Hirsch, *Trinity University*

This article examines match bias arising from earnings imputation. Wage equation parameters are estimated from mixed samples of workers reporting and not reporting earnings, the latter assigned earnings of donors. Regressions including attributes not used as imputation match criteria (e.g., union) are severely biased. Match bias also arises with attributes used as match criteria but matched imperfectly. Imperfect matching on schooling (age) flattens earnings profiles within education (age) groups and creates jumps across groups. Assuming conditional missing at random, a general analytic expression correcting match bias is derived and compared to alternatives. Reweighting a respondent-only sample proves an attractive approach.

I. Introduction

In household surveys conducted by the U.S. Census Bureau, nonresponse rates for most questions are low. The exception is the high rate of nonresponse for questions on earnings and other sources of income. The chief reason for nonresponse is concern about confidentiality, although other reasons, such as insufficient knowledge among surveyed household members, matter as well (Groves and Couper 1998; Groves 2001). The approach

Previous versions of this article were presented at the research conference in honor of Mark Berger (held at the University of Kentucky, October 7–8, 2004); at the meetings of the Midwest Econometrics Group, the Society of Labor Economists,

[*Journal of Labor Economics,* 2006, vol. 24, no. 3]

0734-306X/2006/2403-0004$10.00

most frequently employed by researchers is to use imputed values provided by the Census. The implications of using Census (and other) imputations in estimation, however, are not well understood.[1] Lillard, Smith, and Welch (1986) warned that nonresponse and imputations in the March Current Population Survey (CPS) significantly affected conclusions about income and earnings. Recent work in the statistics literature (e.g., Schafer and Schenker 2000; Wu 2004) has focused upon inference with imputed values. Other work (Horowitz and Manski 1998, 2000) has focused on identification conditions when data are missing, but that does not directly address the issue of using imputations. Hirsch and Schumacher (2004), whose work we extend, show that coefficient bias resulting from imputation of a dependent variable (earnings) can be of first-order importance.

The CPS monthly earnings files have earnings and wages imputed by the Census using a "cell hot deck" procedure, in which the Census "allocates" (assigns) to nonrespondents the reported earnings of a matched donor who has an identical mix of measured attributes. The proportion of imputed earners was approximately 15% from 1979 to 1993, increased as a result of CPS revisions in 1994, and has risen in recent years to almost 30% (Hirsch and Schumacher 2004, table 2). For a variety of reasons, the Census and the Bureau of Labor Statistics (BLS) include earnings of both respondents and nonrespondents in published tabulations of earnings and other outcomes of interest. Researchers typically do the same when estimating earnings equations, under the belief that including individuals with imputed earnings causes little bias in empirical results (Angrist and Krueger 1999, 1352–54). Hirsch and Schumacher (2004) show that, in a standard earnings equation, there exists attenuation or "match bias" toward zero for coefficients on those characteristics that are not imputation match criteria (e.g., union status). The attenuation is severe, roughly equal to the sample proportion with imputed earnings. Match bias operates independently of possible response bias, existing even when nonresponse is random (i.e., missing at random).

Match bias associated with "nonmatch" attributes (i.e., those not included as Census match criteria) is a first-order problem. As shown in this article, serious bias issues also arise with match attributes that are

and the Western Economic Association; and at seminars at Florida State University, Georgia State University, University of Kentucky, and Ohio State University. Helpful comments were received from John Abowd, Dan Black, David Blanchflower, J. S. Butler, Shiferaw Gurma, James Heckman, David Macpherson, Cordelia Reimers, Mary Beth Walker, and Aaron Yelowitz. An unpublished appendix (Bollinger and Hirsch 2006) accompanying this article can be downloaded, or it may be obtained from either of the authors. Contact the corresponding author, Barry Hirsch, at bhirsch@trinity.edu.

[1] For an excellent survey of imputation procedures, see Little and Rubin (2002, 60), who state: "Despite their popularity in practice, the literature on the theoretical properties of the various [hot deck] methods is very sparse."

imperfectly matched. The Census uses broad categories to match donors' earnings with nonrespondents. For example, rather than matching on the exact age, individuals are grouped into six age categories. Similarly, Census uses three education categories—less than high school, high school through some college, and Bachelor of Arts (BA) or above. When researchers include regressors in a wage equation containing greater detail than the match categories, say, detailed age or specific educational attainment levels, match bias can lead to highly misleading results.

This article presents a general framework for examining match bias due to earnings imputation, deriving an analytic general bias measure under the assumption of conditional mean missing at random (CMMAR). Using this framework, we first formalize expressions for bias in the case of dummy variables of nonmatch attributes (e.g., union status), the important case studied by Hirsch and Schumacher (2004). We then examine various cases of incomplete match. Even under the assumption of CMMAR, we show that biased wage regression estimates occur when including match attributes (e.g., schooling) at a level more detailed than that used in the Census imputation match. We derive a set of corrections for incomplete match bias, demonstrate their use in several examples, and compare alternative approaches researchers might take to account for match bias.[2]

Coefficient bias due to imperfect imputation is widespread and often severe. Authors using the CPS need to assess the importance of match bias in their specific application. Use of a full-sample general bias measure developed in this article provides one approach. A simple alternative is to exclude imputed earners, basing estimates on a respondent-only sample. Given standard assumptions, these approaches provide estimates with equivalent expected values. In practice, reweighting the respondent sample by the inverse probability of being in that sample is found to be an attractive approach when response is not random and coefficients vary with sample composition.

II. Census Earnings Imputation Methods in the CPS Monthly Earnings Files

Statistical agencies often impute or assign values to variables when an individual (or other unit of observation) does not provide a response or

[2] We do not directly address match bias in longitudinal analysis. Hirsch and Schumacher (2004) provide an informal discussion. Hirsch (2005) describes a form of bias that arises in longitudinal estimates even when there is "perfect" matching on an attribute, in his case part-time status. Although there is no mismatch between a nonrespondent and a donor's part-time status in any given year, there is a mismatch in that part-time/full-time switchers—from whom change coefficients estimates are identified—are highly likely to be assigned in one year the earnings of a part-time stayer and in the other year the earnings of a full-time stayer. Fixed effects are not zeroed out, and wage change estimates are biased toward the wage level results.

when a reported value cannot be shown because of confidentiality concerns. Imputation is common for earnings and other forms of income where nonresponse rates are high. The appeal of imputation is that it allows data users to retain the full sample of individuals, which, with application of appropriate weights, can provide population counts and other population statistics. Often imputation of one or a few variables makes it practical to retain an observation and use reported (nonimputed) information on other variables. Government agencies typically publish tables with descriptive data at relatively aggregate levels classified by broad categories (e.g., earnings by sex, age, and race). As long as the published classification categories are match criteria used in the imputation and are not presented at a level narrower than in the imputation, inclusion of imputed earners does no harm. There is bias where presentation is for a nonmatch criterion, say, earnings by union status and/or industry, or for classifications at finer levels, such as earnings by detailed rather than broad occupation.[3]

Analysis in this article uses the CPS Outgoing Rotation Group (ORG) monthly earnings files, prepared by the Census for use by the BLS, which then makes these files publicly available. An earnings supplement is administered to the quarter sample of employed wage and salary workers in their outgoing fourth and eighth months included in the survey. The sample design of the CPS is that individuals are included in the survey for 8 months—4 consecutive months in the survey, followed by 8 months out, followed by 4 months in (the same months as in the previous year). The CPS-ORG earnings files begin in January 1979. They are typically used as annual files, including the 12 quarter samples during a calendar year.[4]

During the period 1979-93, approximately 15% of employed wage and salary workers had imputed values included for usual weekly earnings.[5] The CPS earnings questions were revised in 1994. The increased complexity and sequencing of earnings questions led to a substantial increase in imputation rates. Publicly available earnings files for January 1994

[3] The BLS publishes an annual table compiled from the CPS earnings files that compounds these forms of bias, providing median weekly earnings for union and nonunion workers by industry and by occupation (the latter at a level more detailed than the imputation match). See U.S. Department of Labor (various years).

[4] Prior to 1979, the earnings supplement was administered to all rotation groups in May 1973 through May 1978. Nonrespondents are included in the May 1973–78 earnings files, but they do not have their earnings imputed. Approximately 20% of employed wage and salary workers in the May 1973–78 files have no value (or the "missing" value) included in the usual weekly earnings field (Hirsch and Schumacher 2004, table 2).

[5] Earnings allocation flags are not reliable during the period 1989–93. Imputed earners can be identified based on those who do and do not have an entry in the "unedited" usual weekly earnings field (Hirsch and Schumacher 2004).

Table 1
CPS-ORG Cell Hot Deck Match Criteria, 1979 to Present

Match Criterion	Number of Cells	Categories
Gender	2	Male, female
Age	6	14–17, 18–24, 25–34, 35–54, 55–64, 65+
Race	2	Black, nonblack
Education	3	Less than high school
		High school through some college
		BA or above
Occupation (1979–2002)	13	Executive, administrative, and managerial occupations
		Professional, specialty occupations
		Technicians and related support occupations
		Sales occupations
		Administrative support occupations, including clerical
		Private household occupations
		Protective service occupations
		Service occupations, except protective and household
		Precision production, craft and repair occupations
		Machine operators, assemblers, and inspectors
		Transportation and material moving occupations
		Handlers, equipment cleaners, helpers, and laborers
		Farming, forestry, and fishing occupations
Occupation (2003–present)	10	Management, business, and financial occupations
		Professional and related occupations
		Service occupations
		Sales and related occupations
		Office and administrative support occupations
		Farming, fishing, and forestry occupations
		Construction and extraction occupations
		Installation, maintenance, and repair occupations
		Production occupations
		Transportation and material moving occupations
Hours worked:		
Prior to 1994	6	0–20, 21–34, 35–39, 40, 41–49, 50+
Added 1994 and after	8	Hours vary, usually full time; hours vary, usually part time
Overtime, tips, or commissions	2	Usually receive; not usually receive
Total imputation cells:		
1979–93	11,232	
1994–2002	14,976	
2003–present	11,520	

Source.—Hirsch and Schumacher (2004) and information provided by U.S. Census Bureau and Bureau of Labor Statistics economists.

Note.—Total imputation cells is the product of the cell numbers shown. In 1994, the designation for variable hours worked was introduced. In 2003, occupational categories were reduced from 13 to 10. Publicly available earnings files for January 1994 through August 1995 do not identify those with imputed earnings.

through August 1995 do not identify those with imputed earnings. Beginning in September 1995, valid earnings allocation flags are included. Imputation rates rose from about 22% in 1996 to about 30% in the period from 2000 to 2004.

Earnings in the CPS-ORG are imputed using a "cell hot deck" method. There has been minor variation in the hot deck match criteria over time. For the ORG files during the 1979–93 period, the Census created a hot deck, or cells containing 11,232 possible combinations based on the following seven categories: gender (2 cells), age (6), race (2), education (3), occupation (13), hours worked (6), and receipt of tips, commissions, or overtime (2). These categories are shown in table 1. The Census keeps all cells "stocked" with a single donor, ensuring that an exact match is always found. The donor in each cell is the most recent earnings respondent surveyed by the Census with that exact combination of characteristics. As each surveyed worker reports an earnings value, the Census goes to the appropriate cell, removes the previous donor value, and "refreshes" that cell with a new earnings value from the respondent.[6]

As shown in table 1, the selection categories changed slightly in 1994 and 2003. Beginning in 1994, two additional hours cells were added for workers reporting variable hours, one for those who usually work full time and one for those who usually work part time, resulting in eight "hours worked" cells and 14,976 possible combinations. Beginning in January 2003, the CPS adopted the 2000 Census occupation codes (COC), which involved a substantial revision from the 1980 and 1990 COC. Detailed occupation codes are grouped into 10 major categories, in contrast to 13 prior to 2003, resulting in 11,520 match cells.

At the start of each month's survey, cells are stocked with ending donors from the prior month. The Census retains donors until replaced, reaching back for donors as far as necessary, first within a given survey month and then to previous months and years. If needed, a donor value is used more than once. A donor's nominal earnings is assigned to the nonrespondent, with no adjustment for wage growth since the cell was refreshed. The Census does not retain information on cell refresh rates or the average "freshness" of donors. A trade-off exists. Less detailed match characteristics would produce more frequent refreshing of cells but would result in lower quality matches.[7]

[6] A brief discussion of Census/CPS hot deck methods is contained in U.S. Department of Labor 2002, 9.3). The more detailed information appearing here and in Hirsch and Schumacher (2004) was provided by economists at the BLS and the Census Bureau. Unlike the ORGs, the March CPS annual demographic files (ADF) use a "sequential" rather than "cell" hot deck imputation procedure to impute earnings (and income). Nonrespondents are matched to donors from within the same March survey in sequential steps, each step involving a less detailed match requirement. For example, suppose that there were just four matching variables—sex, age, education, and occupation. The matching program would first attempt to find a match on the exact combination of variables using a relatively detailed breakdown. Absent a successful match at that level, matching proceeds to a next step with a less detailed breakdown, e.g., broader occupation and age categories. Earnings imputation rates in the ADF are lower than in the ORGs. As emphasized by Lillard et al. (1986), the probability of a close match declines the less common an individual's characteristics. Although the imputation procedure used in the ADF produces a regression bias similar to that identified for the ORGs, our analysis applies most directly to the ORGs.

[7] Location is not an explicit match criterion. Files are sorted by location, and nonrespondents are matched to the most recent matching donor. Thus, a donor is (roughly) the geographically closest person moving backward in the file. Nonrespondents with an unusually common mix of characteristics may be matched to

III. Imputation Match Bias

A. General Approach

In this section, we derive a general analytic approach to evaluate bias from the inclusion of imputed values in the dependent variable (much of the analysis is in the unpublished appendix [Bollinger and Hirsch 2006]). Following the general case, we examine specific cases of interest. We derive an analytic expression for bias in the case considered by Hirsch and Schumacher (2004), where an explanatory variable that is not an imputation match criterion is entered into a regression. We next consider two types of imperfect match. In the first case, a categorical variable such as educational degree or occupation is collapsed into broader categories for the purpose of imputation. In the second case, an ordinal variable that enters the regression, such as age, is collapsed into a set of categorical variables for the purpose of imputation. Finally, we consider a mixed case where a variable collapsed into broader categories for imputation enters the equation as both a linear term and a categorical term (e.g., years of education coupled with degree dummies).

Throughout this section, the variable y_i is the dependent variable in a linear regression, in this case, the natural log of earnings. The variables $\underline{z}_i$ are the regressors of interest, for example, age and education. The variables $\underline{x}_i$ represent the categories upon which matches are made. These variables are binary indicator (dummy) variables in practice, but our analysis does not rely upon this result. The following assumptions are made:

Assumption 1. Only variable y_i is missing, for some but not all observations.

Assumption 2. $E_O[y_i|\underline{z}_i,\underline{x}_i] = E_M[y_i|\underline{z}_i, \underline{x}_i] = E[y_i|\underline{z}_i, \underline{x}_i]$.

Assumption 3. $\underline{x}_i = h(\underline{z}_i)$, where $h()$ is a known deterministic function.

Assumption 4. $E[y_i|\underline{z}_i,\underline{x}_i] = E[y_i|\underline{z}_i] = \alpha + \underline{z}_i'\underline{\beta}$.

Assumption 5. Imputed values of y_i are randomly drawn from the distribution $f_O(y_i|\underline{x}_i)$.

Assumption 1 is self-explanatory. We examine the effect of imputation in the dependent variable only. If all observations had missing values, there would be no donors from which to draw. The imputation effects are similar to measurement error. There is a large (and not unrelated) literature on right-hand-side measurement error.

Assumption 2 is crucial. In assumption 2 and elsewhere, the notation $E_O[y_i|\underline{z}_i,\underline{x}_i]$ reads as the population expectation of y_i when y_i is observed,

someone in a similar neighborhood. More likely, donors are found in different neighborhoods, cities, states, regions, or months. As seen subsequently, we estimate that 83% of nonrespondents are assigned the earnings of donors from previous survey months. In the March CPS, broad region serves as an explicit match criterion for selecting donors.

while $E_M[y_i|\underline{z}_i, \underline{x}_i]$ is the population expectation of y_i for the missing, those who do not report y_i and have earnings imputed. It states that there is no selection on the y_i variable with respect to unobservables (factors not included in $\underline{z}_i$). Assumption 2 assumes conditional missing at random, albeit in a "weak" form, such that there is no difference in mean earnings between the observed and missing, conditional on $\underline{z}_i$. Assumption 2 allows the distribution of $(\underline{x}_i, \underline{z}_i)$ to differ between those who report earnings and those who do not. We call this a "weak" form of "missing at random" (MAR) because it only requires the mean but not the distribution of earnings within a match cell to be equivalent for those who report and do not report earnings. We refer to this as "conditional mean missing at random," or CMMAR. Although not formally considered here, assumption 2 can be further weakened by allowing an intercept difference. Other research (Molinari 2005) considers cases where variables are not missing at random.[8]

Assumption 3 is innocuous, simply stating that knowing $\underline{z}_i$ gives perfect information about the value of $\underline{x}_i$. That is, if you know the value of a variable at its detailed level, you know its value at an aggregated level. The opposite may not be true. Either $h()$ is many to one, as in the schooling and age cases, so $(\underline{x}_i)$ is a crude measure of $\underline{z}_i$, or there may be variables in $\underline{z}_i$ that are not measured in $(\underline{x}_i)$, for example, nonmatch attributes union status, foreign born, and industry. An important implication for this is that $E[\underline{x}_i|z_i] = \underline{x}_i$, while $E[\underline{z}_i|\underline{x}_i]$ is not specified generally.

Assumption 4 implies that the relationship between y_i and $\underline{z}_i$ is linear in the parameters and that $\underline{x}_i$ do not contain information about y_i beyond what is contained in the more detailed variables $\underline{z}_i$. When $\underline{z}_i$ is categorical to begin with, this is always true, while when $\underline{z}_i$ is an ordinal variable, it implies that the specification is linear and there are no further nonlinearities that are better captured by the collapsed categories. Note that nonlinearities are allowed; the vector $\underline{z}_i$ must simply contain appropriate variables such as quadratic terms. Essentially, the assumption implies that the researcher has the correct specification for the conditional expectation function $E[y_i|\underline{z}_i]$.

Finally, assumption 5 implies that, conditional upon $\underline{x}_i$, the distribution of the imputed y_i is independent of the distribution of $\underline{z}_i$. That is, the imputed data conditioned on $\underline{x}_i$ are independent of the variables not included as imputation match criteria.

We consider the population least squares projection of y_i on $\underline{z}_i$ when imputed values are used for those who do not report y_i. Under general

[8] Although CMMAR is assumed above for the general case and for all empirical work, we subsequently impose MAR in some of our illustrative theory sections in order to simplify results.

assumptions, OLS is consistent for the least squares projection. The unpublished appendix (Bollinger and Hirsch 2006) formally derives the following important result for the population least squares slope coefficients $\underline{b}$ on variables $\underline{z}_i$:

$$\underline{b} = \underline{\beta} - p\left(E\left[\underline{z}_i\underline{z}_i'\right] - E\left[\underline{z}_i\right]E\left[\underline{z}_i\right]'\right)^{-1} \times \left(E_M\left[\underline{z}_i(\underline{z}_i - E_O\left[\underline{z}_i|\underline{x}_i\right])'\right]\right.$$
$$\left. - E[\underline{z}_i]E_M\left[\underline{z}_i - E_O[\underline{z}_i \mid x_i]'\right]\right)\underline{\beta}.$$

The parameter p is the probability of not observing y_i (estimated by the proportion of missing values in the sample). Terms like $E_O\,[\underline{z}_i|\underline{x}_i]$ are the expectation of z_i given x_i for the population who report y_i, while E_M is for the population who do not report y_i. Terms with no subscript are for the full population, including both respondents and nonrespondents. The terms to the right of the initial β produce the match bias resulting from imputation.

The term $(E_M[\underline{z}_i(\underline{z}_i - E_O\,[\underline{z}_i|\underline{x}_i])'] - E[\underline{z}_i]E_M[\underline{z}_i - E_O[\underline{z}_i \mid x_i]'])$ is the covariance between the regressors $\underline{z}_i$ and the prediction error from the relationship between those regressors and the match variables. Hence, the entire term can be thought of in the following way. First, regress z_i on the match variables and take the residuals $(z_i - E_O\,[\underline{z}_i|\underline{x}_i])$. Then regress those residuals back on z_i. This measures the variation in z_i that is not accounted for by the match variables. In essence, this is measuring the omitted information from the imputation procedure, and it behaves like an omitted variable bias term. This can also be viewed as measurement error. The donor's earnings were generated from a particular value of z, which does not necessarily match the value of z_i of the recipient. The measurement error is $(z_i - E_O\,[\underline{z}_i|\underline{x}_i])$, which measures the difference between the recipient's z_i (the mismeasured variable) and the average donor's z_i for donors in the cell. The bias term is similar to the usual attenuation term found with measurement error.

Rearranging the equation above, we arrive at the following expression:

$$\underline{\beta} = \left(I - p\left(E\left[\underline{z}_i\underline{z}_i'\right] - E\left[\underline{z}_i\right]E\left[\underline{z}_i\right]'\right)^{-1} \times \left(E_M\left[\underline{z}_i(\underline{z}_i - E_O\left[\underline{z}_i|\underline{x}_i\right])'\right]\right.\right.$$
$$\left.\left. - E[\underline{z}_i]E_M\left[\underline{z}_i - E_O[\underline{z}_i \mid x_i]'\right]\right)\right)^{-1} \underline{b},$$

where I is the $k \times k$ identity matrix. This is a "general correction" for match bias; it produces consistent estimates of $\underline{\beta}$ and is applicable in all cases discussed in this article.

Two simple cases may illuminate the nature of match bias. First, note that, if $\underline{z}_i = \underline{x}_i$, implying that all variables in the model are included as imputation characteristics and at the same level of detail, then $\underline{b} = \underline{\beta}$ and no bias exists. Another interesting special case is where we have strict missing at random and z_i and x_i are scalars. In that case, $E_M[\underline{z}_i - E_O[\underline{z}_i|\underline{x}_i]'] = 0$ and $E_M[\underline{z}_i(\underline{z}_i - E_O[\underline{z}_i|\underline{x}_i])']$ is the variance of z_i not explained by x_i. So, the ratio $E_M[\underline{z}_i(\underline{z}_i - E_O[\underline{z}_i|\underline{x}_i])]/E[\underline{z}_i\underline{z}_i] - E[\underline{z}_i]E[\underline{z}_i] = 1 - V(z_i|x_i)/V(z_i)$, which is similar in concept to $1 - R^2$ but allows for a fully nonlinear model. Indeed, in a case where x_i is binary (as is often the case for imputation characteristics), this is the R^2 from the regression of z_i on x_i. In the extreme case where $R^2 = 1$, all information in z_i can be accounted for by the imputation match criteria x_i, so there is no bias.

B. Empirical Implementation

All terms in the equation for the slope coefficients (seen in the previous section) are estimable in sample. For example, the term $E_O[\underline{z}_i|\underline{x}_i]$ is the mean of the regressor variables, conditional upon the imputation attributes, using only the sample where earnings are reported. The following six steps are used below to estimate the bias and correct the full sample estimates for imputation bias:

Step 1. Use OLS to estimate the slopes on the full sample (including imputations). Retain the inverse of the variance of $\underline{z}_i$.

Step 2. Using the $R_i = O$ (observed) subsample, estimate $E_O[\underline{z}_i|\underline{x}_i]$. As a practical matter, in the CPS, this can be done using OLS on a full set of interaction terms for the imputation categories: age, education, gender, race, and so forth. Alternatively, this can be done by constructing all imputation cells and averaging within cell.

Step 3. Predict $\underline{z}_i$ using the estimated $E_O[\underline{z}_i|\underline{x}_i]$, for all observations in the $R_i = M$ sample (using the appropriate $\underline{x}_i$ for each observation).

Step 4. Construct $\underline{z}_i(\underline{z}_i - E_O[\underline{z}_i|\underline{x}_i])'$ and $(\underline{z}_i - E_O[\underline{z}_i|\underline{x}_i])$ in the $R_i = M$ sample and average over that sample.

Step 5. Parameter p is estimated by the missing rate in the sample.

Step 6. Use estimated terms to construct estimates of α and β.[9]

Up to this point we have said nothing about bias in coefficient standard errors owing to imputation. Statistical significance is often not an issue in wage analyses owing to large samples. Imputation does bias standard

[9] The expression for β is provided in the previous section. The expression for α is

$$\alpha = a - pE[\underline{z}_i]'\left(E[\underline{z}_i\underline{z}_i'] - E[\underline{z}_i]E[\underline{z}_i']'\right)^{-1}$$
$$\times\left(E_M\left[\underline{z}_i\left(\underline{z}_i - E_O[\underline{z}_i|\underline{x}_i]\right)'\right] - E[\underline{z}_i]E_M\left[\underline{z}_i' - E_O[\underline{z}_i|\underline{x}_i]'\right]\right)\beta$$
$$+\ pE_M\left[\underline{z}_i' - E_O[\underline{z}_i|\underline{x}_i]'\right]\beta.$$

errors, however. Typical estimators of standard errors assume that observations are independent. When imputed values are drawn from other observations included in the sample, that assumption is violated. In general this will cause typical estimated standard errors to understate the true sampling variation. Heckman and LaFontaine (2006, in this issue) address the issue of standard errors in regressions using imputed values by employing the bootstrap algorithm of Shao and Sitter (1996). Little and Rubin (2002) summarize classic work addressing this issue.

Since the imputed observations are not independent of the nonimputed observations, the usual standard errors are not appropriate. Indeed, if the regression is y_i on $\underline{x}_i$, if all imputations are drawn from the observed sample, the standard errors reduce to the standard errors from only the observed sample. In the CPS hot deck procedure, many imputations derive from observations from previous months, some of which may not be included in the estimation sample. If the sample is selected on some $\underline{z}_i$ criteria (including time period), some imputations will be drawn from outside the criteria. In cases where the regression includes variables other than $\underline{x}_i$, as in the case studied here, there is some informational gain to including imputations.

Although one approach to estimating standard errors in this case would be to use a bootstrap, we use estimates based upon standard asymptotic results. Heteroskedastic robust standard errors for the OLS estimates are produced with typical software. To arrive at standard errors for the bias-corrected results, we assume nonstochastic regressors. The variance covariance matrix for the bias-corrected slopes is then simply $A \times V(b) \times A^T$, where A is the bias correction matrix (since the estimates are simply Ab). This may tend to slightly understate the variance since it ignores variation in A. As in most empirical studies, we ignore the issue of sampling variation due to the imputations (Little and Rubin 2002).

In the following sections, we focus on specific forms of match bias, each permitting a simplification from the general case. Following theory presented in each section, we provide illustrative empirical evidence and apply the general bias correction developed here.

C. Match Bias with Nonmatch Attributes: Theory

Here we reconsider the results of Hirsch and Schumacher (2004), who examine the case of coefficient bias on a single nonmatch explanatory variable (e.g., union status). Hirsch and Schumacher present a bias expression for both a simple case where no other covariates are present in the regression and a general case where all other covariates are assumed to be exact match criteria.[10] The second case is an approximation based

[10] In the case of no covariates, Hirsch and Schumacher (2004) show that bias (the sum of match error rates for union and nonunion nonrespondents) is equivalent to

upon the results of Card (1996). We show that the approximation in Hirsch and Schumacher is quite close to the exact analytic result in most cases but that it may differ substantially if a match characteristic is highly correlated with the nonmatch variable.

Let

$$\underline{z}_i = \begin{bmatrix} \underline{z}_{1i} \\ z_{2i} \end{bmatrix},$$

where $\underline{z}_{1i} = \underline{x}_i$ and z_{2i} is a binary variable such as union status. All other covariates are included in the match criteria for imputation, but z_{2i} is not. Let $q = E[z_{2i}] = P(z_{2i})$, $q_M = E_M[z_{2i}]$, $q_O = E_O[z_{2i}]$, $q_M(\underline{z}_{1i}) = P_M[z_{2i}|\underline{z}_{1i}]$, $q_O(\underline{z}_{1i}) = P_O[z_{2i}|\underline{z}_{1i}]$, $V_{11} = V(\underline{z}_{1i})$, and $C = \mathrm{Cov}(\underline{z}_{1i}, z_{2i})$, while R^2 is from the linear regression of z_{2i} on $\underline{z}_{1i}$ in the full population. Then the results in the unpublished appendix (Bollinger and Hirsch 2006) demonstrate that the coefficient from the LS projection of y_i on z_{2i} will be

$$\begin{aligned} b_2 = \beta_2\Bigg(1 - p\Bigg(\Bigg(&\frac{q_M - E_M[q_M(\underline{z}_{1i})q_O(\underline{z}_{1i})] - q(q_M - E_M[q_O[\underline{z}_{1i}]])}{(q - q^2)(1 - R^2)}\Bigg) \\ &- \Bigg(\frac{C'V_{11}^{-1}(E_M[\underline{z}_{1i}(q_M(\underline{z}_{1i}) - q_O(\underline{z}_{1i}))] - E[\underline{z}_{1i}](q_M - E_M[q_O(\underline{z}_{1i})]))}{(q - q^2)(1 - R^2)}\Bigg)\Bigg)\Bigg)\Bigg). \end{aligned}$$

The results of Hirsch and Schumacher (2004) provide an expression that is closely related to this but which is based upon the assumption that the probability of misclassification is independent of the match criteria. This is an assumption of the results derived by Card (1996), which, in turn, were applied by Hirsch and Schumacher. If the strong missing at random assumption is applied, the two expressions are both equal to $\beta_1(1 - p)$. Similarly, if $\underline{z}_{1i}$ and z_{2i} are uncorrelated, the results are equivalent. The Hirsch and Schumacher results also do not extend to the case of multiple nonmatch variables. For these reasons, the general match bias correction derived in this article is preferable.

D. Match Bias with Nonmatch Attributes: Evidence

In this section, we compare alternative methods to correct match bias, providing evidence on wage gap estimates with respect to selected attributes that are not match criteria. These gap estimates include union status, marital status, foreign born, Hispanic, and Asian, as well as wage dispersion across region, city size, and employment sectors (industry, public

that from right-hand-side measurement error of a dummy variables, as shown by Aigner (1973) and extended in subsequent literature (e.g., Bollinger 1996; Black, Berger, and Scott 2000).

sector, and nonprofit status).[11] The sample is drawn from the CPS-ORG for the period 1998–2002. These years provide a convenient time period. Beginning in 1998, added information on education, including the GED, was included. Beginning in 2003, new occupation codes (from the 2000 census) led to a change in the imputation match categories (see table 1). Our estimation sample includes all nonstudent wage and salary employees ages 18 and over. Estimates are provided separately by gender, the sample of men being 388,578 and that of women being 369,762. In the male sample, 28.7% have earnings imputed, as compared to 26.8% of the female sample.

Table 2 provides coefficient estimates obtained from a standard log wage equation estimated using alternative approaches. Included in the equations are potential experience in quartic form (defined as the minimum of age minus years schooling minus 6 or years since age 16) and dummy variables for education (23 dummies), marital status (2), race/ethnicity (4), foreign born, union, metropolitan size (6), region (8), occupation (12), employment sector (17), and year (4). The dependent variable is the natural log of average hourly earnings, including tips, commissions, and overtime, calculated as usual weekly earnings divided by usual weekly hours worked. Top-coded earnings are assigned the estimated mean above the cap ($2,885) based on an assumed Pareto distribution above the median (estimates are gender and year specific and roughly 1.5 times the cap, with small increases by year and higher means for men than for women).[12]

Wage gap estimates in table 2 are drawn from regressions based on the full sample with Census imputations (the standard approach among researchers), the imputed ("missing") sample, the respondent ("observed") sample, the observed sample using inverse probability weighting (IPW) to correct for changes in the sample composition, and the full sample using the general bias correction derived in Section III.A. The IPW estimates require a brief explanation. Although we have assumed no specification error, in practice, coefficients may differ across workers with different characteristics. If individuals are missing at random, the composition of the observed and full samples will be the same. If nonresponse is not random, estimates can differ. To account for the change in sample composition correlated with observables, we first run a probit equation with response as the binary dependent variable and all $\underline{z}_i$ as regressors. We then weight the observed sample by the inverse of the probability of response, thus giving

[11] Nonmatch attributes include not only variables measured in the monthly CPS but also attributes measured in CPS supplements such as job tenure, employer size, and computer use.

[12] Mean earnings above the CPS cap by gender and year (since 1973), calculated by Barry Hirsch and David Macpherson, are posted at http://www.unionstats.com.

Table 2
Wage Gap Estimates Corrected and Uncorrected for Match Bias from Nonmatch Criteria

				IP Weighted	Corrected						
	Full Sample	Imputed	Respondents	Respondents	Full Sample	Ratio	Ratio	Ratio	Ratio	Ratio	Ratio
	(1)	(2)	(3)	(4)	(5)	(1)/(3)	(1)/(4)	(1)/(5)	(3)/(4)	(3)/(5)	(4)/(5)
Men:											
Worker attribute coefficient:											
Union member	.142	.024	.191	.193	.199	.75*	.74*	.71*	.99*	.96*	.97*
Married, spouse present	.096	.021	.127	.130	.132	.76*	.74*	.73*	.97*	.96*	.99
Foreign born	−.099	−.024	−.130	−.133	−.139	.76*	.75*	.71*	.98*	.94*	.96*
Hispanic	−.099	−.029	−.123	−.125	−.128	.81*	.79*	.77*	.98*	.96*	.98
Asian	−.024	−.005	−.033	−.038	−.038	.74*	.63*	.63*	.85*	.86	1.00
Mean absolute deviation of coefficients:											
Sector: industry/public/nonprofit (18)	.090	.031	.117	.117	.124	.77	.77	.72	1.01	.95	.94
Metro size (7)	.094	.011	.125	.124	.129	.75	.76	.73	1.01	.97	.97
Region (9)	.023	.013	.034	.033	.031	.67	.68	.72	1.02	1.08	1.06
N	388,578	111,669	276,909	276,909	388,578						
Wald statistic						285.3†	101.7†	991.2†	39.5†	13.5†	7.0†
Women:											
Worker attribute coefficient:											
Union member	.111	.013	.143	.143	.148	.78*	.78*	.75*	1.00	.97*	.97*
Married, spouse present	.028	.016	.033	.032	.037	.86*	.87*	.76*	1.01	.88*	.87*
Foreign born	−.079	−.015	−.105	−.103	−.110	.76*	.77*	.72*	1.01*	.95*	.94*

Hispanic	−.077	−.019	−.096	−.098	−.100	.80*	.78*	.77*	.98*	.96*	.98
Asian	−.016	.002	−.020	−.023	−.020	.78	.68*	.78*	.87*	.99	1.14
Mean absolute deviation of coefficients:											
Sector: industrial/public/nonprofit (18)	.098	.030	.128	.128	.133	.77	.77	.74	1.00	.96	.96
Metro size (7)	.102	.018	.129	.129	.135	.79	.79	.76	1.00	.96	.96
Region (9)	.040	.012	.052	.051	.053	.78	.78	.76	1.01	.97	.96
N	369,762	99,225	270,537	270,537	369,762						
Wald statistic						200.5†	75.7†	681.5†	24.2†	18.1†	9.8†

NOTE.—The sample includes all nonstudent wage and salary workers ages 18 and over, from the January 1998 to December 2002 monthly CPS-ORG earnings files. The proportion of the full CPS sample with imputed earners is .287 among men and .268 among women. Results are shown for the full sample (respondents plus nonrespondents with Census-imputed earnings), imputed (missing) earners only, earnings respondents (observed) only, respondents with inverse probability (IP) weighting, and the full sample with parameter estimates corrected by the general match bias measure. Included in the wage equation are potential experience in quartic form and dummy variables for education (23 dummies), marital status (2), race/ethnicity (4), foreign born, part time, union, metropolitan size (6), region (8), occupation (12), employment sector (17), and year (4). Sector includes 18 groups: 13 private for-profit industry categories, private nonprofit, and the public sector groups postal, federal nonpostal, state, and local. Shown in the top area are log wage gaps with the following reference groups: union versus nonunion workers, married with spouse present versus single, foreign born versus U.S. born, Hispanic versus non-Hispanic white, and Asian versus non-Hispanic white. Shown in the bottom area is the mean absolute deviation of coefficients (unweighted) with the omitted reference group counted as zero. The first three ratio columns show observed attenuation coefficients, the ratio of the uncorrected to alternative corrected estimates. The last three columns show the ratios of corrected estimates.

* Indicates that the null of equal coefficients on the given variable between the designated columns can be rejected at the .05 significance level.

† Indicates that the null of jointly equivalent coefficients between the designated equations can be rejected at the .05 significance level.

enhanced weight to those most likely to be underrepresented in the observed sample (Wooldridge 2002, 587–88). Reweighting does not correct for possible selection on unobservables (factors correlated with earnings but uncorrelated with $\underline{z}_i$).

Severe match bias is readily evident in the estimates shown in table 2. Focusing first on the male sample, the union-nonunion log wage gap is estimated to be .191 among respondents, only .024 among imputed earners, and .142 in the combined sample, a 25% attenuation (1 − [.142/.191]), as seen in the "Ratio (1)/(3)" column. Similar imputation bias is found for other nonmatch criteria. A "married" coefficient measures the wage gap between married males with spouse present and never-married males. The full CPS sample produces an uncorrected marriage premium estimate of .096, while exclusion of imputed earners increases the estimate to .127, implying attenuation of 24%. The wage disadvantage for foreign-born workers is an estimated −.130 in the respondent sample but only −.099 in the full sample. Hispanic workers have an estimated −.123 wage disadvantage using the respondent sample, compared to −.099 in the full sample. Wage gap estimates for Asian workers (compared to non-Hispanic whites) are small but display similarly large attenuation (26%).

There exists a large literature on industry wage dispersion. Whatever one's interpretation of this literature, failure to account for match bias causes industry differentials (using wage-level analysis) to be understated, since employment sector is not a Census match criterion. Table 2 provides wage dispersion estimates among 18 sectors, 13 private for-profit industry groups, 4 public sector groups (federal nonpostal, postal, state, and local), and the private nonprofit sector. The mean absolute log deviation for these 18 sectors is an estimated .117 based on the respondent sample, but it falls to .090 using the full sample. One observes similar attenuation among wage differences for region and city size, standard control variables in most earnings equations.

Turning to the sample of women, we see exactly the same qualitative pattern as that seen for men. Magnitudes of the "worker attribute" wage gaps are somewhat smaller for women than for men. Interestingly, sectoral, region, and city size gaps are slightly larger among women. Attenuation from match bias is generally a little lower among women than men owing to a lower rate of nonresponse.

How do estimates based on the unweighted respondent sample compare to alternatives? Hirsch and Schumacher (2004) suggest that estimation from a respondent-only sample provides a reasonable first-order approximation of a true parameter but may not fully account for match bias. In table 2, we examine two alternatives to use with an unweighted respondent sample. Focusing on the union wage gap, we obtain a corrected full-sample union gap for men of .199, compared to a .191 based on the unweighted respondent sample; corresponding estimates for women are

.148 and .143, respectively. These qualitative differences comport well with results in Hirsch and Schumacher (2004).[13] If differences between the corrected full samples and unweighted respondent samples are a result of composition differences, an attractive alternative may be to use a respondent sample weighted by the inverse of the probability of being in the respondent sample. These IPW results, shown in table 2, produce a union gap estimate of .193 among men, higher than those obtained from the unweighted respondent sample but less than from the corrected full sample. The IPW union gap estimate is .143 among women, the same as the unweighted respondent estimate.

The patterns found for the union gap appear to be typical. As seen in table 2, in all but one case, the corrected full sample estimates exceed (in absolute value) estimates from the respondent sample (the exception is regional wage dispersion among men). The reweighted respondent sample (IPW) results among men tend to lie between the unweighted respondent and corrected full sample estimates. Among women, the IPW results are highly similar to the unweighted respondent results.

In table 2, we present at the bottom of each ratio column significance tests for differences in all coefficients jointly across samples. For males, we obtain Wald statistics (ordered from high to low) of 991.2 for uncorrected full versus corrected full, 285.3 for uncorrected full versus unweighted respondent, 101.7 for uncorrected full versus weighted respondent, 39.5 for unweighted respondent versus weighted respondent, 13.5 for unweighted respondent versus corrected full, and 7.0 for weighted respondent versus corrected full. Although all differences are significant (the critical value is 1.3), that found between the corrected full sample and weighted respondent sample is relatively small. An identical qualitative pattern is found for women. Table 2 also summarizes results from significance tests (at the .05 level) for differences across regressions in coefficients for the five worker attribute nonmatch characteristics included in table 2. In most cases the null of equality is readily rejected. Estimates are most similar among the corrected full and weighted respondent regressions (the far right column). Based on this comparison, we reject the null for "only" two of five coefficients among men and three of five among women.

Which estimation approach is preferable? This question is not easily answered. If we have the correct specification and conditional mean missing at random, as assumed in our bias correction, then the unweighted

[13] When Hirsch and Schumacher (2004) estimate union wage gaps with the full sample, using either their own imputation procedure or correcting bias based on Card's measure for misclassification error, they obtain larger estimates than those obtained from the respondent sample. They suggest that attributes more common among nonrespondents are associated with larger union gaps. They do not explore whether the union result is common among a broader family of wage gap estimates.

respondent sample, the weighted respondent sample (IPW), and the full sample with bias correction should produce consistent estimates. The only "wrong" approach is the standard one, including the full sample with Census imputations and no match bias correction. Differences between the corrected estimates from the full sample and those from the weighted and unweighted respondent samples result either from a violation of CMMAR or differences across groups in the value of the parameter of interest (i.e., specification error). None of these approaches accounts for a violation of CMMAR.[14]

When there exists specification error, some estimation approaches may be preferable to others. Researchers routinely estimate (for good and bad reasons) simple but misspecified models. If one desires a parameter estimate "averaged" across a representative population, then use of either the full sample with bias correction or the reweighted respondent sample may be preferable to the unweighted respondent sample. Although an important contribution of this article is the derivation and use of the full-sample bias correction approach, it faces limitations for more general use. First, it is not trivial to understand and program, making it an unattractive approach in some cases. Second, the bias correction derived here is designed specifically for the cell hot deck imputation used in CPS-ORG, although the setup and its application can be used more broadly. The weighted respondent sample (IPW) approach may be more general, working well regardless of a survey's imputation methods, which may be highly complex or unknown to the researcher.[15] For these reasons, estimates from

[14] It is possible to account for nonignorable selection bias given appropriate instrument(s), but this is not a topic addressed in this article. Hirsch and Schumacher (2004) estimate a selection model in which nonresponse is identified using as an instrument a variable indicating whether CPS survey questions are being answered by the individual or by another household member.

[15] The bias correction derived in this article can be applied to either the CPS-ORG cell hot deck or to the March CPS Annual Demographic File (ADF) sequential hot deck. Its assumptions, however, are more severely violated in the ADF. The bias correction assumes that the draw for the imputation is from the same distribution as the rest of the sample. The imputation draws from the conditional distribution $f(y|X_1, X_2)$, where the X's are the specific match characteristics. With dated donors from prior months, this is not literally true in the ORGs, since $f(y_t|X_1, X_2)$ may differ from $f(y_{t-1}|X_1, X_2)$, but it is not a bad approximation. With the March ADF, the assumption is violated when we draw from $f(y|X_1)$, the second or subsequent step matching only on some characteristics (an X at a broader level of detail). For both the ORG and the ADF, the question can be thought of as how different $f(y|X_1)$ is from $f(y|X_1, X_2)$. In general, there is probably less of a problem with ORG (last month's distribution is highly similar to this month's) than with the ADF (the earnings distribution of male, high school grads who work in a "narrow" occupation may be quite different than the distribution of male, high school grads for a "broad" occupation). For the ADF, the questions are, how often does the ADF move to matching with broader classifications and how different are

a reweighted respondent sample may be the preferable approach in a majority of applications. All of the approaches address the first-order match bias inherent in using the full uncorrected sample, but only IPW provides an easy and broadly applicable method to reweight the respondent sample to be representative of the full sample.

An alternative that we also briefly considered is to conduct one's own imputation (or multiple imputation) procedure, an approach that can be useful when tailored to a particular question at hand. For example, Hirsch and Schumacher (2004) conduct a simple cell hot deck imputation that adds union status as a match criterion, while Heckman and LaFontaine (2006, in this issue) add the GED as an imputation match variable. Unfortunately, imputation is not an attractive general approach. A hot deck imputation that eliminates (or sharply reduces) discrepancies between the information provided by the included regressors $\underline{z}_i$ and the more limited Census match criteria $\underline{x}_i$ comes at a cost. Adding imputation match criteria to a hot deck procedure leads to many thin and highly dated cells. We explore a simple alternative. We conduct a regression-based imputation for nonrespondents using the predicted value from the observed sample parameters, plus an error term. Not surprisingly, this approach produces estimates that are highly similar to the unweighted respondent sample results. It fails to account for composition bias owing to the use of the observed-only parameters and the absence of the detailed interactions implicit in a cell hot deck.

This section has demonstrated that attenuation of coefficients attached to variables not used as imputation match criteria is a concern of first-order importance and has compared alternative approaches to address match bias. In subsequent sections, the estimation approaches applied above for nonmatch attributes are used to account for bias from various forms of imperfect matching.

E. Imperfect Match on Multiple Categories

1. *Theory*

This section examines a less obvious form of match bias—bias for attributes that are match criteria but that are matched imperfectly. Specifically, we consider categorical variables $\underline{x}_i$ matched at a level more aggregated than that seen among the included $\underline{z}_i$ regressors. The example we emphasize is education, where nonrespondents are assigned earnings from donors within one of three broad education groups. The same logic applies

those distributions? Lillard et al. (1986) show that broad matches are frequent and often poor. Thus, our general full sample correction method is probably not as good applied to the ADF as to the ORG. Weighted (IPW) respondent estimation is likely to be the better (as well as simpler) choice for use with the March CPS.

to other match criteria.[16] We previously presented a general bias formulation for this and other cases of match bias. Discussion below illustrates with some simple cases the nature of the bias in estimating returns to schooling. For simplicity, this section assumes that missing at random holds. The results are qualitatively similar for weaker assumptions (see our unpublished appendix [Bollinger and Hirsch 2006] for more details).

Here we assume that $\underline{z}_i$ is a vector of $k - 1$ binary variables representing k mutually exclusive categories. We assume that $x_i = 1$ represents the "last" J^* categories of $\underline{z}_i$ while $x_i = 0$ represents the reference category and the remaining categories of $\underline{z}_i$. Formally we define

$$x_i = \sum_{j \geq J^*} z_{ji},$$

where z_{ji} is the jth element of $\underline{z}_i$.

We show in the appendix (Bollinger and Hirsch 2006) that

$$\begin{aligned} E_s[y_i | \underline{z}_i] = &\left(\alpha + p \sum_{j=1}^{J^*-1} \Pr[z_{ji} = 1 | x_i = 0] \beta_j \right) \\ &+ \sum_{j=1}^{J^*-1} z_{ji} (1 - p) \beta_j \\ &+ \sum_{j=J^*}^{k-1} z_{ji} \left((1 - p)\beta_j + p \sum_{l=J^*}^{k-1} \Pr[z_{li} = 1 | x_i = 1] \beta_l \right). \end{aligned}$$

Thus, in the regression of y_i on $\underline{z}_i$, the intercept will be α plus p times a weighted average of the β''s for the z_{ji} where $x_i = 0$. The coefficients on $\underline{z}_i$ when $x_i = 0$ will be $(1 - p)\beta_j$ and are simply downwardly biased. Finally, the coefficients on the z_{ij} where $x_i = 1$ will be $(1 - p)\beta_j$ plus p times the weighted average of all the β_j for z_{ji} where $x_i = 1$.

Consider a very simple case where there are four categories ($k = 4$) represented by three indicator variables ($k - 1 = 3$) but two of the categories are combined for the match procedure ($J^* = 2$), which results in a binary match variable x_i. In the regression of y_i on z_{1i}, z_{2i}, and z_{3i}, the intercept will be $\alpha + p\beta_1$. The coefficient on z_{1i} will be simply $(1 - p)\beta_1$. Since $\Pr[z_{2i} = 1 | x_i = 1] + \Pr[z_{3i} = 1 | x_i = 1] = 1$, the coefficient on z_{2i} will be

$$b_2 = \beta_2 + p(\beta_3 - \beta_2)\Pr[z_{3i} = 1 | x_i = 1].$$

[16] Only two imputation match criteria have exact matching, sex and the receipt of overtime, tips, or commissions. Note that some match variables are ordered (e.g., age, hours worked) whereas others are not (e.g., occupation, race).

If $\beta_3 > \beta_2$, the coefficient b_2 will be biased upward, while if $\beta_3 < \beta_2$, b_2 will be biased downward.

In the more general case, we note that $\sum_{l=j^*}^{k-1} \Pr[z_{li} = 1|x_i = 1]\beta_l$ is a weighted average of the β'’s for the $x_i = 1$ group. If β_j is less than this average, then the estimated coefficient will be inflated, while if β_j is more than this average, it will be attenuated.

Since these results generalize in a straightforward way, this indicates that regressions with a full set of education dummy variables will have estimated returns to schooling that are biased. It is not difficult to extend the model to include other match variables. It is important to note that, when other perfectly matched regressors are included as control variables, their coefficients will be biased as well if they are correlated with the mismatched variables.

2. *Evidence: Returns to Schooling*

Beginning in 1992, the CPS substituted an educational degree question for their previous measure of completed years of schooling. In 1998, additional questions were added to the CPS on receipt of a GED and years spent in school for both nondegree and degree students. Based on this information, one can construct detailed schooling degree/years variables that include well over 25 categories. One can also distinguish between years of schooling and highest degree, a “mixed” case examined in Section III.G. The ORG hot-deck imputation used since 1979 includes schooling as a match criterion, but it matches the earnings of donors to nonrespondents based on three broad categories of education, which we label “low” (less than a high school degree), “middle” (a high school degree, including a GED, through some college), and “high” (a BA degree or above).

Were schooling the only match criterion, the expected value of donor earnings matched to nonrespondents would be the average earnings among respondents within each broad schooling category. Donor earnings would increase across the three schooling groups but not within. Because other match criteria, in particular broad occupation, are correlated with schooling and earnings, imputed earnings may increase modestly within schooling groups. The schooling match creates an interesting form of match bias, flattening estimated earnings-schooling profiles within the low, middle, and high education groups and creating large jumps across groups.

Parts *a* and *b* of figure 1 provide separate estimates of schooling returns for respondents and imputed earners. Estimates are from male and female wage equations, using the same 1998–2002 CPS samples seen in the prior section. Shown in the figures are log wage differentials for each schooling group relative to earnings respondents with no zero schooling. Control

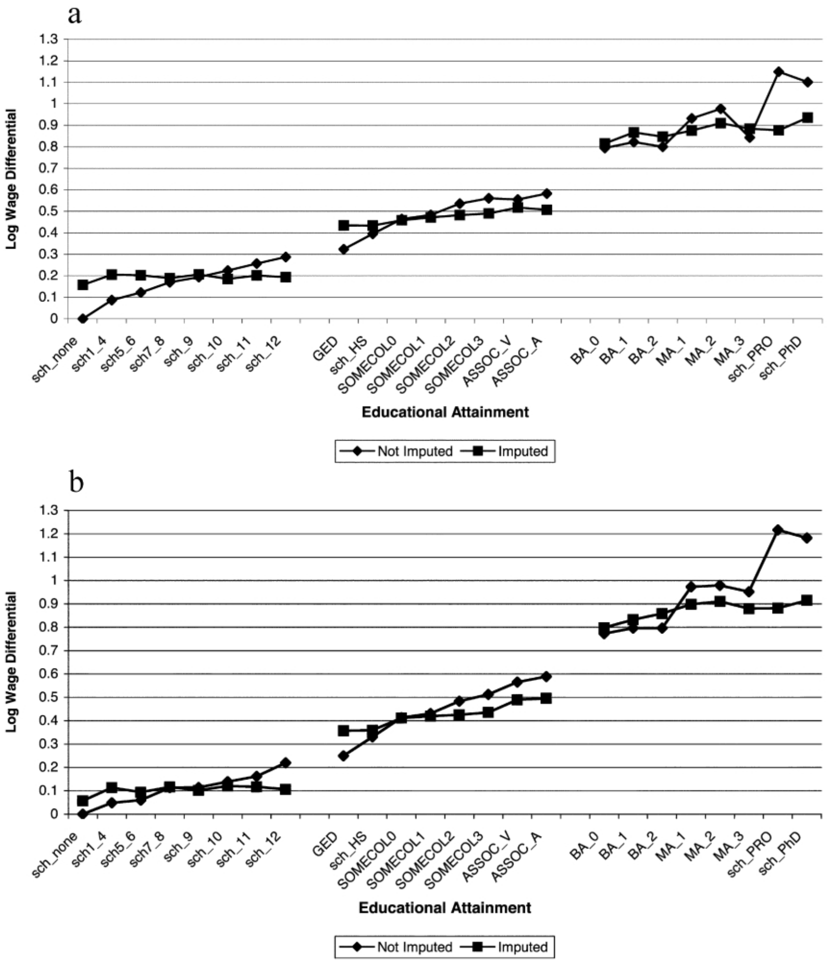

FIG. 1.—Schooling returns among male and female respondents and imputed earners, 1998–2002. Estimates are from a pooled wage equation of respondents and imputed earners using the CPS-ORG for 1998–2002. The male sample size (pt. *a*) is 388,578—276,909 respondents and 111,669 with earnings allocated (imputed) by the Census. The female sample size (pt. *b*) is 369,762—270,537 respondents and 99,225 with earnings allocated (imputed) by the Census. The sample includes all nonstudent wage and salary workers, ages 18 and over. Shown are log wage differentials for each schooling group relative to earnings respondents with no schooling. In addition to the education variables, control variables include potential experience (defined as the minimum of age minus years schooling minus 6 or years since age 16) in quartic form, race-ethnicity (four dummy variables for five categories), foreign born, marital status (2), part time, labor market size (6), region (8), and year (4).

variables are listed in the figure note. Variables that most clearly reflect postmarket outcomes (occupation, industry, union status, etc.) are not included.[17] The basic story seen in the figures is identical for women and men. The earnings of respondents (shown by "diamonds") increase fairly steadily with schooling level. In contrast, imputed earnings among nonrespondents ("squares") are essentially flat in the low education category and increase only slightly within the middle and high education categories. Failure to account for match bias leads to a downward bias in estimates for those at high education levels within each group and an upward bias for those with low education within each group. It leads to upwardly biased "jumps" in earnings as one moves across categories, specifically the movement from high school dropout to GED and from an associate's degree to a BA.

The GED results warrant examination. Here, upward match bias is severe, because the GED is the lowest education level within the middle education match category. Based on the sample of earnings respondents, the earnings gain for a male GED recipient relative to men who stop at 12 years of high school without a degree is a modest .036.[18] The same differential for imputed earners is an incredible .241 log points, as is seen in part *a* of figure 1 as the large jump between the Sch_12 and GED "squares." A standard wage equation using an uncorrected full sample would find a misleadingly large .087 wage gain for the GED (not shown), more than double the .036 estimate found for respondents. Similarly, imputation bias distorts the observed wage advantage for regular high school graduates as compared to GED recipients. The standard biased estimate indicates a .042 GED wage disadvantage, substantially smaller than the .072 GED disadvantage found among those with observed earnings. Among the sample of imputed earners, little wage difference is found between those with GEDs and standard diplomas. The story seen for women is extremely similar. As emphasized by Heckman and LaFontaine (2006, in this issue) and in previous literature, GED estimates are also biased upward by unobserved heterogeneity, a bias that we do not address here.[19]

[17] We do not interpret schooling parameters, even those corrected for match bias, as causal effects. Among other things, the estimates do not account for ability bias or reporting error in education.

[18] The CPS provides information on years of schooling completed prior to receipt of the GED. We do not use that information here, but we do use it in our subsequent analysis of "sheepskin" effects.

[19] Heckman and LaFontaine (2006, in this issue) provide a detailed analysis of the GED and imputation bias, including a critique of misleading results found in Clarke and Jaeger (2006). Using the post-1998 CPS, they show that the positive effect of the GED on earnings is small once one omits imputed earners or, alternatively, uses the GED as an imputation match criterion. Based on additional analysis using the NLSY (National Longitudinal Survey of Youth) and the NALS (National

Equally startling examples of bias from imperfect matching are seen among workers with professional degrees and PhDs. Match bias in this case is downward, owing to these groups having the highest education levels within the "high" schooling category but being matched primarily to donors with the BA as their terminal degree. Estimates from the respondent sample reveal a large .355 log point wage advantage among men with professional degrees as compared to men with BA degrees. Based on a standard full sample without correction, the wage advantage is .241, attenuation being 32%. The bias is similarly large for women, with a professional/BA degree wage advantage of .444 log points among earnings respondents versus .296 using the full sample, attenuation of 33%. A similar pattern of bias is readily evident for those with PhD degrees.

In short, match bias due to incomplete matching on education flattens wage-schooling profiles within educational match categories, while it steepens the jump in wages between categories. Depending on the specific level of schooling attainment being examined, bias can range from small to very large. In a subsequent section, we examine a mixed model with an ordinal schooling variable and categorical degree variables (sheepskin effects).

F. Imperfect Match on Ordinal Variables

1. *Theory*

Here we consider a simplified case where a scalar ordinal variable, such as age, enters a regression linearly but is reduced to two categories for purposes of the imputation match. We use the term *ordinal*, but the analysis applies equally well to ordered categorical variables and cardinal variables. Indeed, age (or experience) is typically treated as cardinal. For simplicity, this section assumes missing at random, but similar results hold for less restrictive assumptions. The specific structure is

$$E[y_i|z_i] = \alpha + \beta z_i$$

and

$$x_i = \begin{cases} 1 & \text{if } z_i > z^* \\ 0 & \text{if } z_i \leq z^* \end{cases}.$$

Given this simple structure, it follows then that

$$E[y_i|x_i] = \alpha + \beta E[z_i|x_i = 0] + \beta(E[z_i|x_i = 1] - E[z_i|x_i = 0])x_i.$$

Adult Literacy Survey), which permits an accounting for ability bias, the authors conclude that the remaining effects of the GED seen in the CPS are unlikely to be causal.

Substitution gives

$$E_s[y_i|x_i, z_i] = \alpha + (1-p)\beta z_i + p\beta(E[z_i|x_i = 0] \\ + (E[z_i|x_i = 1] - E[z_i|x_i = 0])x_i).$$

Then the linear projection of y_i on z_i gives an intercept of

$$a = \alpha - \beta(p(1 - R^2))E[z_i]$$

and a slope coefficient of

$$b = \beta(1 - p(1 - R^2)),$$

where R^2 is the squared correlation between z_i and x_i. The slope coefficient is attenuated by the proportion p imputed, mitigated in part by correlation between the information in match variables x_i and the nonmatch elements of z_i. This result generalizes to multiple categories and to the case of quadratic age: the quadratic profile is flattened relative to the true profile when imputed values are included.

Maintaining the assumption of missing at random, these results can be extended to the case where additional match characteristics are included in the regression. As with the previous case, all coefficients are biased.

2. *Evidence: Wage-Age Profiles*

As seen above, match bias resulting from imperfect matching arises in estimates of earnings profiles with respect to age (or potential experience). In the CPS, nonrespondents are matched to the earnings of donors in six age categories: ages 15–17, 18–24, 25–34, 35–54, 55–64, and 65 and over (our analysis includes nonstudent workers, 18 and over). Thus, the slopes of profiles are flattened within age categories, with jumps in earnings across categories. A simple way to illustrate the bias is to estimate linear wage-age profiles within each of the age categories using the respondent and imputed samples. We use a specification with largely "premarket" demographic and schooling variables, plus location and year controls. These results are shown in table 3.

The most notable bias is for young workers, whose wage-age profiles are steep. Focusing first on men, annual wage growth among respondents is .041 for ages 18–24 and .028 for ages 25–34. Wage growth seen among those with imputed earnings is far lower, .006 for ages 18–24 and .004 for ages 25–34. Wage growth is low in the 35–54 age interval, .005 in the respondent sample versus close to zero in the imputed sample. In the two oldest age categories, inclusion of imputed earnings causes wage decline to be understated. Identical patterns are seen for women, although overall wage-age growth is lower than for men (we observe wage growth with respect to age and not accumulated work experience). Whereas female

Table 3
Wage-Age and Wage-Experience Profile Estimates

	Men	Women
Linear wage growth per year within age groups:		
Respondents:		
18–24	.041	.029
25–34	.028	.020
35–54	.005	.002
55–64	−.021	−.011
65 plus	−.013	−.010
Imputed earners:		
18–24	.006	.001
25–34	.004	.002
35–54	.000	.000
55–64	−.007	−.002
65 plus	−.003	.004
Quadratic potential experience profiles:		
Respondents:		
Exp	.039	.025
$\text{Exp}^2/100$	−.068	−.044
Imputed earners:		
Exp	.035	.023
$\text{Exp}^2/100$	−.057	−.039
Pooled sample:		
Exp	.038	.024
$\text{Exp}^2/100$	−.065	−.042
Sample sizes:		
Respondents	276,909	270,537
Imputed earners	111,669	99,225
Pooled sample	388,578	369,762

NOTE.—The sample is all nonstudent wage and salary workers, ages 18 and over, from the CPS-ORG, 1998–2002. Control variables include a full set of education dummies, demographic variables, region, city size, and year. Specifications including age variables do not include potential experience.

respondents display annual wage growth of .029 for ages 18–24 and .020 for ages 25–34, growth using the imputed sample is effectively zero.

A more general way to illustrate the bias is to include a full set of age dummies and to estimate wage-age profiles for respondents and nonrespondents. These results are shown separately for men and women, respectively, in parts *a* and *b* of figure 2. Imputed earners exhibit substantial flattening of wage-age profiles within each age category, the bias being most serious for ages 18–24 and 25–34 when wage growth is highest. In the imputed worker sample, large wage jumps are observed between ages 24 and 25, between ages 34 and 35, and, going in the opposite direction, between ages 64 and 65. There is no jump between ages 54 and 55, since the weighted means of assigned donor earnings are similar in the adjacent age intervals.

Does inclusion of imputed earners greatly distort coefficients on potential experience in a Mincerian wage equation? The short answer is “a little.” The most typical wage equation includes potential experience as

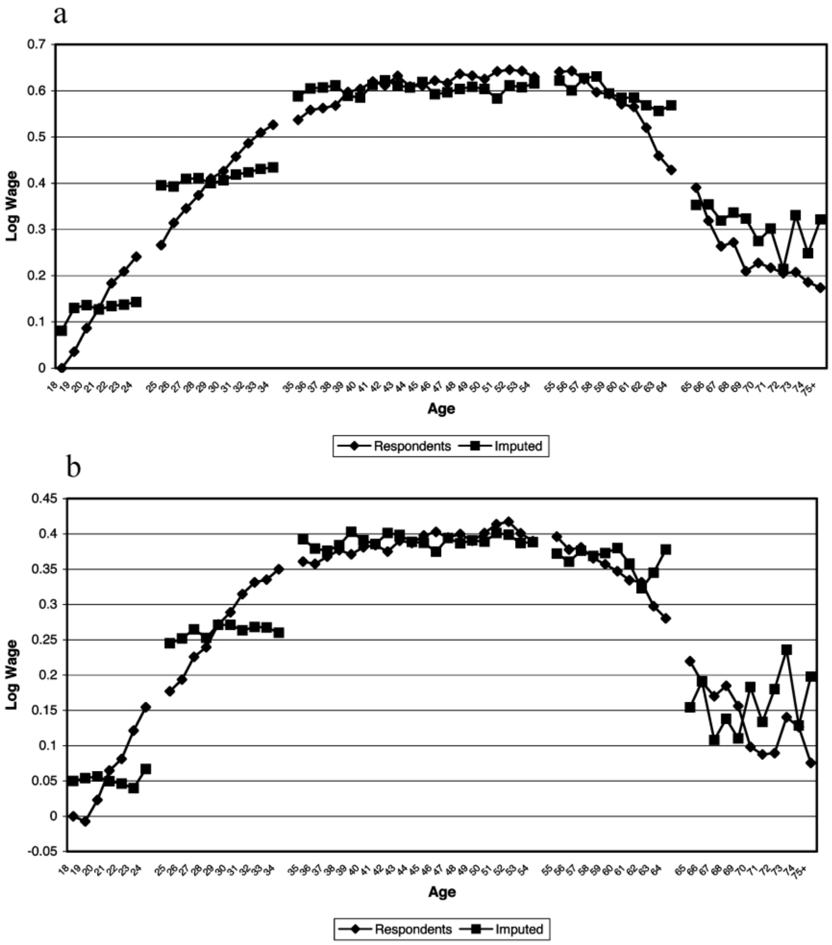

FIG. 2.—Male and female wage-age profiles (pts. *a* and *b*, respectively). Same samples as in figure 1. Shown are log wage differentials at each age relative to earnings of respondents who are age 18. In addition to the education dummies, control variables include race-ethnicity (four dummy variables for five categories), foreign born, labor market size (6), region (8), and year (4).

a quadratic.[20] In a male wage equation, respondents have a quadratic log wage profile of .039 and −.068 (to rescale coefficients, Exp^2 is divided by 100). Estimates for the imputed sample produce a flatter profile, .035 and −.057. Estimating the profile using the full sample without correction,

[20] Murphy and Welch (1990) and Lemieux (2006) make strong arguments for use of higher-order terms (e.g., up to a quartic) in the Mincerian wage equation, as was done in the regressions shown in tables 2 and 4 and fig. 1.

coefficient estimates are .038 and −.065, a profile slightly flatter than the one observed for respondents. Uncorrected standard errors (not shown) are much higher when imputed earners are included. An identical qualitative pattern is seen for women.

In short, bias due to imperfect matching causes wage patterns within and across age-match categories to be meaningless among imputed earners. Failure to account for this form of match bias has a modest effect in most applications, but it should not be ignored in analyses of earnings-experience (age) profiles, particularly those focusing on wage growth among young workers.

G. Mixed Case: Imperfect Matching with Ordinal and Multiple Category Variables

1. *Theory*

Education provides an important example of a mixed case. Some researchers observe that, in addition to a linear return to years of education, there are "sheepskin" effects, which result in jump discontinuities in the earnings-education profile. We examine the implications of match bias for this type of specification. Let z_{1i} be a dummy variable and let z_{2i} be an ordinal variable, with

$$z_{1i} = \begin{cases} 1 & \text{if } z_{2i} > z^* \\ 0 & \text{otherwise} \end{cases}.$$

We assume that

$$E[y_i | \underline{z}_i] = \alpha + \beta_1 z_{1i} + \beta_2 z_{2i}$$

and that $x_i = z_{1i}$. That is, the single match characteristic is the dummy variable. For simplicity, we assume MAR for this result. Following our unpublished appendix (Bollinger and Hirsch 2006) and recognizing that $x_i = z_{1i}$, the bias terms for the two slope coefficients will be

$$\begin{bmatrix} V_1 & C \\ C & V_2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & E[z_{1i}(z_{2i} - E[z_{2i}|z_{1i}])] \\ 0 & E[z_{2i}(z_{2i} - E[z_{2i}|z_{1i}])] \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix},$$

where V_1 is the variance of z_{1i}, V_2 is the variance of z_{2i}, and C is the covariance between z_{1i} and z_{2i}. The term $E[z_{1i}(z_{2i} - E[z_{2i}|z_{1i}])] = 0$, while the term $E[z_{2i}(z_{2i} - E[z_{2i}|z_{1i}])]$ is the variance of z_{2i} conditional on z_{1i}. Define R^2 as the squared correlation between z_{1i} and z_{2i}, and note that $E[z_{2i}(z_{2i} - E[z_{2i}|z_{1i}])] = V_2(1 - R^2)$. Then the above bias equation can be written as

$$\begin{bmatrix} V_1 & C \\ C & V_2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ 0 & V_2(1 - R^2) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

Evaluating leads to the following expressions for the bias from the least squares projection:

$$b_1 = \beta_1 + p\frac{C}{V_1}\beta_2,$$

$$b_2 = \beta_2(1-p).$$

Here we see that the degree effect will be overstated (since by definition of z_{1i} and z_{2i} the covariance will be positive), while the year or marginal effect will be understated. Indeed, if there is no degree effect (if $\beta_1 = 0$), its OLS estimate will still be positive while the marginal effect will be understated. It must be kept in mind that the presence of other variables will alter these results.

2. *Evidence: Sheepskin Effects and Linearity*

A common approach in estimating the returns to schooling is to assume linearity and to include a single schooling variable measuring years of school completed. The schooling coefficient represents the percentage (log) wage gain associated with an additional year of schooling (see Mincer [1974], Willis [1986], and subsequent literature for assumptions necessary to interpret this as a rate of return). A related approach includes indicators for completed degrees, measuring separately the effect of the sheepskin on earnings. This approach can be informative (but not decisive) in determining the extent to which education increases human capital and the extent to which it provides some verifiable signal of innate human capital or motivation. In the extreme (and ignoring complicating factors), if education is exclusively human capital enhancement, then the coefficients on the degree completion indicators should approach zero and years of schooling should measure the full human capital effect. If education provides only a signaling mechanism, then the coefficient on years schooling should approach zero and only the degree effects should matter.[21]

Table 4 provides estimates of a model with these mixed education variables. The sample is restricted to the range of data over which we can clearly distinguish between years of schooling and degree. We omit the relatively few workers with less than 9 years of schooling or with professional and PhD degrees for whom separate information on years

[21] If unmeasured ability differences lead degree recipients to acquire more human capital per year of schooling than do nonrecipients, estimates of degree effects will be positively biased.

Table 4
Estimated Schooling and Sheepskin Effects, 1998–2002

	Full Sample	Imputed	Respondents	Respondents (IP Weighted)	Full Sample (Corrected)
Men:					
School (years completed)	.036	.022	.042	.043	.046
GED	.119	.251	.067	.067	.068
High school	.136	.230	.097	.094	.092
Associate's degree	.190	.270	.156	.151	.160
BA	.367	.549	.294	.287	.268
Master's	.414	.587	.345	.337	.335
N	359,564	103,476	256,088	256,088	359,564
Women:					
School (years completed)	.048	.030	.054	.056	.062
GED	.129	.236	.091	.093	.082
High school	.137	.224	.104	.104	.088
Associate's degree	.237	.290	.215	.213	.214
BA	.368	.562	.297	.293	.252
Master's	.440	.595	.382	.375	.347
N	353,585	95,120	258,465	258,465	353,585

Note.—The sample is drawn from the CPS-ORG, 1998–2002. It includes all nonstudent wage and salary workers, ages 18 and over, with between 9 years of schooling and a master's degree (omitted are those with schooling of less than 9 years, those with professional degrees, and PhDs). Control variables include a full set of demographic variables, region, city size, and year. The full sample includes both the respondent (observed) and imputed (missing) samples with Census imputation. Corrected estimates are based on the full sample and the general bias correction shown in the text. The IP-weighted column reports least squares estimates from the respondent sample reweighed by the inverse probability that an individual's earnings are reported.

schooling is not provided.[22] Estimates are provided using the full sample with Census imputations and no bias correction (the standard approach), the respondent ("observed") sample, the observed sample with inverse probability weighting (IPW), and the full sample using the general correction measure derived in Section III.C.

School is the measure of years of schooling completed. The full sample estimate for men suggests a rate of return of .036 (in log points) for a year of schooling, holding degree constant. The estimate on the observed sample is .042 absent weights and .043 reweighted to adjust for a changed sample composition. The corrected full sample estimate is .046, a percentage point larger than the uncorrected estimate. Some of the degree indicators, absent correction, are very misleading. For example, the coefficient on high school degree in the full sample is .136. The estimates from the observed sample, the IPW observed sample, and the corrected full sample are much smaller at .097, .094, and .092, respectively. Similarly,

[22] MA recipients designate their program as a 1, 2, or 3+ year program. Information on additional years schooling is provided for those with some college and no degree and for BA degree recipients with graduate course work but no degree. Those with some college but no postsecondary degree are coded as having received a regular high school diploma (information on the GED is provided only for those without education beyond high school).

the estimated effect of a GED (years constant) is overstated due to match bias. The full sample estimate places the value of a GED at .119, while the observed, IPW observed, and full corrected sample estimates are only .067, .067, and .068, respectively.[23]

The results for women follow a similar pattern. The full sample return estimate of .048 is less than estimates from the observed sample of .054, the reweighted observed sample of .056, and the corrected full sample of .062. The GED full sample estimate of .129 compares to estimates of .091, .093, and .082 from the unweighted observed, IPW observed, and corrected full samples. Estimates of the value of a high school degree are very similar to those for men.

The results confirm that imperfect group matching using the Census imputation procedure biases rate of return estimates, trivially for some schooling groups but substantially for others. In a sheepskin model, the Census imputation tends to understate the returns to years of schooling while generally overstating degree effects. Sheepskin effects are still evident, but these are less pronounced than those seen with observed Census earnings or from estimates corrected for match bias.

IV. Dated Donors

Earnings nonrespondents are assigned the nominal earnings of the donor who is the most recent respondent with an identical mix of match attributes. During the 1994–2002 period, the Census match procedure included 14,976 cells or combinations of match characteristics. For match cells with a relatively uncommon mix, donor earnings may be relatively dated, biasing downward imputed earnings owing to nominal and real wage growth. Stated alternatively, the survey month can be considered a wage determinant in $\underline{z}_i$ that, for nonrespondents, is imperfectly mapped from $\underline{x}_i$.

How serious is the dated donor problem? The Census does not record the "shelf age" of donor earnings assigned to nonrespondents. To assess this issue, one must approximate Census hot deck methods and measure the datedness of donor earnings. Our analysis begins with all employed wage and salary workers, ages 18 and over, from the December 2002 CPS. That month's file contains 4,759 nonrespondents. Some of these individuals will be matched to donor earnings in the current month, while most will reach back to donors in previous months and years. Each nonrespondent in December 2002 is given a unique match number corresponding to the 14,976 possible combinations of match attributes. Likewise, po-

[23] Note that these estimates account for the years of schooling completed by GED recipients (mostly 9–12 years). Prior estimates of a GED effect, drawn from fig. 1, did not include a separate years schooling variable and compared GED recipients to those with 12 years schooling but no degree.

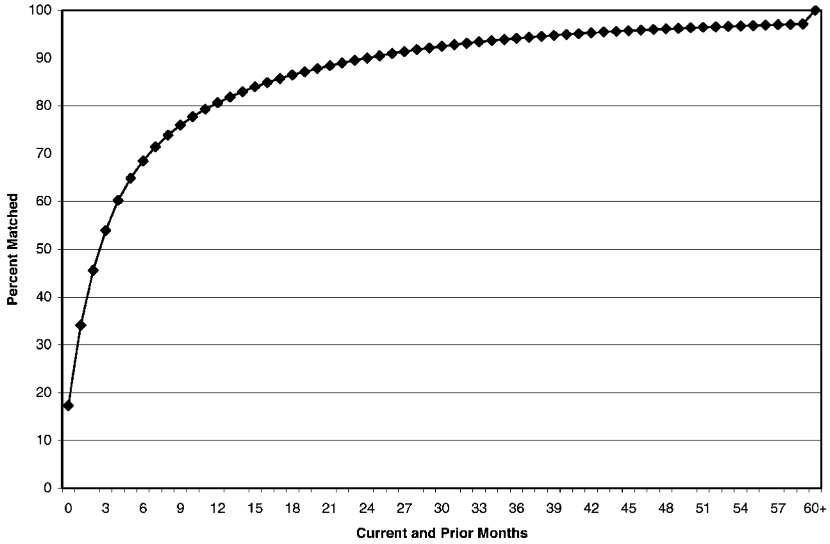

FIG. 3.—Dated donors: CPS cumulative imputation match rate for current and prior month donors. Cumulative monthly match rates of CPS-ORG nonrespondents in 2002 to 1998–2002 potential donors. Period 0 represents donor matches in the current survey month, while period *n* represents donor matches in the *n*th prior month. Period 60+ figures represent all nonrespondents not finding a donor match during the period 1998–2002.

tential donors (respondents) in 60 monthly CPS earnings files (December 2002 back to January 1998) are assigned attribute match numbers on the same basis. We first examine whether at least one donor match exists for each nonrespondent in December 2002. Those not finding a donor are retained, and a search for a donor in November 2002 is executed. This process continues back to January 1998. In order to increase the size and representativeness of the nonrespondent sample, we conduct the identical analysis for nonrespondents during from January to November 2002. The total number of nonrespondents during 2002 is 55,902.[24]

Cumulative match rates resulting from the donor match exercise are shown in figure 3. In the initial month, just 17.3% of 2002 nonrespondents find a same-month donor.[25] Reaching back 1 month, an additional 16.8%

[24] For ease of programming, nonrespondents during each month of 2002 are treated as if they were December nonrespondents. That is, for each 2002 nonrespondent, we first search for matching donors in December 2002, and then we reach back in time as far as January 1998.

[25] To approximate the Census match rate in the initial month, the donor pool is constructed by taking a 50% random sample of December 2002 respondents. The Census searches for donors among those who are listed in the file layout prior to the nonrespondent. Thus, nonrespondents at the beginning of the December 2002 file are assigned donors from November 2002 or earlier, whereas nonrespondents

are matched, followed by 11.5% and 8.3% reaching back 2 and 3 months. Within these first 4 survey months (the sample month plus 3 months back), over half (53.9%) of all nonrespondents are assigned donor earnings. Those not finding matches have decreasing match hazards (probabilities of finding a match) in subsequent months. Even after 5 years, reaching back 59 months from month zero to January 1998, 2.85% of nonrespondents remain without an earnings assignment and are assigned donor earnings in excess of 5 years old. In figure 3, we add the residual monthly match rate of 2.85% to the prior month labeled 60 plus.

Beginning in 2003, the number of occupation categories in the Census match algorithm was reduced from 13 to 10, reducing the number of hot deck cells from 14,976 to 11,520. In order to see how this affects donor datedness, we provide an analysis matching the 17,864 earnings nonrespondents in the January–April 2004 period to donors beginning in April 2004 and reaching back to January 2003 (the first month with the new occupation codes). We find little change in average donor datedness. Whereas 53.9% of the 2002 nonrespondents found donors during the current or 3 previous months, the corresponding number for the January–April 2004 nonrespondents is 53.1%. Reaching back 15 months, 84.0% of the 2002 nonrespondents found a match; the corresponding number for 2004 respondents is 83.4%. We conclude that donor datedness has not appreciably changed as a result of the revised occupational match categories beginning in 2003.

How serious is the problem of dated donor earnings? Combining information on average donor age with the rate of wage growth, one can estimate the downward bias in average earnings. To calculate mean donor age one must assume an average match date for the nonrespondents who have failed to find a match in the previous 5 years. For the 2002 sample of nonrespondents, we assume that the 2.85% not matched going back to January 1998 would on average find a match in 6 additional months. Using this assumption, the average age or datedness of all donor earnings is 8.6 months, or nearly three-quarters of a year, substantially larger than the median age of 3 months (the current month and 3 months back).[26] If nominal wage growth were, say, 3% annually, this would imply that the average earnings of donors are understated by 2.25%. With approximately 30% of the CPS sample being nonrespondents, the CPS understates average earnings by .675% (three-quarters of a year times 3% annual wage

at the end of the file can be matched to the full month donor sample. We approximate this by using a half donor sample in the initial month (and full samples thereafter). If we instead search through all December respondents for donors, the initial match rate increases by several percentage points and the next month rate falls, with quick convergence in subsequent months to the rates in fig. 3.

[26] The estimate of an 8.6 month mean donor age is sensitive to the assumed average match date for those relatively few (2.85%) nonrespondents remaining unmatched.

growth times .30 proportion donors), or two-thirds of a percentage point. In 2004, average hourly earnings compiled from the CPS, including imputed earners, is $17.69, 2.85% higher than the 2003 average of $17.20. Multiplying by .0064 (.75 times 2.85% times .30), earnings are understated by $.11, with the true average wage closer to $17.80. This was a period of modest nominal wage growth; the bias increases proportionately with the growth rate.

Do dated donors affect wage gap estimates? To the extent that a "treatment" group of workers has more (less) dated donors than a comparison group, the treatment group wage gap will be understated (overstated). Comparisons of the average datedness of donors across various groups of workers based on gender and race suggest that standard wage gap estimates in the literature have been affected little by bias from dated donors.

Most CPS nonrespondents are matched to the nominal earnings of donors from prior months rather than the current month, causing earnings to be understated. The resulting bias for most labor market studies, however, is modest and does not warrant serious concern. If nominal wage growth were to increase sharply in future years, this conclusion would warrant reconsideration.

V. Conclusion

Match bias arising from Census earnings imputation is an issue of some consequence, but it is not one that generally has been considered by labor economists. Given the assumption of conditional mean missing at random (CMMAR), this article derives a general analytic solution that measures match bias in its multiple forms. Bias is of first-order concern in studies estimating wage gaps with respect to attributes that are not Census match criteria (union status, foreign born, etc.). Attenuation in this case is roughly equal to the imputation rate, nearly 30% in recent CPS earnings surveys. Consistent estimates can be obtained from samples including only earnings respondents (weighted or unweighted) or from the full sample corrected for match bias.

This article shows that earnings imputation also warrants concern where there is matching on an attribute but the match is imperfect (e.g., education, age, occupation). Matching across a range of values flattens estimated earnings profiles within match categories (say, low, middle, and high education), while creating jumps across categories. Such match bias can be modest or severe, leading to overstatement (e.g., returns to the GED) or understatement (e.g., returns to professional and doctoral degrees). We also draw attention to rather subtle forms of match bias, for example, understatement of imputed earnings due to the datedness of donors (also see Hirsch 2005).

For the applied researcher, the simplest approach to account for match bias is to omit imputed earners from wage equation (and other) analyses. Alternatively, one can retain the full sample and calculate corrected parameter estimates as shown in this article. Under the assumption of CMMAR and absent specification error, either set of parameter estimates is consistent. In practice, these approaches differ a bit. If one is concerned about composition effects but does not wish to implement the analytic match bias correction outlined in this article, a simple alternative is inverse probability weighted least squares estimation on the respondent sample. The IPW method has the added advantage of greater generality, being appropriate with surveys whose imputation methods differ substantively from the Census cell hot deck.[27]

Discussion in this article has examined the CPS-ORG earnings files and the estimation of earnings equations. Similar issues arise with the March CPS ADF and other household surveys, although rates of nonresponse are generally lower than in the ORGs and imputation methods (where used) differ from the cell hot deck. Although our focus has been on earnings imputation, similar issues arise for other variables whose values are imputed and are used as outcome (dependent) variables in empirical work. Fortunately, nonresponse rates on nonincome related variables tend to be small. Finally, earnings (income) is often used as an explanatory variable. If the dependent variable is not a Census match criterion, there will exist attenuation in the earnings coefficient for precisely the same reason as seen in our discussion of match bias.

Ultimately, the moral of this story is that earnings imputation must be given serious consideration by applied researchers. Match bias resulting from imputation is often large and shows up in surprising places. Authors should add match bias to their already long checklist of issues to consider. Census and the Bureau of Labor Statistics should be more forthcoming about the methods used to impute earnings (income).[28] Where an earnings variable is used as a dependent or a key independent variable, researchers should use a sample of earnings respondents (unweighted or reweighted)

[27] Even ignoring match bias, a case can be made to use WLS with Census weights when using the full sample, given that the CPS is not fully representative (Polivka 2000; Helwig, Ilg, and Mason 2001). Because our results were affected little by the use of Census weights, we have not followed that approach. As discussed in Sec. III.D, it is sometimes practical to retain the full sample and implement one's own imputation procedure, using the particular characteristic of interest as a match variable.

[28] Our focus is on how researchers can deal with Census imputation methods. Given the severity of the match bias problem, attention ought to be given as well to possible changes in these methods. Given current methods, we recommend that the BLS enter missing values in the edited weekly (and hourly) earnings fields typically used by researchers, while providing imputed values in separate fields. Use of imputed values would require an explicit decision to do so.

or provide corrected full sample coefficient estimates. Inclusion of imputed earners absent bias correction should not occur, absent a persuasive argument for doing so. Such arguments are not easy to make.

References

Aigner, Dennis J. 1973. Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics* 1, no. 1:49–59.

Angrist, Joshua D., and Alan B. Krueger. 1999. Empirical strategies in labor economics. In *Handbook of labor economics*, vol. 3A, ed. Orley C. Ashenfelter and David Card, 1277–1366. Amsterdam: Elsevier.

Black, Dan A., Mark C. Berger, and Frank A. Scott. 2000. Bounding parameter estimates with non-classical measurement error. *Journal of the American Statistical Association* 95 (September): 739–48.

Bollinger, Christopher R. 1996. Bounding mean regressions when a binary regressor is mismeasured. *Journal of Econometrics* 73 (August): 387–99.

Bollinger, Christopher R., and Barry T. Hirsch. 2006. Appendix accompanying "Match bias from earning imputation in the Current Population Survey," *Journal of Labor Economics* 24 (July). Available at either http://gatton.uky.edu/faculty/bollinger/Workingpapers/jole_app.pdf or http://www.trinity.edu/bhirsch/jole_app.pdf.

Card, David. 1996. The effect of unions on the structure of wages: A longitudinal analysis. *Econometrica* 64 (July): 957–79.

Clarke, Melissa A., and David A. Jaeger. 2006. Natives, the foreign-born, and high school equivalents: New evidence on the returns to the GED. *Journal of Population Economics* (forthcoming).

Groves, Robert M. 2001. *Survey nonresponse*. New York: Wiley.

Groves, Robert M., and Mick P. Couper. 1998. *Nonresponse in household interview surveys*. New York: Wiley.

Heckman, James J., and Paul A. LaFontaine. 2006. Bias-corrected estimates of GED returns. *Journal of Labor Economics* 24, no. 3:661–700.

Helwig, Ryan T., Randy E. Ilg, and Sandra L. Mason. 2001. Expansion of the Current Population Survey sample effective July 2001. *Employment and Earnings* 48 (August): 3–7.

Hirsch, Barry T. 2005. Why do part-time workers earn less? The role of worker and job skills. *Industrial and Labor Relations Review* 58 (July): 525–51.

Hirsch, Barry T., and Edward J. Schumacher. 2004. Match bias in wage gap estimates due to earnings imputation. *Journal of Labor Economics* 22 (July): 689–722.

Horowitz, Joel L., and Charles F. Manski. 1998. Censoring of outcomes and regressors due to survey non-response: Identification and estimation using weights and imputations. *Journal of Econometrics* 84 (May): 37–58.

———. 2000. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association* 95 (March): 77–84.

Lemieux, Thomas. 2006. The "Mincer Equation" thirty years after *Schooling, experience, and earnings*. In *Jacob Mincer, a pioneer of modern labor economics*, ed. Shoshana Grossbard. New York: Springer Verlag.

Lillard, Lee, James P. Smith, and Finis Welch. 1986. What do we really know about wages? The importance of nonreporting and Census imputation. *Journal of Political Economy* 94 (June): 489–506.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical analysis with missing data*. New York: Wiley.

Mincer, Jacob. 1974. *Schooling, experience, and earnings*. New York: Columbia University Press.

Molinari, Francesca. 2005. Missing treatments. Photocopy, Department of Economics, Cornell University (June).

Murphy, Kevin M., and Finis Welch. 1990. Empirical age-earnings profiles. *Journal of Labor Economics* 8 (April): 202–29.

Polivka, Anne E. 2000. Using earnings data from the Monthly Current Population Survey. Photocopy, Bureau of Labor Statistics (October).

Schafer, Joseph L., and Nathaniel Schenker. 2000. Inference with imputed conditional means. *Journal of the American Statistical Association* 95 (March): 144–54.

Shao, J., and R. R. Sitter. 1996. Bootstrap for imputed survey data. *Journal of the American Statistical Association* 91 (September): 1278–88.

U.S. Department of Labor, Bureau of Labor Statistics. Various years. Median weekly earnings of full-time wage and salary workers by union affiliation, occupation and industry, Washington, DC. http://www.bls.gov/cps/cpsaat43.pdf.

———. 2002. *Current Population Survey: Design and methodology*. Technical Paper 63RV, Bureau of Labor Statistics, Washington, DC (March). http://www.bls.census.gov/cps/tp/tp63.htm.

Willis, Robert J. 1986. Wage determinants: A survey and reinterpretation of human capital earnings functions. In *Handbook of labor economics*, vol. 1, ed. Orley C. Ashenfelter and Richard Layard. Amsterdam: Elsevier.

Wooldridge, Jeffrey M. 2002. *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wu, Lang. 2004. Exact and approximate inferences for nonlinear mixed effects models with missing covariates. *Journal of the American Statistical Association* 99 (September): 700–709.